Multiscale and integrative single-cell Hi-C analysis with Higashi

Ruochi Zhang[1], Tianming Zhou[1], and Jian Ma[1,*]

[1]Computational Biology Department, School of Computer Science,
Carnegie Mellon University, Pittsburgh, PA 15213, USA
[*]Correspondence: jianma@cs.cmu.edu

Abstract

The advent of single-cell Hi-C (scHi-C) technologies offers an unprecedented opportunity to unveil cell-to-cell variability of 3D genome organization. However, the development of computational methods that can effectively enhance scHi-C data quality and extract 3D genome features in single cells remains a major challenge. Here, we report Higashi, a new algorithm that enables robust analysis of scHi-C data based on hypergraph representation learning. Extensive evaluation and integrative analysis demonstrate that Higashi significantly outperforms existing methods for embedding and imputation of scHi-C data. Higashi is uniquely able to reveal multiscale 3D genome features (such as compartmentalization and TAD-like domain boundaries), allowing markedly refined analysis of structure and function connections at single-cell resolution. We show that Higashi can also be applied to the recently developed single-cell co-assays that jointly profile scHi-C and additional epigenomic features. Higashi enables much improved delineation of functionally important cell-to-cell variation of 3D genome organization using scHi-C data.

Introduction

The rapid development of whole-genome mapping methods such as Hi-C (Lieberman-Aiden et al., 2009) for probing the three-dimensional (3D) genome organization inside the nucleus has revealed multiscale higher-order chromatin structures (Kempfer and Pombo, 2020), including A/B compartments (Lieberman-Aiden et al., 2009), more refined nuclear compartmentalization (Rao et al., 2014; Xiong and Ma, 2019; Wang et al., 2020), topologically-associating domains (TADs) (Dixon et al., 2012; Nora et al., 2012), and chromatin loops (Rao et al., 2014). These 3D genome features in different scales are interconnected with vital genome functions such as gene transcription and DNA replication (Dekker et al., 2017; Marchal et al., 2019), yet the variation of 3D genome structures and its functional implication in single cells remain mostly unclear (Misteli, 2020). The emerging single-cell Hi-C (scHi-C) technologies have enabled genomic mapping of 3D chromatin structures in individual cells (Nagano et al., 2013; Stevens et al., 2017; Flyamer et al., 2017; Ramani et al., 2017; Nagano et al., 2017; Tan et al., 2018) and, more recently, joint profiling of chromosome conformation with other epigenomic features (Lee et al., 2019; Li et al., 2019). These exciting scHi-C assays have the potential to comprehensively reveal fundamental genome structure and function connections at single-cell resolution in a wide range of biological contexts.

However, computational methods that can fully leverage the scHi-C data to analyze the cell-to-cell variability of 3D genome features are significantly lacking. To account for the sparseness and high-dimensionality of scHi-C data, methods have been developed to represent the chromatin interaction map (i.e., contact map) of each cell with a lower-dimensional vector (i.e., embedding) (Liu et al., 2018; Kim et al., 2020). Meanwhile, computational methods for the analysis of 3D genome features in single cells based on noisy and sparse scHi-C data are extremely limited. scHiCluster (Zhou et al., 2019) uses linear convolution and "random walk with restart" for data imputation and identifies TAD-like structures from scHi-C data. However, the imputation effectiveness of scHiCluster is diminished for more distal chromatin interactions. It also remains unclear how to reliably compare TAD-like domain boundaries across single cells. Additionally, no method currently exits for robust analysis of compartmentalization based on scHi-C data. Therefore, new algorithms are urgently needed to fill these important gaps to uncover cell-to-cell variability of multiscale 3D genome features.

Here, we report Higashi, a new computational method for multiscale and integrative single-cell Hi-C analysis using hypergraph representation learning. One critical novel design in Higashi is to allow neighboring cells in the embedding space to share information, synergizing the imputation among different single cells based on their latent correlations. Our comprehensive evaluations show that Higashi consistently achieves better performance compared with existing methods for both embedding and data imputation. Our method is also able to leverage co-assayed scHi-C data with additional epigenomic feature. Crucially, Higashi is uniquely able to uncover cell-to-cell variability of A/B compartment scores and TAD-like domain boundaries that exhibits significant correlation with gene transcription. Higashi enables much improved analysis of scHi-C data with the potential to shed new light on the dynamics of 3D genome structures and their functional implications in different biological processes.

Results

Overview of Higashi

The key algorithmic design of Higashi is to transform the scHi-C data into a hypergraph (Fig. 1a). Such transformation preserves the single-cell resolution and 3D genome features of the scHi-C contact maps. Specifically, the process of embedding the scHi-C data is now equivalent to learning node embeddings of the hypergraph, while imputing the scHi-C contact maps becomes predicting missing hyperedges within the hypergraph. In Higashi, we leverage our recently developed Hyper-SAGNN architecture (Zhang et al., 2020), which is a generic hypergraph representation learning framework, with substantial new development specifically for scHi-C analysis (see Methods).

Higashi has four main components. (1) We represent the scHi-C dataset as a hypergraph, where each cell and each genomic bin are represented as cell node and genomic bin node, respectively. Each non-zero entry in the single-cell contact map is modeled as a hyperedge connecting the corresponding cell and the two genomic loci of that particular chromatin interaction (Fig. 1a). The read count for each chromatin interaction is used as attribute of the hyperedge. (2) We train a hypergraph neural network based on the constructed hypergraph (Fig. 1b). (3) We extract the embedding vectors of cell nodes from the trained hypergraph neural network for downstream analysis. (4) We then use the trained hypergraph neural network to impute single-cell Hi-C contact maps with the flexibility to incorporate the latent correlations between cells to enhance overall imputation, enabling detailed characterization of 3D genome features such as TAD-like domain boundaries and A/B compartment scores at single-cell resolution. In addition, as part of Higashi, we have developed new computational strategies that can reliably compare A/B compartment scores and TAD-like domain boundaries across individual cells. The details are described in Methods.

Higashi embeddings reflect cell types and cellular states

We sought to demonstrate that Higashi effectively captures the variability of 3D genome structures from the sparse scHi-C data with the embedding vectors. We tested our method on three scHi-C datasets with multiple cell types or known cell state information at 1Mb resolution. These datasets include: 4DN sci-Hi-C dataset (Kim et al., 2020) (5 cell types: GM12878, H1ESC, HAP1, HFFc6, IMR90), Ramani et al. dataset (Ramani et al., 2017) (4 cell types: GM12878, HAP1, HeLa, K562), and Nagano et al. dataset (Nagano et al., 2017) (cell cycle stages). See Methods for data processing and Tables S1 and S2 for statistics of theses datasets. After training, the Higashi embeddings are projected to a two-dimensional space with UMAP (McInnes et al., 2018) for visualization (Fig. 2a-c). We quantified the effectiveness of the embeddings with various supervised and unsupervised evaluation settings and made direct comparisons to two existing scHi-C embedding methods, HiCRep/MDS (Liu et al., 2018) and scHiCluster (Zhou et al., 2019) (Supplementary Methods B.2). All evaluation results are in Fig. 2d-f.

For both the 4DN sci-Hi-C dataset and Ramani et al. dataset, we observed that the UMAP visualizations of the Higashi embeddings have clear patterns where the cells of the same cell type are clustered together (Fig. 2a-b). The quantitative results further suggest the effectiveness of Higashi embeddings as it consistently outperforms the other two methods in both supervised and unsupervised tasks (Fig. 2d-e). For the Nagano et al. dataset where the cells were labeled according to their cell-cycle stage, we found that the Higashi embeddings exhibit a circular pattern that reflect the cell cycle (Fig. 2c). Both

Higashi and HiCRep/MDS achieve the highest ACROC score, while Higashi reaches the highest Micro and Macro F1 scores (Fig. 2f). Overall, Higashi consistently achieves the best performance on scHi-C datasets with either categorical cell types or continuous cell states. Although these results are based on the embedding with dimension size 64, our sensitivity analysis on the embedding dimension shows that Higashi is more robust to the choice of dimension size (Supplementary Results A.1 and Fig. S1a).

Together, these results demonstrate that Higashi effectively captures the cell-to-cell variability of 3D genome structures with scHi-C embeddings that accurately reflect the underlying cellular states.

Higashi robustly imputes scHi-C contact maps

In addition to dimension reduction of scHi-C data for cell type identification, Higashi also imputes missing data in sparse scHi-C contact maps. Here, we sought to demonstrate the imputation accuracy with several evaluations. For comparisons, we included the imputed results from scHiCluster. Note that scHiCluster represents each scHi-C contact map as an individual graph, whereas Higashi represents the whole scHi-C dataset as a hypergraph, allowing imputation to be potentially coordinated across different cells. Specifically, in Higashi, when imputing the contact map of cell i, its k-nearest neighbors in the embedding space would contribute to the imputation by taking advantage of their latent correlations (see Methods). To demonstrate the advantages of this important design employed in Higashi, we included the imputed results from Higashi with k equals 0 and 4 (referred to as Higashi(0) and Higashi(4), respectively). Importantly, we performed sensitivity analysis on the hyperparameter k and showed that Higashi is highly robust to the choice of k (Supplementary Methods A.1 and Fig. S1b).

We generated simulated scHi-C data based on the multiplexed STORM imaging of chromatin conformation, which provides high-resolution physical views of 3D organization of genomic loci in individual cells (Bintu et al., 2018). We turned the imaging data of a 2.5Mb region on chr21 from 11,631 cells at 30Kb resolution into scHi-C contact maps with various simulation coverage (Supplementary Methods B.1). We found that Higashi(0), i.e., no information sharing among different cells, can already consistently outperform scHiCluster. In addition, we found that Higashi(4) improves the imputation most significantly. To illustrate why using neighboring cells in the embedding space improves imputation, we show a typical example from the simulated data with contact maps before and after imputation (Fig. 3a and Fig. S3). Consistent with the quantitative evaluation, Higashi(4) reveals the clearest patterns and domain boundaries across all coverage. Importantly, the neighboring cells in the embedding space that contribute to the imputation have similar 3D chromatin interactions compared with the selected cell, while the furthest cells do not. We carried out a similar set of evaluation using the more recent multiplexed imaging data of 3D genome structure (Su et al., 2020) (3,029 simulated contact maps of chr2 at 1Mb resolution) and reached the same conclusion of Higashi's clear advantage (Fig. S4). These simulation evaluations based on imaging data demonstrate that the embeddings generated by Higashi accurately reflect the cell-to-cell variability of the chromatin structures and the strong advantage of sharing information of neighboring cells in the embedding space for imputation.

We performed additional evaluation via downsampling the existing scHi-C datasets with relatively higher coverage (Supplementary Methods B.1). We used the WTC-11 scHi-C dataset (personal communication with Bing Ren) of chr1 at 1Mb resolution and downsampled the sequencing reads of each cell at different rate (Methods; Table S1 and Table S4). We again observed clear advantage of Higashi for imputation with the strongest performance achieved by Higashi(4) (Fig. 3b and Fig. S5).

Together, these evaluations demonstrate that Higashi achieves improved imputation of scHi-C contact

maps robustly. The performance is further empowered by the unique mechanism of sharing information among neighboring cells in the embedding space. The improved imputation enables more reliable analysis of 3D genome structural features of each individual cell with higher accuracy.

Higashi reveals compartmentalization variability at single-cell resolution

Next, we explored how the enhanced contact maps produced by Higashi facilitate multiscale 3D genome analysis at single-cell resolution. A/B compartments reflect large-scale chromosome spatial segregation with distinct connections to genome function (Lieberman-Aiden et al., 2009). To date, no systematic A/B compartment annotation has been performed using scHi-C data primarily because of the data sparseness. Here, we applied Higashi to impute the WTC-11 scHi-C at 50Kb resolution to balance between computational efficiency and precision (see examples of the imputation results in Fig. S6). We designed a new method to calculate continuous compartment scores such that the scores are directly comparable across the cell population and reflect detailed variation (Supplementary Methods B.3).

Fig. 4a shows the merged correlation matrices (Pearson correlation of the merged contact maps) before and after Higashi imputation as well as the compartment scores from the pooled scHi-C (i.e., bulk) and the single-cell compartment scores of chr21. After imputation, the merged scHi-C correlation matrix has much clearer checkerboard patterns that correspond to A/B compartments. The calculated single-cell compartment scores are overall consistent with the bulk compartment scores while showing cell-to-cell variability. Note that we identified one cluster of cells in the heatmap that have distinct patterns and are likely near the mitosis stage (marked with "*" in the bottom panel of Fig. 4a).

We then asked if the variability of compartment scores across the cell population correlates with transcriptional activity in different cells. We compared the compartment scores with the scRNA-seq from WTC-11 (Friedman et al., 2018). For this analysis, the cells that are likely near the mitosis stage were removed. For each gene, we quantified its compartment variability as the standard deviation of the single-cell compartment scores. The transcriptional variability was calculated using the coefficient of variation (CV) (Supplementary Methods B.4). Based on the quantile of compartment variability, we classified each expressed gene as compartment variable or stable with a cut-off 50%. Compared with the transcriptional variability within these two groups (Fig. 4b), we observed that the genes in more variable compartments have higher transcriptional variability (P-value$<$0.001). We then used the 50Mb window resolution to assess if such structure-function variability correlation can also be observed at a finer scale. We used a 50Mb sliding window with a 1Mb step size on each chromosome and calculated the log difference of the median transcriptional variability between the variable and stable compartment regions within this window. As shown in Fig. 4c, among all windows, 71% of them follow the trend that genes in compartment variable regions have higher transcriptional variability. As a comparison, $\sim$76% of the genomic windows exhibit that the bulk compartment A correlates with higher expression levels (Lieberman-Aiden et al., 2009) (Fig. S8d). In addition, we made a step further to increase the resolution to individual genes. We classified genes as locally variable or stable by identifying the local minima/maxima of the transcriptional variability. We found that for the genes that are locally variable in terms of transcription, their compartment variability scores also tend to be the local maximum (Fig. 4d).

These results further demonstrate the reliability of Higashi imputations, revealing cell-to-cell variability of compartment scores that are also functionally correlated. To confirm the robustness of these observations, in addition to using CV to measure transcriptional variability, we used another metric based a variance stabling algorithm (Supplementary Methods B.4) and reached similar observations (Fig. S8).

Higashi unveils TAD-like domain boundary variation across single cells

Recent work based on multiplexed STORM imaging of chromatin conformation revealed the existence and cell-to-cell variability of TAD-like structures in single cells (Bintu et al., 2018). However, the identification of TAD-like domains remains challenging for sparse scHi-C data with very limited available methods (e.g., Zhou et al. (2019), which only analyzed TAD-like domain differences between major cell types rather than the variability at single-cell resolution). We developed a new approach to analyzing TAD-like domain boundary variability from single cells based on the Higashi imputations (Supplementary Methods B.5 and B.6). The analysis was conducted on the WTC-11 scHi-C dataset at 50Kb resolution.

We calculated single-cell insulation scores in which the local minima correspond to TAD-like domain boundaries (Crane et al., 2015) (Fig. 4e). We again observed a cluster of cells likely near the mitosis stage showing unidentifiable domain boundaries (marked with "*" in the bottom panel of Fig. 4e). We also observed that the local minima of the single-cell insulation scores often center around the domain boundaries observed in the merged imputed scHi-C, while the exact locations of the single-cell boundaries vary across the cell population (Fig. 4e). The dynamics of the single-cell domain boundaries have two main patterns: (1) present/absent across the population (marked with a yellow box in Fig. 4e); and (2) sliding along the genome (marked with an orange box in Fig. 4e). The first pattern reflects that a domain boundary only occurs in a fraction of the cells. The second pattern manifests the shift of domain boundary along the genome, suggesting more gradual cell-to-cell variability. Comparison with scRNA-seq following the same approach used for single-cell compartment scores reached similar conclusions that domain boundary variability is strongly correlated with transcriptional variability at different scales (Fig. S8e-j).

Next, we made direct comparisons of TAD-like domain boundaries by calibrating the single-cell domain boundaries (Supplementary Methods B.6). As shown in Fig. 4f, where each dot corresponds to a single-cell boundary, we observed a negative correlation between the occurrence frequency of a domain boundary with its median single-cell insulation scores. This suggests that the more stable domain boundaries (i.e., higher occurrence frequency) from the cell population tend to be "stronger" boundaries in single cells associated with lower insulation scores (Gong et al., 2018). We also observed positive correlation between the domain boundaries with higher occurrence frequency and the number of CTCF binding peaks as well as the average CTCF peak intensity in the boundaries (Fig. 4g and Supplementary Results A.2). This result is consistent with the observation based on multiplexed STORM imaging (Bintu et al., 2018).

As an iPS cell type, WTC-11 can undergo cell differentiation. We identified differentially expressed genes (DEGs) from a scRNA-seq dataset of WTC-11 cells at 5 differentiation stages (Friedman et al., 2018) (Supplementary Methods B.7) and asked if DEGs would have certain 3D genome variability in WTC-11. Using hypergeometric test, we found that DEGs are over-represented in genes located near more variable domain boundaries (cut-off 50% on the quantile of insulation score std, P-value $\leq 7.9 \times 10^{-8}$) (Fig. 4h). In addition, we compared the variability of insulation scores between DEGs and non-DEGs and found that DEGs have markedly higher standard deviation (t-test, P-value<0.001) (Fig. 4i). This suggests that the cell-to-cell variability of domain boundaries in WTC-11 may have important implication in cell differentiation.

Taken together, by analyzing the TAD-like domain boundaries across single cells with Higashi, we revealed the correlation between domain boundary variability and other functional genomic features. These results further demonstrate the robustness and applicability of Higashi results to facilitate the

analysis of structure-function connections of 3D genome organization in single cells.

Higashi enables analysis of scHi-C data co-assayed with additional epigenomic feature

The emerging new technologies that can jointly profile chromosome conformation and other epigenomic features have provided unique opportunities to directly analyze 3D genome structure and other modalities at single-cell resolution (Lee et al., 2019; Li et al., 2019). However, existing methods developed for scHi-C cannot handle the co-assayed data, which leads to separate processing and analysis of two different modalities. We extended Higashi to incorporate the co-assayed signals into the hypergraph representation learning framework, thereby taking full advantage of the co-assayed data (see Methods). We then applied Higashi to a recently generated co-assayed dataset called single-nucleus methyl-3C (sn-m3C-seq) that jointly profiles Hi-C and DNA methylation in individual human frontal cortex cells (Lee et al., 2019).

We first analyzed the embeddings generated by this joint modeling scheme and made comparisons to the embeddings based on only one modality. We observed that in the UMAP visualization of the Higashi embeddings when only using scHi-C as input (referred to as Higashi(hic)), the non-neuronal brain cell types are separated into individual clusters and the neuronal cell types (L2/3, L4, L5, L6, Pvalb, Sst, Ndnf and Vip) are mixed together (Fig. 5a). Quantitative evaluation shows that the Higashi embeddings from scHi-C outperforms the embeddings generated by scHiCluster for this dataset (Fig. 5d). Meanwhile, the single-cell methylation profiles can resolve the differences within the neuronal cell types, showing better separation between the clusters of neuronal cell types (Fig. 5b) and achieving higher scores in all evaluation metrics (Fig. 5d). This observation is consistent with the findings reported in Lee et al. (2019), where the cell-type specific methylation signatures can be more effectively captured by the existing embedding methods. Importantly, when using Higashi to jointly model both signals (the embeddings referred to as Higashi(joint)), we observed that Higashi reaches the overall best performance with even clearer patterns in the UMAP with cells being aggregated according to their cell types, which is also confirmed by quantitative evaluations (Fig. 5c-d). Note that here the co-asssayed methylation profiles are not part of the input but serve as the targets to approximate (see Methods). In other words, the input to the model only contains scHi-C. Overall, the joint modeling of both scHi-C and methylation profiles enhances the scHi-C embedding by further preserving the chromatin conformation information predictive of DNA methylation profile.

Together, these results demonstrate the ability and advantages of Higashi to effectively analyze co-assayed scHi-C with additional epigenomic feature, capturing intrinsic connections between different modalities at single-cell resolution.

Discussion

In this work, we developed Higashi for multiscale and integrative scHi-C analysis. Our extensive evaluation demonstrated the advantage of Higashi over existing methods for both embedding and imputation. Additionally, we developed new methods in Higashi to analyze variable multiscale 3D genome features (A/B compartment scores and TAD-like domain boundaries) based on the imputed scHi-C contact maps, revealing their implications in gene transcription.

The key algorithmic innovation of Higashi is the transformation of scHi-C data into a hypergraph, which has unique advantages as compared to existing methods. First, this transformation preserves the single-cell precision and 3D genome features from scHi-C. Second, modeling the whole scHi-C

datasets as a hypergraph instead of modeling each contact map as individual graphs allows information to be coordinated across cells to improve both embedding and imputation by taking advantage of the latent correlations between cells. Third, although we mainly focused on scHi-C data, the hypergraph representation in Higashi is highly generalizable to other single-cell data types. As a proof-of-principle, we showed that Higashi can be extended to analyze co-assayed scHi-C data with methylation in an integrated manner, showing markedly improved performance as compared to separate analysis of the two modalities.

There are several directions that Higashi can be further enhanced. As a data-driven method, despite the unique ability of leveraging information from neighboring cells in the embedding space, Higashi requires at least a moderate size scHi-C dataset to achieve high performance. Further optimization in neural network architecture and implementation may also improve scalability. In addition, to achieve more comprehensive delineation of 3D genome organization at single-cell resolution, Higashi can be potentially extended to analyze other types of single-cell assays of higher-order chromatin structures, e.g., the recently developed scSPRITE (Arrastia et al., 2020) that probes multi-way chromatin interactions.

It is anticipated that many exciting new scHi-C technologies including co-assay methods that jointly profile single-cell 3D genome and other epigenetic features will emerge and become widely available in the coming years. Higashi is a robust and powerful new framework that can serve as an essential tool to effectively analyze such data, providing uniquely new insights into nuclear genome structure and function at single-cell resolution in various biological contexts.

Methods

scHi-C data and other genomic data processing

In this work, we used several publicly available single-cell Hi-C datasets. We refer to them as: Ramani et al. (Ramani et al., 2017), Nagano et al. (Nagano et al., 2017), 4DN sci-Hi-C (Kim et al., 2020) (4DN Data Portal: 4DNES4D5MWEZ, 4DNESUE2NSGS, 4DNESIKGI39T, 4DNES1BK1RMQ, 4DNESTVIP977). We also used a new scHi-C dataset generated from the WTC-11 iPSC line (4DN Data Portal: 4DNESF829JOW, 4DNESJQ4RXY5).

For all scHi-C datasets, we only kept the cells with more than 2,000 read pairs that have genomic span greater than 500Kb. At a given resolution, we define the number of contacts per cell as the number of interaction pairs (read count) assigned to the non-diagonal entries of the intrachromosomal contact maps. The Ramani et al. dataset and the 4DN sci-Hi-C dataset used single-cell combinatorial indexed Hi-C (sci-Hi-C). After filtering, the Ramani et al. dataset contains 620 cells of four human cell types (GM12878, HAP1, HeLa, K562) with 7,800 median contacts per cell, while the 4DN sci-Hi-C dataset contains 6,388 cells of five human cell types (GM12878, H1ESC, HAP1, HFFc6, IMR90) with 3,800 median contacts per cell. The Nagano et al. dataset used a different protocol with 1,171 cells and 56,800 median contacts per cell. The WTC-11 scHi-C data (188 cells in total) was generated using single-nucleus Hi-C with 144,800 median contacts per cell. The co-assayed single-cell methylation and Hi-C dataset (sn-m3C-seq) was from Lee et al. (2019). We followed the same processing pipeline as sn-m3C-seq for processing the methylation signals.

We also used other publicly accessible genomic datasets in this work. The scRNA-seq of WTC-11 was from Friedman et al. (2018). The details on calculating transcriptional variability based on scRNA-seq can be found in Supplementary Methods B.4. We also analyzed the CTCF binding near the identified single-cell TAD-like domain boundaries in WTC-11 cells. We used the WTC-11 CTCF ChIA-PET data (4DN Data portal: 4DNES8MZ76GP) and called peaks based on the singleton reads from the dataset following the ENCODE ChIP-seq peak calling pipeline (Moore et al., 2020). Specifically, peaks were generated for individual replicates and merged by keeping only the reproducible peaks.

Overall structure of the hypergraph neural network in Higashi

A *hypergraph* G is a generalization of graph and can be formally defined as a pair of sets $G = (V, E)$, where $V = \{v_i\}$ represents the set of nodes in the graph, and $E = \{e_i = (v_1^{(i)}, ..., v_k^{(i)})\}$ represents the set of hyperedges. For any hyperedge $e \in E$, it connects two or more nodes ($|e| \geq 2$). Both nodes or hyperedges can have attributes reflecting the associated properties such as node type or the strength of a hyperedge. The *hyperedge prediction problem* aims to learn a function f that can predict the probability of a group of nodes $(v_1, v_2, ..., v_k)$ forming a hyperedge or the attributes associated with the hyperedge. For simplicity, we refer to both cases as predicting the probabilities of forming a hyperedge.

The core part of Higashi is a hypergraph representation learning framework, extending our recently developed Hyper-SAGNN (Zhang et al., 2020), that models higher-order interaction patterns from the hypergraph constructed from the scHi-C data. The overall structure of the hypergraph neural network is illustrated in Fig. 1b. We use x_i to represent the attributes of node v_i. The input to the model is a triplet, i.e., (x_1, x_2, x_3) consisting of attributes of one cell node and two genomic bin nodes. We do not differentiate between these two types of nodes in this section for simplicity. Each node within a

triplet passes through a neural network (NN), respectively, to produce (s_1, s_2, s_3), where $s_i = \text{NN}_1(x_i)$. The structure of NN_1 used in this work is a position-wise feed-forward neural network with one fully connected layer. By definition, each s_i remains the same for node v_i independent to the given triplet, and is thus called the "static embedding", reflecting the general topological properties of a node in the given hypergraph. In addition, the triplet as a whole also passes through another transformation, leading to a new set of vectors (d_1, d_2, d_3), where $d_i = \text{NN}_2(x_i \mid (x_1, x_2, x_3))$. The structure of NN_2 will be discussed later. The definition of d_i depends on all the node features within this triplet that reflects the specific properties of a node v_i in a particular hyperedge and is thus called the "dynamic embedding".

Next, the model utilizes the difference between the static and dynamic embeddings to produce $\hat{y}_i$. Additional features, including the genomic distance between the bin pair, one hot encoded chromosome ID, and also the total read number per cell, are concatenated and sent to a multi-layer perceptron with output $\hat{y}_{\text{ext}}$. All the output $\hat{y}_i$ and $\hat{y}_{\text{ext}}$ are further aggregated to produce the final result $\hat{y}$, i.e., the predicted probability for this triplet to be a hyperedge:

$$\hat{y} = \hat{y}_{\text{ext}} + \sum_{i=1}^{3} \hat{y}_i = \hat{y}_{\text{ext}} + \sum_{i=1}^{3} \text{FC}\left[(\vec{d_i} - \vec{s_i})^{\circ 2}\right] \tag{1}$$

where FC is the fully-connected layer.

In the following sections, we describe how the node attributes are generated, the structure of NN_2, the model training, and how we incorporate co-assayed signals into Higashi.

Node attribute generation in Higashi

As mentioned, the input to the hypergraph neural network model is a triplet consisting of attributes of one cell node and two genomic bin nodes. For the bin nodes, we use the corresponding rows of the merged scHi-C contact maps as the attributes. For the cell nodes, we calculate a feature vector based on its single-cell contact maps as its attributes. This process is as follows:

1. Each contact map is transformed into log-scale and then normalized based on the total read count.
2. (optional) Contact maps are smoothed via linear convolution to reach similar sparsity.
3. Contact maps are flattened into 1D vectors and concatenated across the cell population.
4. (optional) Quantile normalization is applied to the matrix.
5. (optional SVD (Singular value decomposition) is used to reduce dimensions for computational efficiency.
6. The corresponding row in the feature matrix is used as the attributes for the corresponding cell.

For all the analysis except the co-assayed scHi-C, we did not apply steps 2 and 4 for computational efficiency. These two steps are useful when the sequencing depths vary significantly across batches.

Cell dependent graph neural network with self-attention for dynamic embeddings

Here, we introduce the neural network NN_2 (mentioned above) that transforms the attributes of a node given a node triplet to the corresponding dynamic embeddings. In the original Hyper-SAGNN, this was accomplished by the multi-head self-attention layer (Vaswani et al., 2017) (see Zhang et al. (2020)). However, the representation capacity of using self-attention layers to calculate dynamic embeddings is

constrained by the embedding dimensions and the depth of self-attention layers, which would lead to high computational cost and increased training difficulty.

To increase the expressiveness of this neural network for generating dynamic embeddings while maintaining small embedding dimensions and fewer layers, we developed a cell dependent graph neural network (GNN) (Hamilton et al., 2017) that transforms the attributes of bin nodes before passing to the self-attention layer. For a node triplet (c_i, b_j, b_k), where c_i corresponds to a cell node and b_j, b_k are bin nodes, a graph $G(c_i)$ (where both b_j, b_k are nodes in it) is constructed by taking c_i as input. Details on the construction of $G(c_i)$, which is shared for all triplets that contain the cell node c_i, is discussed in the next section. For each layer in the GNN, to generate the output vector for bin node b_j, the information of its neighbors in the graph $\boldsymbol{\mathcal{N}}_{G(c_i)}(b_j)$ is aggregated:

$$H^{(n)}_{\boldsymbol{\mathcal{N}}_{G(c_i)}(b_j)} = \text{Average}\left(\{H^{(n-1)}_u e(u, b_j|c_i), u \sim \boldsymbol{\mathcal{N}}_{G(c_i)}(b_j), u \neq b_k\}\right) \tag{2}$$

$$H^{(n)}_{b_j} = \sigma\left\{W^{(n)}_{\text{GNN}} \cdot \text{Concat}\left[H^{(n-1)}_{b_j}, H^{(n)}_{\boldsymbol{\mathcal{N}}_{G(c_i)}(b_j)}\right]\right\} \tag{3}$$

where $H^{(n)}_{b_j}$ is the output vector of the node v_i at the n-th layer of the GNN and $H^{(0)}_{b_j}$ is the attributes of the node b_j before passing to the GNN. $W^{(n)}_{\text{GNN}}$ represents the weight matrix to be optimized at the n-th layer and σ is the non-linear activation function. We call this GNN cell dependent because the structure of the graph depends on the cell, although the weight matrix $W^{(n)}_{\text{GNN}}$ is shared across all cells. This cell dependent GNN can improve the expressiveness of the neural network by incorporating a large amount of single-cell information (contact maps) into the structure of the model instead of entirely relying on the embeddings of the cell nodes. The GNN is trained to reconstruct the interaction between a pair of bin nodes by using only information of themselves and their neighborhood (but not including each other). The attributes of both b_j and b_k are transformed by this cell dependent GNN into $\hat{b}_j$ and $\hat{b}_k$, respectively, and the triplet of $(c_i, \hat{b}_j, \hat{b}_k)$ passes through the aforementioned self-attention layer to generate the final dynamic embeddings.

Information sharing among cells with similar genome structures

Higashi has a unique capability for cells to share information with each other in the embedding space to enhance imputation by taking advantage of the latent correlations between cells. Specifically, we first train Higashi until convergence without the cell dependent GNN to allow the self-attention layer to capture cell-type specific information and reflect in the embeddings through back-propagation. We then calculate the pairwise distances of cell embeddings that indicate the similarities among cells. Given a hyperparameter k, we construct a graph $G(c_i)$ based on the contact maps of c_i and its k-nearest neighbors in the embedding space. It is crucial to clarify that, when we mention the neighbor of a cell $\boldsymbol{\mathcal{N}}(c_i)$, we are referring to other cells that have small pairwise distances of embedding vectors instead of other nodes that have connections to the cell in the hypergraph. We name the contact maps of c_i as $M(c_i)$. The new $G(c_i)$ is constructed as the weighted sum of $M(u), u \in \{c_i\} \cup \boldsymbol{\mathcal{N}}(c_i)$, where the weight is calculated based on the pairwise distance $d(u, c_i)$ in the embedding space, i.e.,

$$G(c_i) \sim \sum_u w(u, c_i) M(u), \quad u \in \{c_i\} \cup \boldsymbol{\mathcal{N}}(c_i) \tag{4}$$

$$w(u, c_i) \propto \exp\left[-d(u, c_i)\right] \tag{5}$$

Each embedding is normalized by the maximum ℓ^2 norm. Note that although contact maps of different cells are mixed in this step, we do not mix the prediction results from different cells or directly use the mixed contact maps as output. This differentiates our method from the k-NN based smoothing methods fundamentally. The Higashi model is trained with only the observed interactions in each single cell, together with the interactions in cells that share overall similar structures serving as auxiliary information for synergistic prediction in a cell population.

Model training details in Higashi

The hypergraph neural network within Higashi would produce a score $\hat{y}$ for any triplet (c_i, b_j, b_k). The neural network is trained to minimize the difference between the predicted score $\hat{y}$ and the ground truth score y, indicating the probability of the pairwise interaction between bin nodes b_j and b_k in cell c_i. In Higashi, we offer two choices of loss function for scHi-C datasets with different coverage. For scHi-C datasets with relatively low sequencing depths or the analysis resolution is high (hence fewer reads in each genomic bin), the model is trained with a binary classification loss (cross entropy) where the triplets corresponding to all non-zero entries in the single-cell contact maps are treated as positive samples and the rest are considered as the negative samples (i.e., $y(c_i, b_j, b_k) \in \{0, 1\}$). The classification loss is:

$$\text{Loss}_{\text{class}} = -\sum_{i,j,k} y(c_i, b_j, b_k) \log \hat{y}(c_i, b_j, b_k) + [1 - y(c_i, b_j, b_k)] \log [1 - \hat{y}(c_i, b_j, b_k)] \tag{6}$$

For datasets with relatively high sequencing depths or when the analysis resolution is low (hence more reads in each genomic bin), we further differentiate among the non-zero values by training the model with a ranking loss, which maintains consistent ranking of predicted scores versus the continuous ground truth scores (i.e., $y(c_i, b_j, b_k) \in \mathbb{R}$). The ranking loss can be described as a binary classification problem aiming to identify the triplet with larger ground truth score in a pair of selected triplets. For simplicity, we denote two triplets as t_i, t_j and the corresponding ground truth scores as $y(t_i), y(t_j)$. The ranking loss is:

$$l_{ij} = \mathbb{I}\left[y(t_i) > y(t_j)\right] \tag{7}$$

$$p_{ij} = \text{Sigmoid}\left[\hat{y}(t_i) - \hat{y}(t_j)\right] \tag{8}$$

$$\text{Loss}_{\text{rank}} = -\sum_{|y(t_i)-y(t_j)|\geq\alpha} l_{ij} \log p_{ij} + (1 - l_{ij}) \log (1 - p_{ij}) \tag{9}$$

where α defines whether the order of $y(t_i), y(t_j)$ can be reliably called and is set to 2 in this work. Note that l_{ij}, p_{ij} are intermediate variables only used in this definition.

Using either the classification loss or the ranking loss requires negative samples in the training data. We designed an effective negative sampling approach. Specifically, at each epoch, we randomly sample a batch of triplets and make sure that these triplets do not overlap with the positive samples. To reflect the similarity of 3D genome structures of flanking genomic bins, we also exclude triplet (c_i, b_j, b_k) from the negative samples if triplets such as $(c_i, b_j + 1, b_k)$ belong to the positive samples. The number of negative samples generated for each batch is guided by the sparsity of the input data. When studying a scHi-C data where $s\%$ of the contact map entries are zeros, for a batch of n positive triplets, $\min\left[s/(100 - s), 5\right] n$ negative samples will be generated. The number of negative samples is no more than 5 times of the number of positive samples for computational efficiency. The model is optimized by the Adam algo-

rithm (Kingma and Ba, 2014) with learning rate of 1e-3. We train Higashi for 80 epochs with batch size of 192 or until the model converges on an individual validation set before 80 epochs.

Incorporating co-assayed signals in Higashi

The unique design of Higashi allows joint modeling of co-assayed scHi-C and the corresponding one dimensional signals (e.g., from sn-m3C-seq (Lee et al., 2019)). We add an auxiliary task for Higashi by using the learned embeddings for cell nodes c_i to accurately reconstruct the co-assayed signals m_i through a multi-layer perceptron (MLP). The auxiliary loss term is added to the main loss function and optimized jointly. The model thus builds an integrated connection between chromatin conformation and the co-assayed signals, guiding the embedding of the scHi-C data, i.e.,

$$\text{Loss}_{\text{auxi}} = \text{MSE}\left[m_i, \text{MLP}(c_i)\right] \tag{10}$$

where MSE refers to the mean squared error between the ground truth of co-assayed signals and the estimate.

Methods for the analysis of variable compartment and domain boundary from scHi-C

In Higashi, we have developed new strategies for reliable analysis of 3D genome features in different scales across the cell population. We developed a method to calculate continuous compartment scores for the imputed single-cell contact maps such that these scores are directly comparable across different cells in the population to assess variability (see Supplementary Methods B.3). For single-cell TAD-like domain boundary analysis, we developed a new calibration method with a novel optimization scheme to achieve comparative analysis of domain boundary variability from single cells based on insulation scores (see Supplementary Methods B.5 and B.6). These algorithms greatly enhance the analysis of variable multiscale 3D genome structures at single-cell resolution.

Code Availability

The source code of Higashi can be accessed at: https://github.com/ma-compbio/Higashi.

Acknowledgements

The authors would like to thank Bing Ren for sharing the WTC-11 single-cell Hi-C data before publication, Yijun Ruan for making the WTC-11 CTCF ChIA-PET data available, and Yang Zhang for suggestions that improved the manuscript. The authors are also grateful to Bing Ren, Job Dekker, William Noble, Zhijun Duan, and other members of the NOFIC-AICS Collaborative Project Working Group of the NIH 4DN Consortium for discussions and feedback. This work was supported in part by the National Institutes of Health Common Fund 4D Nucleome Program grants U54DK107965 (J.M.) and UM1HG011593 (J.M.) and National Institutes of Health grant R01HG007352 (J.M.). J.M. is additionally supported by a Guggenheim Fellowship from the John Simon Guggenheim Memorial Foundation.

Author Contributions

Conceptualization, R.Z. and J.M.; Methodology, R.Z. and J.M.; Software, R.Z.; Investigation, R.Z., T.Z., and J.M.; Writing – Original Draft, R.Z. and J.M.; Writing – Review & Editing, R.Z., T.Z., and J.M.; Funding Acquisition, J.M.

Competing Interests

The authors declare no competing interests.

References

M. V. Arrastia, J. W. Jachowicz, N. Ollikainen, M. S. Curtis, C. Lai, S. A. Quinodoz, D. A. Selck, M. Guttman, and R. F. Ismagilov. A single-cell method to map higher-order 3D genome organization in thousands of individual cells reveals structural heterogeneity in mouse ES cells. *bioRxiv*, 2020.

B. Bintu, L. J. Mateo, J.-H. Su, N. A. Sinnott-Armstrong, M. Parker, S. Kinrot, K. Yamaya, A. N. Boettiger, and X. Zhuang. Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells. *Science*, 362(6413), 2018.

E. Crane, Q. Bian, R. P. McCord, B. R. Lajoie, B. S. Wheeler, E. J. Ralston, S. Uzawa, J. Dekker, and B. J. Meyer. Condensin-driven remodelling of X chromosome topology during dosage compensation. *Nature*, 523(7559):240–244, 2015.

J. Dekker, A. S. Belmont, M. Guttman, V. O. Leshyk, J. T. Lis, S. Lomvardas, L. A. Mirny, C. C. O'shea, P. J. Park, B. Ren, et al. The 4D nucleome project. *Nature*, 549(7671):219–226, 2017.

J. R. Dixon, S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. S. Liu, and B. Ren. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398): 376, 2012.

I. M. Flyamer, J. Gassler, M. Imakaev, H. B. Brandão, S. V. Ulianov, N. Abdennur, S. V. Razin, L. A. Mirny, and K. Tachibana-Konwalski. Single-nucleus Hi-C reveals unique chromatin reorganization at oocyte-to-zygote transition. *Nature*, 544(7648):110–114, 2017.

C. E. Friedman, Q. Nguyen, S. W. Lukowski, A. Helfer, H. S. Chiu, J. Miklas, S. Levy, S. Suo, J.-D. J. Han, P. Osteil, et al. Single-cell transcriptomic analysis of cardiac differentiation from human PSCs reveals HOPX-dependent cardiomyocyte maturation. *Cell Stem Cell*, 23(4):586–598, 2018.

Y. Gong, C. Lazaris, T. Sakellaropoulos, A. Lozano, P. Kambadur, P. Ntziachristos, I. Aifantis, and A. Tsirigos. Stratification of TAD boundaries reveals preferential insulation of super-enhancers by strong boundaries. *Nature Communications*, 9(1):1–12, 2018.

W. Hamilton, Z. Ying, and J. Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, pages 1024–1034, 2017.

R. Kempfer and A. Pombo. Methods for mapping 3D chromosome architecture. *Nature Reviews Genetics*, 21(4):207–226, 2020.

H.-J. Kim, G. G. Yardımcı, G. Bonora, V. Ramani, J. Liu, R. Qiu, C. Lee, J. Hesson, C. B. Ware, J. Shendure, et al. Capturing cell type-specific chromatin compartment patterns by applying topic modeling to single-cell Hi-C data. *PLoS Computational Biology*, 16(9):e1008173, 2020.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

D.-S. Lee, C. Luo, J. Zhou, S. Chandran, A. Rivkin, A. Bartlett, J. R. Nery, C. Fitzpatrick, C. O'Connor, J. R. Dixon, et al. Simultaneous profiling of 3D genome structure and DNA methylation in single human cells. *Nature Methods*, pages 1–8, 2019.

G. Li, Y. Liu, Y. Zhang, N. Kubo, M. Yu, R. Fang, M. Kellis, and B. Ren. Joint profiling of DNA methylation and chromatin architecture in single cells. *Nature Methods*, 16(10):991–993, 2019.

E. Lieberman-Aiden, N. L. Van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–293, 2009.

J. Liu, D. Lin, G. G. Yardımcı, and W. S. Noble. Unsupervised embedding of single-cell Hi-C data.

Bioinformatics, 34(13):i96–i104, 2018.

C. Marchal, J. Sima, and D. M. Gilbert. Control of DNA replication timing in the 3D genome. *Nature Reviews Molecular Cell Biology*, 20(12):721–737, 2019.

L. McInnes, J. Healy, and J. Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

T. Misteli. The self-organizing genome: Principles of genome architecture and function. *Cell*, (1):28–45, 2020.

J. E. Moore, M. J. Purcaro, H. E. Pratt, C. B. Epstein, N. Shoresh, J. Adrian, T. Kawli, C. A. Davis, A. Dobin, R. Kaul, et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature*, 583(7818):699–710, 2020.

T. Nagano, Y. Lubling, T. J. Stevens, S. Schoenfelder, E. Yaffe, W. Dean, E. D. Laue, A. Tanay, and P. Fraser. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature*, 502(7469): 59–64, 2013.

T. Nagano, Y. Lubling, C. Várnai, C. Dudley, W. Leung, Y. Baran, N. M. Cohen, S. Wingett, P. Fraser, and A. Tanay. Cell-cycle dynamics of chromosomal organization at single-cell resolution. *Nature*, 547 (7661):61, 2017.

E. P. Nora, B. R. Lajoie, E. G. Schulz, L. Giorgetti, I. Okamoto, N. Servant, T. Piolot, N. L. van Berkum, J. Meisig, J. Sedat, et al. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*, 485(7398):381, 2012.

V. Ramani, X. Deng, R. Qiu, K. L. Gunderson, F. J. Steemers, C. M. Disteche, W. S. Noble, Z. Duan, and J. Shendure. Massively multiplex single-cell hi-c. *Nature Methods*, 14(3):263, 2017.

S. S. Rao, M. H. Huntley, N. C. Durand, E. K. Stamenova, I. D. Bochkov, J. T. Robinson, A. L. Sanborn, I. Machol, A. D. Omer, E. S. Lander, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159(7):1665–1680, 2014.

T. J. Stevens, D. Lando, S. Basu, L. P. Atkinson, Y. Cao, S. F. Lee, M. Leeb, K. J. Wohlfahrt, W. Boucher, A. O'Shaughnessy-Kirwan, et al. 3D structures of individual mammalian genomes studied by single-cell Hi-C. *Nature*, 544(7648):59–64, 2017.

J.-H. Su, P. Zheng, S. S. Kinrot, B. Bintu, and X. Zhuang. Genome-scale imaging of the 3D organization and transcriptional activity of chromatin. *Cell*, 2020.

L. Tan, D. Xing, C.-H. Chang, H. Li, and X. S. Xie. Three-dimensional genome structures of single diploid human cells. *Science*, 361(6405):924–928, 2018.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30:5998–6008, 2017.

Y. Wang, Y. Zhang, R. Zhang, T. van Schaik, L. Zhang, T. Sasaki, D. Peric-Hupkes, Y. Chen, D. M. Gilbert, B. van Steensel, et al. SPIN reveals genome-wide landscape of nuclear compartmentalization. *bioRxiv*, 2020.

K. Xiong and J. Ma. Revealing Hi-C subcompartments by imputing inter-chromosomal chromatin interactions. *Nature Communications*, 10, 2019.

R. Zhang, Y. Zou, and J. Ma. Hyper-SAGNN: a self-attention based graph neural network for hypergraphs. In *International Conference on Learning Representations (ICLR)*, 2020.

J. Zhou, J. Ma, Y. Chen, C. Cheng, B. Bao, J. Peng, T. J. Sejnowski, J. R. Dixon, and J. R. Ecker. Robust single-cell hi-c clustering by convolution-and random-walk–based imputation. *Proceedings of the National Academy of Sciences*, page 201901423, 2019.

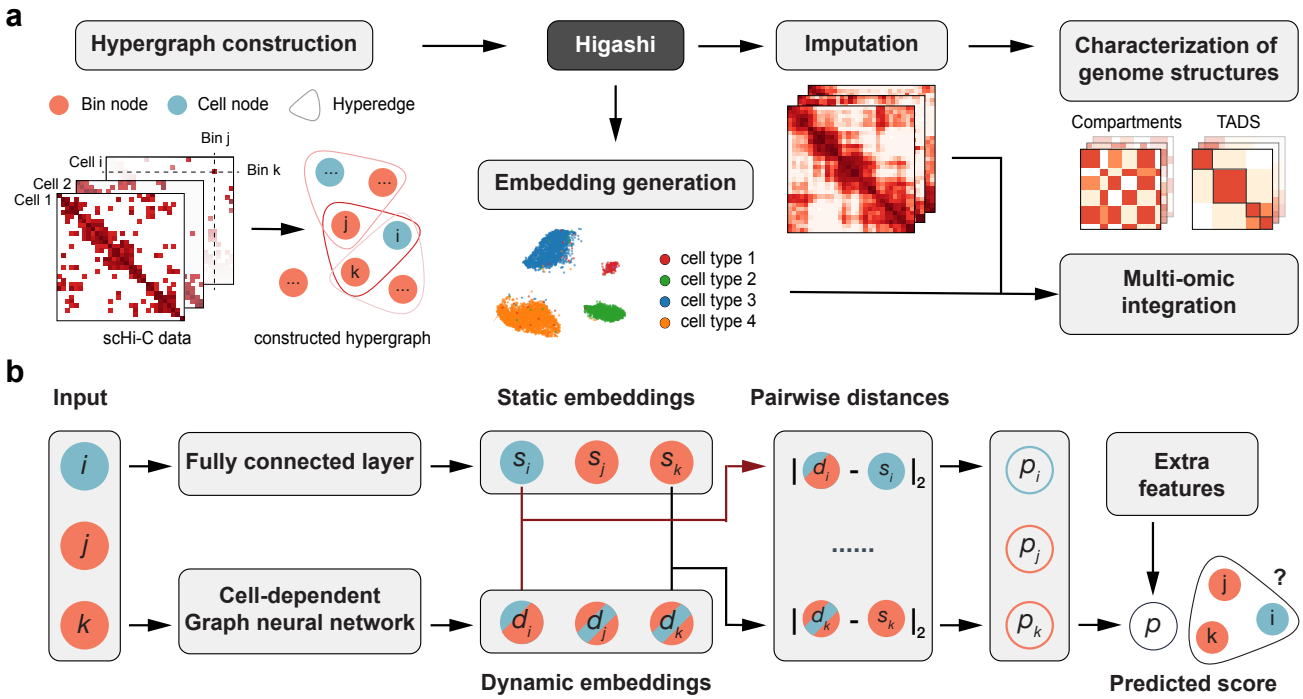

Figure 1: Overall design of Higashi. **a.** Overview of the Higashi framework for scHi-C analysis. The input scHi-C dataset is transformed into a hypergraph where each hyperedge connects one cell node and two bin nodes. A hypergraph neural network is trained to capture high-order interaction patterns within the constructed hypergraph. The trained neural network is able to generate embeddings for scHi-C data as well as impute the sparse scHi-C contact maps. The imputed contact maps and the embeddings allow detailed characterization of multiscale 3D genome features and also multiomic integrative analysis. **b.** The structure of the hypergraph neural network used in Higashi. The input triplet consisting of one cell node and two bin nodes passes through two branches of the network to generate static embeddings and dynamic embeddings for each node, respectively. Then the pairwise distances between static and dynamic embedding pairs are calculated. These pairwise distances are combined with extra features such as genomic distance between the two bins to produce the final predicted score for the input triplet, which represents the probability of an entry in the single-cell contact map.

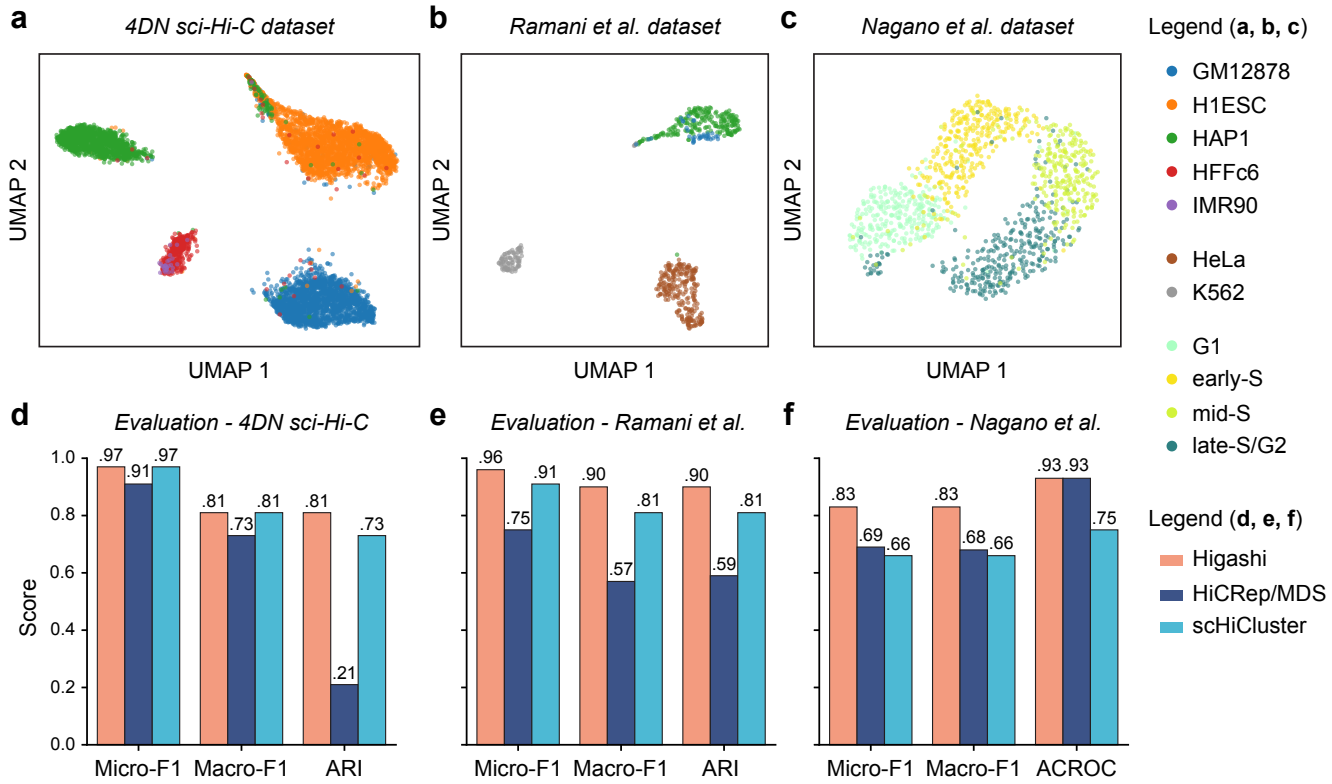

Figure 2: Evaluation of the embeddings generated by Higashi. **a, b, c.** The UMAP visualization of the embeddings produced by Higashi for three scHi-C datasets (4DN sci-Hi-C (Kim et al., 2020), Ramani et al. (Ramani et al., 2017), and Nagano et al. (Nagano et al., 2017) datasets). **(d, e, f).** Quantitative evaluation of Higashi on the three scHi-C datasets by comparing to HiCRep/MDS (Liu et al., 2018) and scHiCluster (Zhou et al., 2019). The performances are measured by Micro-F1, Macro-F1, and Adjusted Rand Index (ARI) scores from the supervised and unsupervised cell type identification tasks.

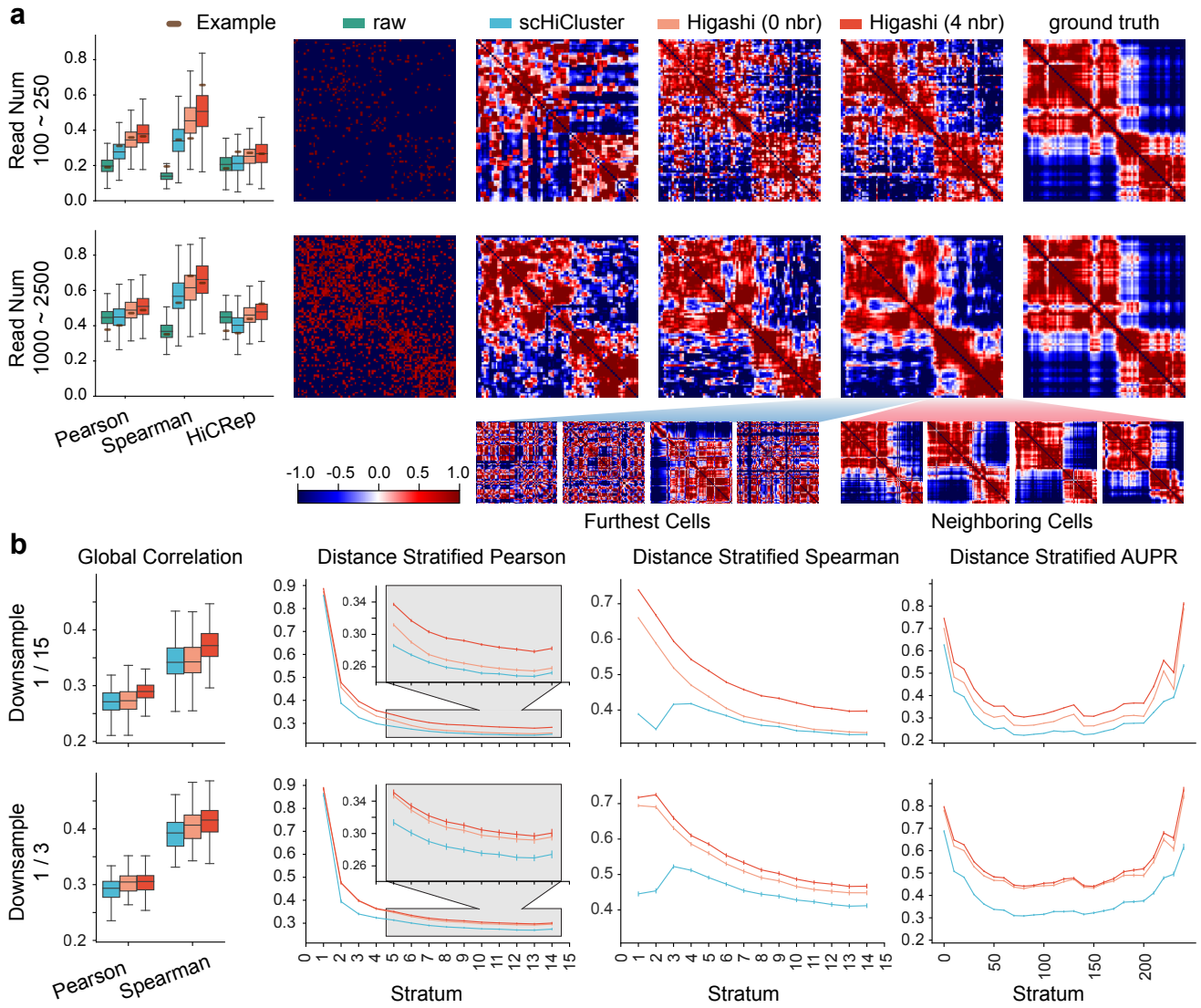

Figure 3: Evaluation of the data imputation produced by Higashi. **a.** Evaluation and visualization of different imputation methods on scHi-C data simulated from multiplexed STORM 3D genome imaging data (Bintu et al., 2018). For Higashi, results by using information from four neighboring cells (4 nbr) or without using neighboring cell information (0 nbr) in the embedding space are both included. Each row corresponds to one set of simulation data with a chosen range of read numbers. The boxplots illustrate the quantitative evaluation of the similarities between the raw (input), the scHiCluster enhanced, and the Higashi enhanced contact maps versus the ground truth (inverse distance map). The heatmaps visualize the contact map before and after imputation as well as the ground truth. The contact maps of both the neighboring cells (in the embedding space) that contribute to the imputation and the cells that are furthest (in the embedding space) are shown. See also Fig. S3. **b.** Evaluation of different imputation methods on scHi-C data simulated based on downsampling the WTC-11 scHi-C data. Each row corresponds to a specific downsampling rate. The boxplots illustrate the global correlation scores of the contact maps before and after imputation versus the ground truth. The line plots visualize the evaluation metrics over different strata of the matrix. The evaluation metrics include the Pearson correlation and the Spearman correlation for short-range interactions as well as the AUPR (area under precision-recall) score for long-range interactions. See also Fig. S5.

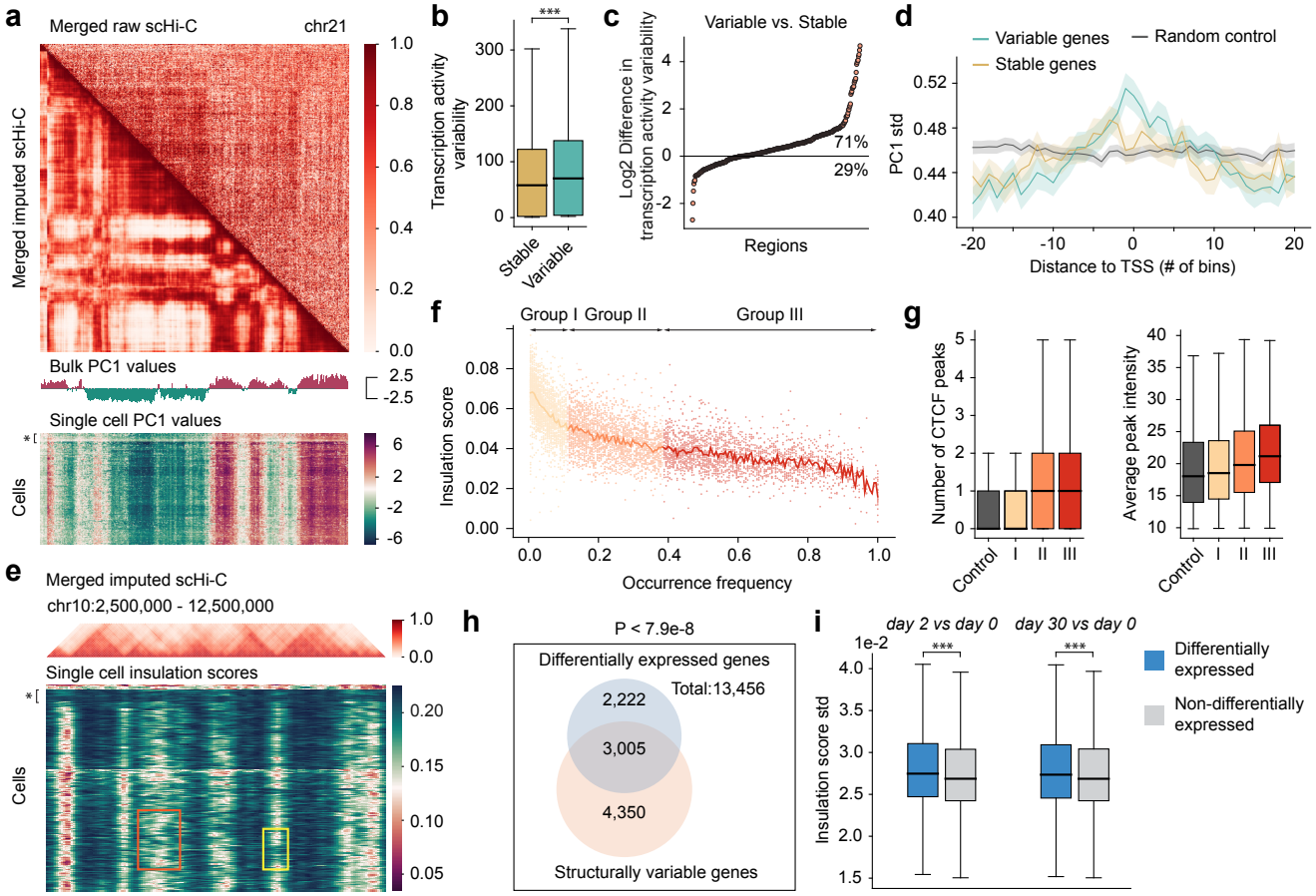

Figure 4: Higashi enables detailed characterization of 3D genome features and their connections to gene transcription at single-cell resolution. **a.** Compartment score annotations for WTC-11 scHi-C data at 50Kb resolution. The merged scHi-C correlation matrix on chr21 before and after imputation as well as compartment scores called from the pooled WTC-11 scHi-C data and each single-cell contact map are visualized. The cells that are likely near the mitosis stage are marked with "*" in the single-cell PC1 heatmap. **b.** Global comparisons of transcriptional variability on regions with variable and stable compartment annotations (*** indicates P-value$<$1e-3). **c.** Log2 difference of transcriptional variability of genes with variable versus stable compartment annotations within a Mb scale window. **d.** Visualization of standard deviation of compartment scores around genes with variable or stable transcriptional level. In **b,c,d**, the transcriptional variability is quantified as the coefficient of variation of the imputed scRNA-seq data. **e.** TAD-like domain boundary calling for WTC-11 scHi-C at 50Kb resolution. The merged scHi-C contact maps at chr10:2,500,000-12,500,000 and the calculated insulation scores are shown. The cells that are likely near the mitosis stage are marked with "*" in the single-cell insulation score heatmap. Regions that represent the present/absent dynamics of single-cell domain boundaries are marked with a yellow box. Regions that represent the sliding dynamics of single-cell domain boundaries are marked with an orange box. **f.** Scatter plot of the single-cell insulation scores versus the occurrence frequency in the cell population of shared domain boundaries. For each cell, only the insulation scores of presented shared boundaries are visualized, i.e., each dot corresponds to a single-cell domain boundary. **g.** CTCF binding at domain boundaries from different occurrence frequency groups. **h.** Venn diagram of the overlap between genes near variable domain boundary in WTC-11 (light red) and differentially expressed genes during cell differentiation (light blue). **i.** Comparison of cell-to-cell variability of insulation scores between differentially expressed genes and non-differentially expressed genes. The high variance of insulation scores of differentially expressed genes indicates that differentially expressed genes are enriched near domain boundaries with higher variability (*** indicates P-value$<$1e-3).

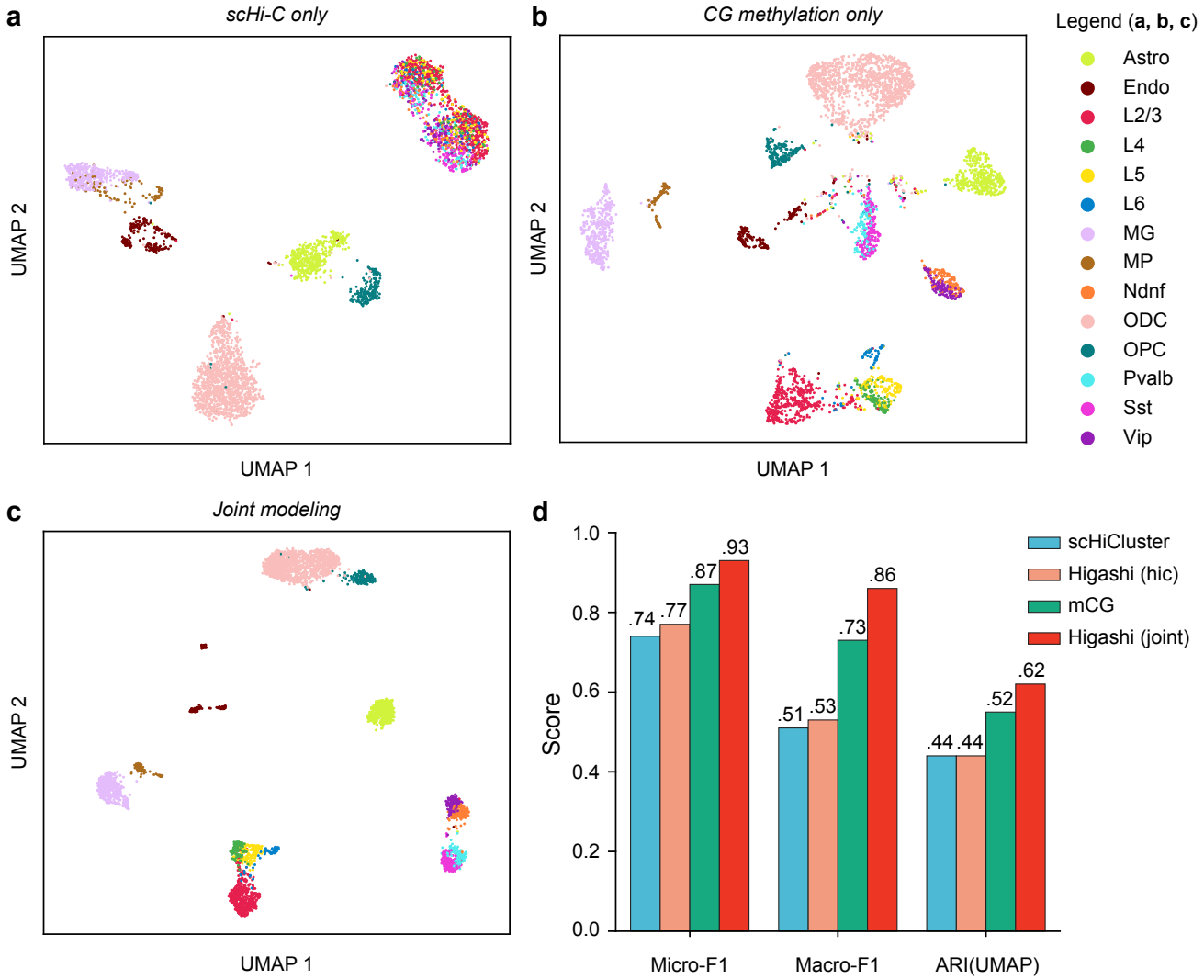

Figure 5: Higashi enables joint modeling of co-assayed scHi-C data with other epigenomic feature. **a, b, c.** UMAP visualization of the human brain prefrontal cortex sn-m3C-seq dataset (Lee et al., 2019) with the Higashi embeddings of chromatin interactions only **(a)**, CG methylation profiles only **(b)**, and the Higashi embeddings of joint modeling of both chromatin conformation and methylation **(c)**. **d.** Quantitative evaluation of different embeddings of the human brain prefrontal cortex sn-m3C-seq dataset. The performances are measured by Micro-F1, Macro-F1, and Adjusted Rand Index (ARI) scores from the supervised and unsupervised cell type identification tasks.