Integrative Machine Learning Models for Spatial Biology

Tianming Zhou

CMU-CB-XX

September 2024

Computational Biology Department
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Jian Ma, Chair
Kathryn Roeder
Ivet Bahar
Fei Chen

*Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy.*

This research is funded in part by XX. The department specifically disclaims responsibility for any analyses, interpretations, or conclusions. The views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of any sponsoring institution, donors or the U.S. Government.

Keywords: 3D genome organization, spatial transcriptome, machine learning

Abstract

Recent advances in single-cell technologies have provided unprecedented opportunities to study genome structure, gene expression, and cellular organization in complex tissues. However, the cell type-specific connections between genome structure and genome function, as well as the molecular mechanisms underlying cellular spatial organization, remain poorly understood. In particular, modeling the spatial patterns of biomolecules and cells and cohesively integrating genome architecture and transcriptome data for single-cell analysis of complex tissues is a major challenge. In this Ph.D. dissertation, I develop a series of new algorithms based on representation learning and probabilistic graphical models for modeling multimodal spatial omics. First, I develop a probabilistic, latent variable modeling framework to model cell identity using single-cell spatial transcriptomic data. Second, I create a tensor decomposition framework to jointly infer cell embeddings and cell type-specific 3D genome features based on single-cell 3D genome mapping data. Third, I introduce an integrative analysis framework for a new single-cell co-assay of 3D genome and transcriptome. Fourth, I develop a transformer-based machine learning model to understand the interplay between DNA sequence, 3D genome structure, and gene expression in a cell type-specific manner. These new machine-learning models are expected to unveil cell-to-cell variability and the spatiotemporal dynamics of 3D genome structures and their connections with gene expression in various biological contexts. Together, the computational methods developed in this dissertation have the potential to shed new light on the spatial organization of the genome and cells and their functional implications in health and disease.

Acknowledgement

Throughout my six-year PhD journey, I had the privilege of sharing experiences and building friendships with many extraordinary people. I am fortunate to have received support and advice along the way. I am grateful to those who helped me reach this point.

I would like to express my gratitude to Professor Jian Ma, my advisor and committee chair. I could not have achieved many Ph.D. milestones without Prof. Ma's support and guidance. I am indebted to Prof. Ma for guiding me to choose the industrial career path with confidence and excitement, and for training me to be independent and proactive, which prepares myself for the next chapters of my life.

I am thankful to Professor Kathryn Roeder, Professor Ivet Bahar, and Professor Fei Chen for serving on my committee and providing feedback on my thesis research.

I am grateful for my former and current lab mates. In particular, I appreciate my coauthors, Ruochi Zhang, Ben Chidester, and Shahul Alam. Additionally, I appreciate Yang Zhang for his helpful advice and unwavering support.

I am fortunate to have friends and fellow students by my side during my Ph.D. and during my 9-month job hunting in the most challenging year, including Ruochi Zhang, Minxue Jia, Yan Zhang, Yifeng Tao, Yuchuan Wang, Muyu Yang, Xinyue Lu, Shike Wang, Shahul Alam, Yihang Shen, Ruogu Lin, Yang Ping Kuo, Yutong Qiu, Dongshunyi Li, Xuecong Fu, Cong Ma, Hongyu Zheng, Haoyun Lei, Xiangrui Zeng, Peiran Jiang, Haotian Teng, Jingjing Tang, Donghui Yan, Xiaoyue Cui, Siping Wang, Endong Zhu, Ruihao Huang, Verna Tian, Yi Zhou.

Lastly, I appreciate the support of my parents and grandparents as well as the company of my partner, Yanhan Tang, and our two cats, Kirby (Rong Rong) and Lucky, through the most challenging years.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background and related work

The cell-type-specific connections between genome structure and function, and their influence on cellular phenotypes remain mostly unclear, especially in complex tissue, such as the brain [1]. In addition, the compositions of different cell types in mammalian tissues, such as the brain, remain poorly understood, due to the complex interplay between the intrinsic multi-scale epigenome and cell-to-cell interaction that collectively contribute to the cell identity [2–4]. The advent of high-throughput whole-genome mapping methods for the three-dimensional (3D) genome organization such as Hi-C [5] has revealed distinct features of chromatin folding in various scales within the cell nucleus, including A/B compartments [5], subcompartments [6, 7], topologically associating domains (TADs) [8, 9], and chromatin loops [6]. These multiscale 3D genome features collectively contribute to vital genome functions such as transcription and DNA replication [10–13]. However, the variation of 3D genome features and their functional significance in single cells remain poorly understood [1, 14]. The recent developments of single-cell Hi-C (scHi-C) technologies have enabled us to probe chromatin interactions at single-cell resolution, from a few cells [15–18] to thousands of cells from complex tissues [19–21]. These new technologies and datasets have the promise to reveal the multi-scale connections between genome structure and function for a wide range of biological contexts [14].

The emerging spatial transcriptomics technologies based on multiplexed imaging and sequencing [22–33] are able to reveal spatial information of gene expression of up to tens of thousands of genes in individual cells *in situ* within the tissue context. These technologies and datasets enabled us to study cell-cell communications mediated by ligand-receptor pairs [27], gap junction [31], and estrogen-based paracrine signaling [26]. At a higher level, spatial transcriptome sheds light on critical biological processes, such as the spatiotemporal developmental trajectory in the brain [29], cancer cell state transitions [34], heterogeneous immune responses of T cells to cancer [35], and amyloid deposition in Alzheimer's Disease [36],

It remains a challenge to develop computational methods for drawing a holistic picture of the intra-cellular multi-scale epigenome and inter-cellular spatial organization of complex tissues [14, 37, 38]. Current methods for the analysis of single-cell Hi-C datasets, such as HiCRep/MDS [39], scHiCluster [40], LDA [17], and [41], have limited ability to infer informative and interpretable embedding spaces that can delineate rare cell types in complex tissues, due to the high dimensionality and high sparsity of scHi-C data. These existing methods also cannot directly reveal 3D genome structures related to cell-type-specific genome functions, or be scaled up to large-scale datasets with limited computational resources. Computational methods have been developed to use spatial transcriptome data to identify spatial domains and cell types in tissues [42–46], to explore the spatial variance of genes [47–50], to align scRNA-seq with spatial transcriptome data [51–54], and to model the inter-cellular spatial dependencies [38, 43, 44, 46]. The existing methods typically do not integrate the modeling of the spatial variability of genes with their contribution to cell identity, and the learned cell embeddings are generally hard to interpret. Therefore, there is an urgent need for robust, interpretable methods that can jointly model both the spatial organization and multi-scale intra-cellular epigenome for complex tissues, which is of vital importance to draw a holistic picture of living organisms and to shed light on health and disease.

From an algorithmic view, the cell type-specific relation between genome structure and function in complex tissue can be modeled by integration of multilinear models and probabilistic graphical models. The latent representation of single cells and important patterns in genome structure and function can be jointly modeled can inferred. The latent representation of single cells and the spatially variable features can be jointly captured by model inference. The overall goal of my Ph.D. thesis is focused on addressing these challenges by developing machine learning models for the study of multiple modalities at single-cell resolution, even in the spatial context when available. Ultimately, we aim to provide novel insights into the cell type-specific relation between genome structure and function in complex tissue.

1.2 Structure of the thesis

This thesis begins with an introduction to the background (Chapter 1), then presents our contributions to addressing critical challenges in this field (Chapter 2, 3, 4, and 5), and finally summarizes our findings and future directions (Chapter 6). Fig. 1.1 illustrates the structure of the thesis and the development and biological applications of computational methods.

In Chapter 2, we develop a new latent representation learning method for the analysis of spatial transcriptomic data that integrates gene expression and spatial patterns of cells in complex tissues. Our method aims to reveal refined cell types, spatially variable metagenes, and spatial patterns of biological processes. This chapter is based on the work:

- Benjamin Chidester, Tianming Zhou, Shahul Alam, and Jian Ma. "SPICEMIX enables integrative single-cell spatial modeling of cell identity." *Nature Genetics* (2023) [55].

In Chapter 3, we developed a new method for the analysis of single-cell Hi-C data based on tensor decomposition. Our model aims to jointly infer cell embeddings and 3D genome features that capture critical structural features of single-cell chromatin that are related to genome functions. This chapter is based on the work:

- Ruochi Zhang, Tianming Zhou, and Jian Ma. "Ultrafast and interpretable single-cell 3D genome analysis with Fast-Higashi." *Cell Systems* (2022) [56].

In Chapter 4, we collaborated with Zhijun Duan from the University of Washington and developed GAGE-seq for concurrent profiling of scHi-C and scRNA-seq. We also developed generic computational algorithms for integrating co-assayed single-cell technologies with one shared modality. Our algorithm enabled the joint analysis of transcriptome, DNA methylation, and 3D genome structure in mouse brain. Our algorithm also overlaid single-cell 3D genome structure and the full transcriptome onto the spatial contexts from mouse brains, revealing the intricate spatial patterns of transcriptome and 3D genome structure. This chapter is based on the following work and the experimental details can be referred to the published paper:

- Tianming Zhou, Ruochi Zhang, Deyong Jia, Raymond T. Doty, Adam D. Munday, Daniel Gao, Li Xin, Janis L. Abkowitz, Zhijun Duan, and Jian Ma. "Concurrent profiling of multiscale 3D genome organization and gene expression in single mammalian cells." *Nature Genetics* (2024) [57].

In Chapter 5, we developed a transformer-based predictive model, called Hi-CFormer, for understanding the intricate interplay of DNA sequence, 3D genome structure, and transcriptome. Hi-CFormer predicts mRNA signals from DNA sequence and cell-type-specific 3D genome structure. We developed novel variants of the transformer layer and the attention layer in order to effectively learn information from 2D Hi-C contact maps. Evaluation on a GAGE-seq dataset from mouse brain data [57] demonstrated the superior performance of Hi-CFormer compared to sequence-only baselines. The interpretation of trained Hi-CFormer revealed cell-type-specific interplay between DNA sequence, 3D genome structure, and transcriptome.

In the final chapter, we revisit these algorithms and the GAGE-seq technology and discuss how they collectively deepen our understanding of the complex interplay between 3D genome structure, transcriptome, and spatial context. We also create blueprints for the future of these directions.

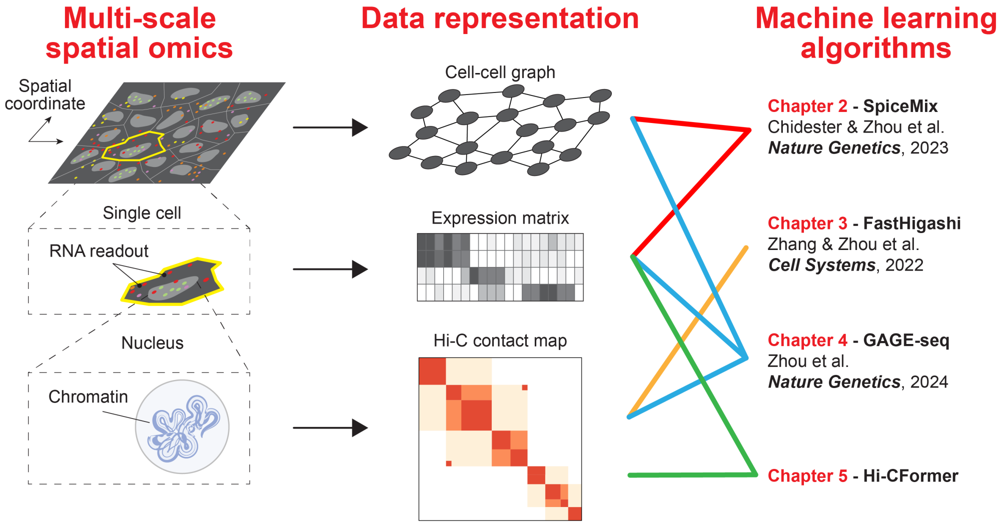

Figure 1.1: Overview and structure of this thesis. The left column illustrates the multi-scale structures in complex tissues, including the spatial arrangement of cells, cellular structures, and chromatin organizations. The middle column shows the data representation used in this thesis. On the right are the algorithms with the association to the aspects of the multi-scale structures.

1.3 Introduction to relevant biological technologies

1.3.1 scHi-C technology

New technologies have been developed to study 3D chromatin organization at single-cell resolution in recent years [15, 16, 58–70]. In sequencing-based methods, interacting genomic loci are labeled in an unbiased manner, and their identities are then obtained by sequencing. Most sequencing-based methods utilize the pairwise proximity ligation to capture pairwise interactions between genomic loci, except scSPRITE [63] which tags loci within one complex with the same barcode and thus is able to capture multi-way interactions directly. Note that the sequencing-based methods cannot capture the exact spatial positions of genomic loci, requiring computational tools to reconstruct 3D conformation from contact maps. An exception is the IGS method [70], where a UMI (unique molecular identifier) is tagged to each locus and sequenced *in situ*, allowing the identification of spatial positions of genomic loci. In imaging-based methods, spatial positions of genomic loci are visualized by fluorescent *in situ* hybridization (FISH). However, the imaging technologies are typically restricted to pre-selected loci. As a result, the resolution can be as high as 5kb, but the coverage remains relatively low because only around a thousand loci can be detected under the current barcoding schemes.

Computational methods play key roles in revealing multiple aspects of single-cell 3D genomes based on scHi-C data. A typical set of analysis tasks includes data processing, dimension reduction for identifying cell clusters with distinct 3D genome organization, improving data quality, and various downstream analyses (see Fig. 1.2). As compared to one-dimensional single-cell assays such as single-cell RNA-seq (scRNA-seq) and single-cell ATAC-seq (scATAC-seq), scHi-C produces two-dimensional contact maps for each cell, which is naturally much sparser and noisier even with similar sequencing depth. More importantly, the contact map representation differs from the vector representation, making it difficult to directly re-purpose the existing computational methods developed for scRNA-seq and scATAC-seq. Chapter 3 and 4 are related to scHi-C.

1.3.2 Spatial transcriptome technology

Spatial transcriptome (ST) technologies combine genomic and spatial data to provide a comprehensive understanding of gene expression patterns within their native tissue environments. This integration is crucial for studying complex biological systems and diseases, as it reveals not only what genes are active, but also where in the tissue these activities occur. Spatial transcriptome technologies can be largely categorized into imaging-based technologies and sequencing-based technologies.

Imaging-based technologies rely on microscopic imaging techniques to visualize RNA molecules within tissue sections. The core method, known as single-molecule fluorescent in situ hybridization (smFISH) [71], involves using fluorescent probes that bind to specific RNA sequences, allowing for the visualization of individual RNA molecules at the sub-cellular level. Advances in this area have led to the development of multiplexed versions like seqFISH [27, 43] and MERFISH [72]. SeqFISH utilizes a barcode system where each RNA molecule is labeled over multiple rounds of hybridization, enabling the simultaneous detection of many different RNA species. MERFISH further improves on this by using a combination of error-robust barcoding and successive imaging to enhance both the multiplexing capabilities and the accuracy of RNA detection.

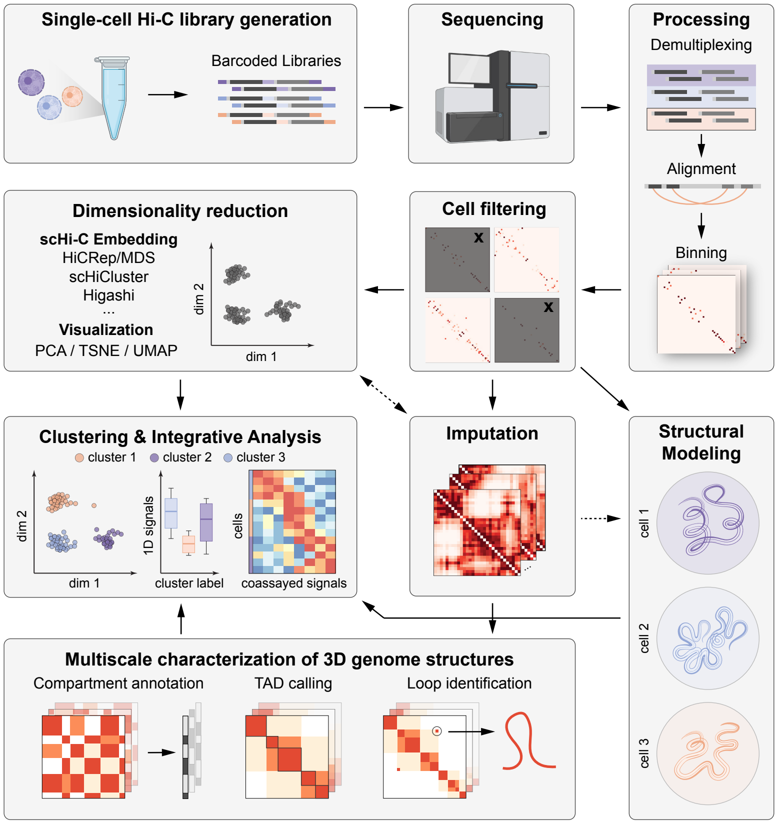

Figure 1.2: A typical workflow for scHi-C data analysis. For a scHi-C sequencing library, the sequencing reads are demultiplexed based on the cellular barcodes, aligned to the reference genome, and binned into single-cell contact maps. After removing low-quality cells, computational methods are used to reduce the dimensionality (i.e., embedding), enhance the data quality (i.e., contact map imputation), and analyze the 3D genome structures. Based on the learned embeddings and the imputed contact maps, downstream analysis such as characterizing multiscale 3D genome features for single cells and clustering cells into distinct cellular states can be performed to reveal the heterogeneity and dynamics of 3D genome organization.

Sequencing-based technologies integrate traditional RNA sequencing methods with spatial resolution techniques. One common method involves the use of microdissection tools to isolate specific tissue regions followed by RNA sequencing, which allows researchers to associate gene expression profiles with their precise locations within the tissue. More advanced technologies, such as Slide-seq [29, 33] and 10x Genomics' Visium [73], employ bead-based or array-based capture systems that retain spatial coordinates and facilitate high-throughput sequencing. These platforms can profile thousands of genes across large tissue areas, providing a detailed spatial map of gene expression that is invaluable for understanding tissue structure and function at the molecular level.

Both imaging and sequencing technologies in spatial transcriptomics are pivotal for advancing our understanding of biological tissues, enabling detailed studies on development, disease pathology, and gene expression dynamics. Chapter 2 and 4 are related to the spatial transcriptome.

1.4 Introduction to relevant computational teqchniques

1.4.1 Unsupervised learning

Unsupervised learning is a branch of machine learning that focuses on identifying patterns and structures from datasets without labeled outcomes. This method operates without guidance, finding hidden structures in unlabeled data, which is particularly useful where the true labels are unknown or hard to obtain.

In unsupervised learning, techniques such as clustering help group similar data points together, revealing inherent groupings within the data [74, 75]. Dimensionality reduction simplifies data by reducing its features to the most significant ones, maintaining the essence while removing noise [76].

In computational biology, unsupervised learning is invaluable for deciphering complex biological data. For example, it is employed in genomics to identify cell types from single-cell gene expression data without prior knowledge [77]. It helps in understanding disease mechanisms by clustering patients based on their genomic profiles, which can uncover new biological insights or subtypes of diseases without predefined labels. Unsupervised learning thereby acts as a powerful tool for hypothesis generation and the discovery of novel biological insights. Chapters 2, 3, and 4 are related to unsupervised learning.

1.4.2 Supervised learning

Supervised learning is a core branch of machine learning where models are trained using labeled data—data that includes the answer or outcome for each example. This method relies on using known data inputs (features) and outputs (labels) to train a model that can predict the output for new, unseen data. The process involves learning a function that maps input features to outputs, which is then validated and refined using test data.

In computational biology, supervised learning plays a crucial role in numerous applications. For example, supervised models are trained to understand the relationships between genetic modifications and their impact on gene expression [78], aiding in the discovery of genetic markers linked to particular traits or diseases. Chapter 5 is related to supervised learning.

1.4.3 Predictive modeling

Predictive modeling is a technique in data science that uses statistical algorithms and machine learning techniques to predict outcomes based on input data. In computational biology, this approach is particularly valuable as it allows researchers to forecast biological behaviors and properties from complex biological data sets.

In computational biology, predictive models are used to understand genomic function. Predictive models help in identifying the functions of various genomic regions by analyzing sequence data and correlating it with phenotypic outcomes [78]. This is essential for understanding gene regulation, expression patterns, and ultimately, their role in health and disease. Chapter 5 is related to predictive modeling.

1.4.4 Representation learning

Representation learning, a subset of machine learning, focuses on automatically discovering the representations needed for feature detection or classification from raw data [79]. This approach enables a machine to identify the most informative features from large and complex datasets without extensive human intervention.

Representation learning has a broad application in computational biology. In genomics, representation learning can be used to transform raw DNA sequences into a lower-dimensional space that captures essential biological features [80]. This is crucial for tasks like predicting gene function, identifying regulatory motifs, or understanding the genetic basis of diseases. In single-cell sequencing, representation learning is used to process high-dimensional data from thousands of cells. Models can learn to represent cell states, types, or developmental stages, facilitating downstream analysis like clustering, trajectory inference, and differential expression analysis [81]. Chapters 2 and 3 are related to representation learning.

1.4.5 Multi-modal integration

Multi-modal integration refers to the process of combining information from multiple different sources or modalities to improve the understanding, analysis, or performance of a system [82]. This approach is widely used across various fields including data science, artificial intelligence, and healthcare, where it leverages the strengths of different data types to provide a more comprehensive view than any single source could offer.

Multi-modal integration in computational biology refers to the synthesis of diverse types of biological data to gain a comprehensive understanding of biological systems and processes [83]. This approach is increasingly critical as the field of biology has expanded to generate voluminous datasets from different biological levels, such as genomic, transcriptomic, proteomic, and metabolomic data. The integration of these diverse data types allows researchers to obtain a more holistic view of how complex biological systems function at multiple levels of regulation. For example, integrating genomic, transcriptomic, and proteomic data can help identify the mechanisms underlying complex diseases by providing insights into how changes at the DNA level affect RNA and protein expressions, which in turn influence cellular behavior and disease phenotypes. This can be particularly valuable in cancer research, where such integrative analyses can reveal how genetic mutations influence tumor progression and response to treatment. Chapter 4 is related to multi-modal integration.

1.4.6 Probabilistic graphical model

Probabilistic graphical models (PGMs) are a sophisticated framework in statistics that combines probability theory and graph theory to model complex networks of variables with uncertainty [84]. They graphically represent the conditional dependencies among multiple variables, which can be either directed (Bayesian networks) or undirected (Markov networks). By encapsulating the dependencies in a visual structure, these models facilitate efficient computation of joint probabilities and provide a systematic approach for inference and learning in large datasets.

In computational biology, PGMs have been pivotal due to their ability to model the stochastic nature of biological processes and integrate diverse types of data. They enable researchers to infer gene regulatory networks from expression data, predict protein structures from amino acid sequences, and understand genetic linkages and their effects on phenotypes. Hidden Markov models [85], a type of PGM, are widely used in bioinformatics for sequence alignment and protein domain prediction, crucial for understanding genetic sequences and their functional implications. Chapter 2 is related to probabilistic graphical models.

1.4.7 Tensor decomposition

Tensor decomposition is a mathematical technique that generalizes matrix decomposition to higher-dimensional data, referred to as tensors [86]. A tensor is a multi-dimensional array, and its decomposition involves breaking it down into simpler, interpretable components. This process can uncover hidden patterns in the data across multiple dimensions, which is particularly useful when dealing with complex datasets. Chapter 3 is related to tensor decomposition.

1.4.8 Non-negativity

Non-negativity is a mathematical constraint applied in various analytical techniques, where it restricts the values of variables or elements to be non-negative (zero or positive). This principle is crucial in many computational models [87], particularly when the variables under study inherently cannot assume negative values, such as concentrations of molecules, expression levels of genes, or counts of biological entities.

In computational biology, non-negativity plays a pivotal role in data analysis and modeling. For example, non-negative Matrix Factorization (NMF) is a widely used technique in computational biology for decomposing high-dimensional datasets into a lower-dimensional space while maintaining non-negativity, making the components easier to interpret biologically [52]. It's particularly useful in gene expression analysis, where it helps in identifying patterns and clusters in the data that correspond to biological pathways or cellular processes. Chapter 2 is related to non-negativity.

1.4.9 Interpretability

Interpretability in machine learning and computational models refers to the ability to understand and explain how decisions or predictions are made by a model. In the context of computational biology, where models are often complex and deal with high-dimensional data, interpretability is crucial for validating the scientific and clinical relevance of the findings. For example, in gene expression analysis, models that are interpretable can help biologists understand which genes or regulatory elements are driving particular phenotypes or disease outcomes. This can be particularly useful in identifying new therapeutic targets or understanding disease mechanisms.

The push for greater interpretability in computational biology not only aids scientific discovery but also increases trust in machine learning models, particularly in genomics where understanding the basis for understanding the complex biological mechanism. Chapters 2, 3, and 5 are related to interpretability.

Chapter 2

SpiceMix enables integrative single-cell spatial modeling of cell identity

2.1 Introduction

The compositions of different cell types in mammalian tissues, such as brain, remain poorly understood, due to the complex interplay among intrinsic, spatial, and temporal factors that collectively contribute to cell identity [2–4]. The emerging spatial transcriptomics technologies based on multiplexed imaging and sequencing [22–32] are able to reveal spatial information of gene expression of dozens to tens of thousands of genes in individual cells *in situ* within the tissue context. However, the development of computational methods that can incorporate the unique properties of spatially-resolved transcriptome data to unveil cell identities and spatially-variable features remains a challenge [37, 38].

Computational methods have been developed to use spatial transcriptome data to identify spatial domains and cell types in tissues [42–46], to explore the spatial variance of genes [47–50], and to align scRNA-seq with spatial transcriptome data [51–54]. To model spatial dependencies, methods using hidden Markov random fields (HMRFs) have been proposed [43, 46]. However, the conventional HMRF has two major limiting assumptions for modeling cell identity: that cell types or spatial domains are discrete, thereby ignoring the interplay of intrinsic and spatial factors, and that they exhibit smooth spatial patterns, which is not true of many cell types, such as inhibitory neurons with sparse spatial patterns. More recently, graph convolution neural networks, such as SpaGCN [44], have been used for identifying spatial domains, but such methods are more susceptible to overfitting and their learned latent representations are not easily interpreted, in comparison to effective linear latent variable models for scRNA-seq data, such as non-negative matrix factorization (NMF) [88]. In addition, the existing methods typically do not integrate the modeling of the spatial variability of genes with their contribution to cell identity. Therefore, there is a need for robust, interpretable methods that can jointly model both the spatial and intrinsic factors of cell identity, which is of vital importance to fully utilize the novel properties of spatial transcriptome data.

Here, we introduce SPICEMIX (Spatial Identification of Cells using Matrix Factorization), an interpretable and integrative framework to model cellular diversity based on spatial transcriptome data. SPICEMIX uses latent variable modeling to elucidate the interplay of spatial and intrinsic factors of cell identity. Crucially, SPICEMIX enhances the NMF [88] model of gene expression by integrating with a graphical model of the spatial organization of cells, leading to more meaningful latent representations. Applications to the spatial transcriptome datasets of brain regions in human and mouse acquired by seqFISH+ [27], STARmap [28], and Visium [73] demonstrate, on both imaging-based and spatial-barcoding-based sequencing technologies, that the enhanced SPICEMIX model of cell identity can uncover complex spatially-variable metagenes and unveil important biological processes.

2.2 Methods

2.2.1 The probabilistic graphical model NMF-HMRF in SPICEMIX

Gene expression as matrix factorization

We consider the expression of individual cells $Y = [y_1, \ldots, y_N] \in \mathbb{R}_+^{G\times N}$, where constants G and N denote the number of genes and cells, respectively, to be the product of K underlying factors (i.e., metagenes), $M = [m_1, \ldots, m_K] \in \mathbb{R}^{G\times K}$, $m_k \in \mathbb{S}_{G-1}$, and weights, $X = [x_1, \ldots, x_N] \in \mathbb{R}_+^{K\times N}$, i.e.,

$$Y = MX + E. \tag{2.1}$$

This follows the non-negative matrix factorization (NMF) formulation of expression of prior work [89]. The term $E = [e_1, \ldots, e_N] \in \mathbb{R}^{G \times N}$ captures unexplained variation or noise, which we model as i.i.d. Gaussian, i.e., $e_i \sim \boldsymbol{\mathcal{N}}(0, \sigma_y^2 I)$. To resolve the scaling ambiguity between M and X, we constrain the columns of M to sum to one, so as to lie in the $(G-1)$-dimensional simplex, $\mathbb{S}_{G-1}$. For notational consistency, we use capital letters to denote matrices and use lowercase letters denote their column vectors.

Graphical model formulation

The formulation for our probabilistic graphical model NMF-HMRF in SPICEMIX enhances standard NMF by modeling the spatial correlations among samples (i.e., cells or spots in this context) via the HMRF [90]. This novel integration aids inference of the latent M and X by enforcing spatial consistency. The spatial relationship between cells in tissue is represented as a graph $\boldsymbol{\mathcal{G}} = (\boldsymbol{\mathcal{V}}, \boldsymbol{\mathcal{E}})$ of nodes $\boldsymbol{\mathcal{V}}$ and edges $\boldsymbol{\mathcal{E}}$, where each cell is a node and edges are determined from the spatial locations. Any graph construction algorithm, such as distance thresholding or Delaunay triangulation, can be used for determining edges. For each node i in the graph, the measured gene expression vector, y_i, is the set of observed variables and the weights, x_i, describing the mixture of metagenes are the hidden states. The observations are related to the hidden variables via the potential function ϕ, which captures the NMF formulation. The spatial affinity between the metagene proportions of neighboring cells is captured by the potential function φ. Together, these elements constitute the HMRF.

More specifically, the potential function ϕ measures the squared reconstruction error of the observed expression of cell i according to the estimated x_i and M,

$$\phi(y_i, x_i) = \exp\left(-U_y(y_i, x_i)\right), \quad U_y(y_i, x_i) = \frac{(y_i - Mx_i)^2}{2\sigma_y^2}, \tag{2.2}$$

where σ_y^2 represents the variation of expression, or noise, of the NMF. The spatial potential function φ measures the inner-product between the metagene proportions of neighboring cells i and j, weighted by the learned, pairwise correlation matrix Σ_x^{-1}, which captures the spatial affinity of metagenes, i.e.,

$$\varphi\left(x_i, x_j\right) = \exp\left(-U_x(x_i, x_j)\right), \quad U_x(x_i, x_j) = \frac{x_i^\top}{\|x_i\|_1}\Sigma_x^{-1}\frac{x_j}{\|x_j\|_1}. \tag{2.3}$$

This form for φ has several motivations. The weighted inner-product allows the affinity between two cells to be decomposed simply as the weighted sum of affinities between metagenes and for the metagenes to have different and learnable affinities between each other. It also allows the model to capture both positive and negative affinities between metagenes. By normalizing the weights x_i of each cell, any scaling effects, such as cell size, are removed. In this way, the similarity that is measured is purely a function of the relative proportions of metagenes. This form also affords a straightforward interpretation for the affinity matrix Σ_x^{-1}. Lastly, it is more convenient for optimization.

Given an observed dataset, the model can be learned by maximizing the likelihood of the data. By the Hammersley-Clifford theorem [91], the likelihood of the data for the pairwise HMRF can be formulated as the product of pairwise dependencies between nodes,

$$P(Y, X|\Theta) = \frac{1}{Z(\Theta)}\prod_{(i,j)\in\boldsymbol{\mathcal{E}}}\varphi(x_i, x_j)\prod_{i\in\boldsymbol{\mathcal{V}}}\phi(y_i, x_i)\pi(x_i), \tag{2.4}$$

where $\Theta = \{\Delta, M\}$ is the set of model parameters and metagenes and $Z(\Theta)$ is the normalizing partition function that ensures P is a proper probability distribution. The potential function π is added to capture an exponential prior on the hidden states X,

$$\lambda_x = 1, \quad \pi(x_i) = \exp\left(-\lambda_x\|x_i\|_1\right), \tag{2.5}$$

with scale parameter 1. We normalize the average of the total normalized expression levels in individual cells to K correspondingly.

Parameter priors

We introduce a regularization hyperparameter λ_Σ on the spatial affinities, which allows the users to control the importance of the spatial relationships during inference to suit the dataset of interest. As the parameter decreases, the influence of spatial affinities during inference diminishes and the model becomes more similar to standard NMF. If we represent

λ_Σ in the form $\lambda_\Sigma = 1/(2\sigma_\Sigma^2)$, we can treat it as a Gaussian prior, with zero mean and σ_Σ^2 variance, on the elements of the spatial affinity matrix Σ_x^{-1},

$$P\left(\Sigma_x^{-1}\right) = \left(\sqrt{\pi/\lambda_\Sigma}\right)^{-K^2} \exp\left(-\lambda_\Sigma \left\|\Sigma_x^{-1}\right\|_F^2\right), \tag{2.6}$$

where F denotes the Frobenius norm. Note that the matrix Σ_x^{-1} is forced to be transpose symmetric.

2.2.2 Alternating estimation of hidden states and parameters

To infer the hidden states and model parameters of the NMF-HMRF model in SPICEMIX, we optimize the data likelihood via coordinate ascent, alternating between optimizing hidden states and model parameters. First, to make inference tractable, we approximate the joint probability of the hidden states by the pseudo-likelihood [91], which is the product of conditional probabilities of the hidden state of individual nodes given that of their neighbors,

$$P(X|\Theta) \approx \prod_{i\in\boldsymbol{\mathcal{V}}} P(x_i|x_{\eta(i)}, \Theta), \tag{2.7}$$

where $\eta(i)$ is the set of neighbors to node i.

Estimation of hidden states

Given parameters Θ of the model, we estimate the factorizations X by maximizing their posterior distribution. The maximum a posteriori (MAP) estimate of X is given by:

$$\hat{X} = \underset{X\in\mathbb{R}_+^{K\times N}}{\arg\max}\, P(X|Y,\Theta) = \underset{X\in\mathbb{R}_+^{K\times N}}{\arg\max}\, P(Y,X|\Theta) = \underset{X\in\mathbb{R}_+^{K\times N}}{\arg\max}\, \{\log P(Y,X|\Theta)\} \tag{2.8}$$

$$= \underset{X\in\mathbb{R}_+^{K\times N}}{\arg\max} \left\{ \sum_{i\in\boldsymbol{\mathcal{V}}} \left[-U_y(y_i, x_i) + \log \pi(x_i)\right] - \sum_{(i,j)\in\boldsymbol{\mathcal{E}}} U_x(x_i, x_j) \right\}. \tag{2.9}$$

This is a quadratic program and can be solved efficiently via the iterated conditional model (ICM) [92] using the software package Gurobi [93].

Estimation of model parameters

Given an estimate of the hidden states X, we can likewise solve for the unknown model parameters Θ by maximizing their posterior distribution. The MAP estimate of the parameters Θ is given by:

$$\hat{\Theta} = \arg\max_{\Theta} P(\Theta|Y,X) = \arg\max_{\Theta} P(Y,X|\Theta)P(\Theta) = \arg\max_{\Theta} \{\log P(Y,X|\Theta) + \log P(\Theta)\} \tag{2.10}$$

$$= \arg\max_{\Theta} \left\{ \sum_{i \in \boldsymbol{\mathcal{V}}} [-U_y(y_i,x_i) + \log \pi(x_i)] - \sum_{(i,j)\in\boldsymbol{\mathcal{E}}} U_x(x_i,x_j) - \log Z(\Theta) + \log P(\Theta) \right\} \tag{2.11}$$

$$\approx \arg\max_{\Theta} \left\{ \sum_{i \in \boldsymbol{\mathcal{V}}} [-U_y(y_i,x_i) + \log \pi(x_i) - \log Z_i(\Theta)] - \sum_{(i,j)\in\boldsymbol{\mathcal{E}}} U_x(x_i,x_j) + \log P(\Theta) \right\}. \tag{2.12}$$

Eqn. 2.12 is an approximation by the mean-field assumption [91], which is used, in addition to the pseudo-likelihood assumption, to make the inference of model parameters tractable. We note that we can estimate metagenes, spatial affinity, and the noise level independently. The MAP estimate of the metagenes M is a quadratic program, which is efficient to solve. The MAP estimate of Σ_x^{-1} is convex and is solved by the optimizer Adam [94]. Due to the complexity of the partition function $Z_i(\Theta)$ of the likelihood, which includes integration over X, it is approximated by Taylor's expansion. Since it is a function of Θ, this computation must be performed at each optimization iteration.

Initialization

To produce the initial estimates of the model parameters and hidden states, we do the following. First, we use a common strategy for initializing NMF, which is to cluster the data using K-means clustering, with K equal to the number of metagenes, and use the means of the clusters as an estimate of the metagenes. We then alternate for T_0 iterations between solving the NMF objective for X and M. This produces, in only a few quick iterations, an appropriate initial estimate for the algorithm, which will be subsequently refined. We observed that if T_0 is too large, it can cause the algorithm to prematurely reach a local minimum before spatial relationships are considered. However, this value can be easily tuned by experimentation, and in our analysis, we found that just 5 iterations were necessary.

2.2.3 Empirical running time

On a CentOS 7 machine with sixteen 2.30GHz Intel(R) Xeon(R) Gold 5218 CPUs and one GeForce 2080 Ti GPU, SPICEMIX takes 0.5-2 hours to run on a typical spatial transcriptome dataset with 2,000 genes and 1,000 cells. The GPU is used for the first 5 iterations, or around that number, only, when the spatial affinity matrix Σ_x^{-1} is changed significantly. In subsequent iterations, most time is spent solving quadratic programs. Since the algorithm uses a few iterations of NMF to provide an initial estimate, which is a reasonable starting point, it is expected to find a good initial estimate of metagenes and latent states efficiently.

2.2.4 Generation and analysis of simulated data

We generated simulated spatial transcriptomic data following expression and spatial patterns of cells of the mouse primary visual cortex. Cells in the mouse cortex are classified into three primary categories: inhibitory neurons, excitatory neurons, and non-neurons or glial cells [95, 96]. Excitatory neurons in the cortex exhibit dense, concentrated, layer-wise specificity, whereas inhibitory neurons are sparse and can be spread across several layers. Non-neuronal cells can be either layer-specific or scattered across layers. We simulated single-cell data from an imaging-based method applied to a slice of tissue, which consists of four distinct vertical layers and eight cell types: four excitatory, two inhibitory, and two glial (Fig. 2.2a). Each layer was densely populated by one layer-specific excitatory neuron type. The two inhibitory neuron types were scattered sparsely throughout several layers. One non-neuronal type was restricted to the first layer and the other was scattered sparsely throughout several layers. For each simulated image, or tissue sample, 500 cells were created with locations generated randomly in such a way so as to maintain a minimum distance between any two cells, so that the density of cells across the sample was roughly constant. With this spatial layout of cells, we devised two methodologies for generating gene expression data for individual cells. The first uses a metagene-based formulation and the second uses a recent method, scDesign2 [97], which we fit to real scRNA-seq data of the mouse cortex [95].

2.2.5 Data processing for the used spatial transcriptome datasets

Preprocessing and analysis of seqFISH+ data

We applied SpiceMix on a seqFISH+ dataset that profiled the mouse primary visual cortex [27]. We first removed genes which had non-zero expression in less than 40% of cells, which yielded an unbiased set of 2,470 genes. We then normalized the expression of these genes by scaling the total counts to 10,000 per cell, adding one, and applying the log transform: $E'_{ig} := \log\left(1 + \left(10^4 \frac{E_{ig}}{\sum_{g'} E_{ig'}}\right)\right)$. To generate a graphical representation of the cells, we applied Delaunay triangulation to physical coordinates of cells, and then removed edges of length larger than 300 pixels (30.9 μm).

For the regularization parameter of the spatial pairwise dependency, λ_Σ, we considered possible values in the set $\{2|\boldsymbol{\mathcal{E}}| \times 10^{-2}, 2|\boldsymbol{\mathcal{E}}| \times 10^{-4}, 2|\boldsymbol{\mathcal{E}}| \times 10^{-6}\}$. We found $2|\boldsymbol{\mathcal{E}}| \times 10^{-4}$ to yield the desired balance of spatial regularization based upon visual inspection. We experimented with the number of metagenes, K, and chose the highest value before the expression of metagenes became too sparse. We also examined the UMAP plots of latent states, without annotations from the original analysis, to guide our selection. This led us to use $K = 20$ metagenes for both SpiceMix and NMF. For each hyperparameter configuration, we ran several iterations of the algorithm with different initial random seeds and chose the random seed that resulted in the highest value of the objective function, Q. After learning the latent states, we z-score normalized the latent states along the cell dimension and performed hierarchical clustering on the normalized latent states to define cell type assignment using Ward's method and the Euclidean distance [98]. We used the Calinski-Harabasz (CH) index [99] as the criterion for determining the optimal number of clusters. Before downstream analysis, we repeatedly merged the two clusters with the lowest threshold form hierarchical clustering until the last 3 splits did not create any cluster with less than five cells. We then eliminated outlier SpiceMix cell types that had less than five cells. This led to 15 cell types for SpiceMix and 13 cell types for NMF.

Preprocessing and analysis of STARmap data

We also applied SpiceMix on a STARmap dataset that profiled the mouse primary visual cortex [28]. We normalized the data by scaling the total counts to 10,000 per cell, adding one, and applying the log transform: $E'_{ig} := \log\left(1 + \left(10^4 \frac{E_{ig}}{\sum_{g'} E_{ig'}}\right)\right)$. To generate a graphical representation of the cells, we applied Delaunay triangulation to physical coordinates of cells, and then removed edges of length larger than 600 pixels.

For the regularization parameter of the spatial pairwise dependency, λ_Σ, we considered possible values in the set $\{2|\boldsymbol{\mathcal{E}}| \times 10^{-2}, 2|\boldsymbol{\mathcal{E}}| \times 10^{-4}, 2|\boldsymbol{\mathcal{E}}| \times 10^{-6}\}$. We found $2|\boldsymbol{\mathcal{E}}| \times 10^{-4}$ to yield the desired balance of spatial regularization based upon visual inspection. We experimented with the number of metagenes, K, and chose the highest value for each algorithm before the expression of metagenes became too sparse. We also examined the UMAP plots of latent states, without annotations from the original analysis, to guide our selection. This led us to use $K = 20$ metagenes for SpiceMix and $K = 15$ metagenes for NMF. For each hyperparameter configuration, we ran several iterations of the algorithm with different initial random seeds and chose the random seed that resulted in the highest value of the objective function, Q. After learning the latent states, we z-score normalized the latent states along the cell dimension and performed hierarchical clustering on the normalized latent states to define cell type assignment using Ward's method and the Euclidean distance [98]. We used the CH index as the criterion for determining the optimal number of clusters. Before downstream analysis, we removed an outlier SpiceMix cell type that had only one cell. This led to 16 cell types for SpiceMix and 11 cell types for NMF.

Preprocessing and analysis of Visium data

Lastly, we applied SpiceMix to a dataset acquired from the 10x Genomics Visium platform that profiled spatial transcriptome of the human DLPFC [73]. For analysis with SpiceMix, we removed genes which had non-zero expression in less than 10% of spots, which yielded an unbiased set of 3,194 genes. We did not apply this filtering when using SpaGCN or BayesSpace. We then normalized the expression of these genes by scaling the total counts to 10,000 per spot, adding one, and applying the log transform: $E'_{ig} :=$

$\log\left(1+\left(10^4\frac{E_{ig}}{\sum_{g'}E_{ig'}}\right)\right)$. To generate a graphical representation of the spots, we defined the neighborhood of a spot to be the set of directly adjacent spots in the hexagonal grid, since the spots in each FOV form a hexagonal grid. Therefore, except for spots on the edge of the grid, each spot has exactly 6 neighbors.

2.3 Results

2.3.1 Overview of SpiceMix

SpiceMix models spatial transcriptome data by a probabilistic graphical model, which we call NMF-HMRF (Fig. 2.1 and **Methods**). Our model has a natural interpretation for single-cell spatial transcriptome data, where each node in the graph represents a cell and edges capture nearby cell-to-cell relationships, but it can also be applied to *in situ* sequencing-based methods (e.g., Visium [25]), where each node represents a spatially-barcoded spot that consists of potentially multiple cells.

For each node i in the graphical model, a latent state vector x_i represents the mixture of weights for K different intrinsic or extrinsic factors of cell identity (Fig. 2.1). To capture the continuous nature of cell state, our model extends the standard HMRF by allowing these latent states to be continuous. Importantly, different types of correlations of latent states in nearby cells are captured by the matrix Σ_x^{-1}, which, unlike a conventional HMRF and many other spatial models, does not exclusively assume smooth spatial patterns, but instead has the flexibility to represent both the smooth and sparse spatial patterns that compose real tissue. Each element of the $K \times K$ matrix Σ_x^{-1} represents the pairwise affinity between two factors, providing an intuitive interpretation of the spatial patterns of cells in tissue. For each factor, a "metagene" in the $G \times K$ matrix M captures the expression of its associated genes, where G denotes the number of genes. The observed expression from spatial transcriptome data, $y_i = Mx_i$ for node i, follows a robust linear mixing model, which gives an intuitive interpretation of the relationship of gene expression to the different latent factors representing cell identities and critical genes. Thus, the NMF-HMRF model in SpiceMix is able to uniquely integrate the spatial modeling of the HMRF with the NMF formulation

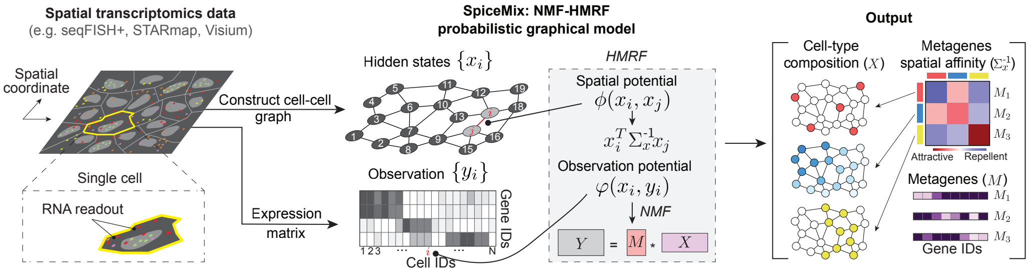

Figure 2.1: Overview of SpiceMix. Gene expression measurements and a neighbor graph are extracted from spatial trancriptome data and fed into the SpiceMix framework. SpiceMix decomposes the expression y_i in cell (or spot) i into a mixture of metagenes weighted by the hidden state x_i. Spatial interaction between neighboring cells (or spots) i and j is modeled by an inner product of their hidden states, weighted by Σ_x^{-1}, the inferred spatial affinities between metagenes. The hidden mixture weights X, the metagene spatial affinity Σ_x^{-1}, and K metagenes M, all inferred by SpiceMix, provide unique insight into the spatially variable features that collectively constitute the identity of each cell.

for gene expression into a single model for spatial transcriptome data.

Given an input spatial transcriptome dataset, SpiceMix simultaneously learns the intuitive metagenes M of latent factors, the latent states X for all nodes, and their spatial affinity Σ_x^{-1}. This is achieved by our alternating maximum *a posteriori* (MAP) optimization algorithm. Importantly, in SpiceMix, metagenes are an integral part of the model outcome, which presents a methodological advance in comparison to the calculation of spatially-variable genes as a post-processing step in other recent methods (such as SpaGCN [44]). A regularizing parameter allows users to control the weight given to the spatial information during optimization to suit the input data. The detailed description of the NMF-HMRF model is provided in the **Methods** section with additional details of optimization in the published version.

2.3.2 Evaluation using simulated spatial transcriptome data

We first evaluated SpiceMix using simulations that model the mouse cortex, a featured region for many spatial transcriptomic studies (Fig. 2.2a-b; see **Methods** for the simulation method details). We devised two methods of generating expression based on the position and type of each cell: Approach I follows a metagene-based simulation; Approach II uses scDesign2 [97] trained on real scRNA-seq data [95]. For Approach II, we introduced two forms of spatial noise: leakage, which randomly swaps some reads of neighboring cells, to mimic challenges of processing real spatial transcriptomics data; and additive noise that follows random, spatially-smooth patterns. We compared the results from SpiceMix to that of NMF, HMRF, Seurat [100], and the recent SpaGCN [44]. We evaluated different methods by comparing the inferred cell types with the true cell types using the adjusted Rand index (ARI) metric. For SpiceMix and NMF, we subsequently applied Louvain clustering to the learned latent representations. The approaches for preprocessing the data and for choosing other hyperparameters for each method are provided in the published version.

For both simulation approaches, we found that SpiceMix consistently outperformed other methods (Fig. 2.2c-e). For Approach I, SpiceMix achieved the highest average ARI scores (0.65-0.82) across scenarios. For lower noise settings ($\sigma_y = 0.2$), the ARI of SpiceMix was 9-18% higher than that of SpaGCN or NMF (Fig. 2.2d). SpiceMix, SpaGCN, and NMF all outperformed Seurat and HMRF. For the higher noise setting ($\sigma_y = 0.3$), SpiceMix clearly outperformed all methods (Fig. 2.2d). We found that SpiceMix was able to recover both the layer-specific and sparse metagenes that underlie the identity of cells. For example, SpiceMix successfully recovered metagene 7, which is specific to layer L1 (Fig. 2.2c) and is enriched in eL1 excitatory neurons (blue in Fig. 2.2a). Notably, SpiceMix was able to reveal nearly all excitatory neurons (Fig. 2.2e). SpiceMix also recovered metagene 6 (Fig. 2.2c), which captures intrinsic factors of the sparse inhibitory neuron subtype i1 (red in Fig. 2.2a). In contrast, the equivalent of metagene 7 for NMF is strongly expressed across layers L1-L3 (Fig. 2.2c), and NMF confused some eL3 excitatory neurons (light green) with eL1 excitatory neurons (Fig. 2.2e). The equivalent of metagene 6 for NMF shows a more diffuse pattern (Fig. 2.2c). Additional evaluation by varying the parameter λ_x or zero-thresholding to reflect different sparsity of the latent variables of NMF further demonstrated the robust advantage of SpiceMix. In addition, SpaGCN, Seurat, and HMRF all incorrectly assigned the spatial patterns for many more excitatory neurons (Fig. 2.2e).

For simulation Approach II, SpiceMix performed the best for all but one scenario, for which it tied with NMF, and the advantage of SpiceMix became more significant as the influence of noise and leakage on spatial expression patterns became more prevalent. We found that the spatial metagenes from SpiceMix reliably reflect both cell type composition and spatial noise . Overall, SpiceMix achieved much more accurate spatial assignments of cells than all other methods.

Taken together, we showed that the integration of matrix factorization and spatial modeling in SpiceMix yields better and robust inference of spatially variable features (both sparse and layer-specific) that underlie cell identities as compared to existing methods.

2.3.3 Improving cell identity modeling of seqFISH+ data

We applied SpiceMix to a recent single-cell spatial transcriptomic dataset of the primary visual cortex of a mouse (five samples of nearby regions), acquired by seqFISH+ [27], with single-cell expression of 2,470 genes in 523 cells [27]. We compared the spatial patterns revealed by SpiceMix to those produced by NMF with various levels of sparsity via λ_x and zero-thresholding, as well as Louvain clustering and the HMRF-based method of Zhu et al. [43], both reported in Eng et al. [27]. In addition, SpiceMix revealed spatially-informed metagenes capturing biological processes in the cortex.

We first clustered the cells in the latent representation of SpiceMix using hierarchical clustering, which revealed five excitatory neural subtypes, two inhibitory neural subtypes, and eight glial types (Fig. 2.3a), supported by known marker genes [95] (Fig. 2.3b (left)). Major cell type assignments were generally consistent among SpiceMix, NMF, and Louvain clustering (Fig. 2.3b (middle)). However, SpiceMix uncovered more refined cell subtypes and states. Notably, SpiceMix identified three distinct clusters following known stages of oligodendrocyte maturation [101], from oligodendrocyte precursor cells (OPCs) to mature, myelin-sheath forming oligodendrocytes, throughout the five samples, as reflected by the spatially-informed metagenes. Metagene 8 is enriched among oligodendrocytes, distinguishing them from OPCs (Fig. 2.3b (right)), while metagene 7, which is also in OPCs, separates a cluster of early-stage oligodendrocytes (Oligo-E) from later-stage oligodendrocytes (Oligo-L), suggesting that these metagenes capture their maturation trajectory. These stages are supported by the expression patterns of the OPC marker gene *Cspg4*, the differentiating oligodendrocyte marker gene *Tcf7l2* [102], and the mature oligo-

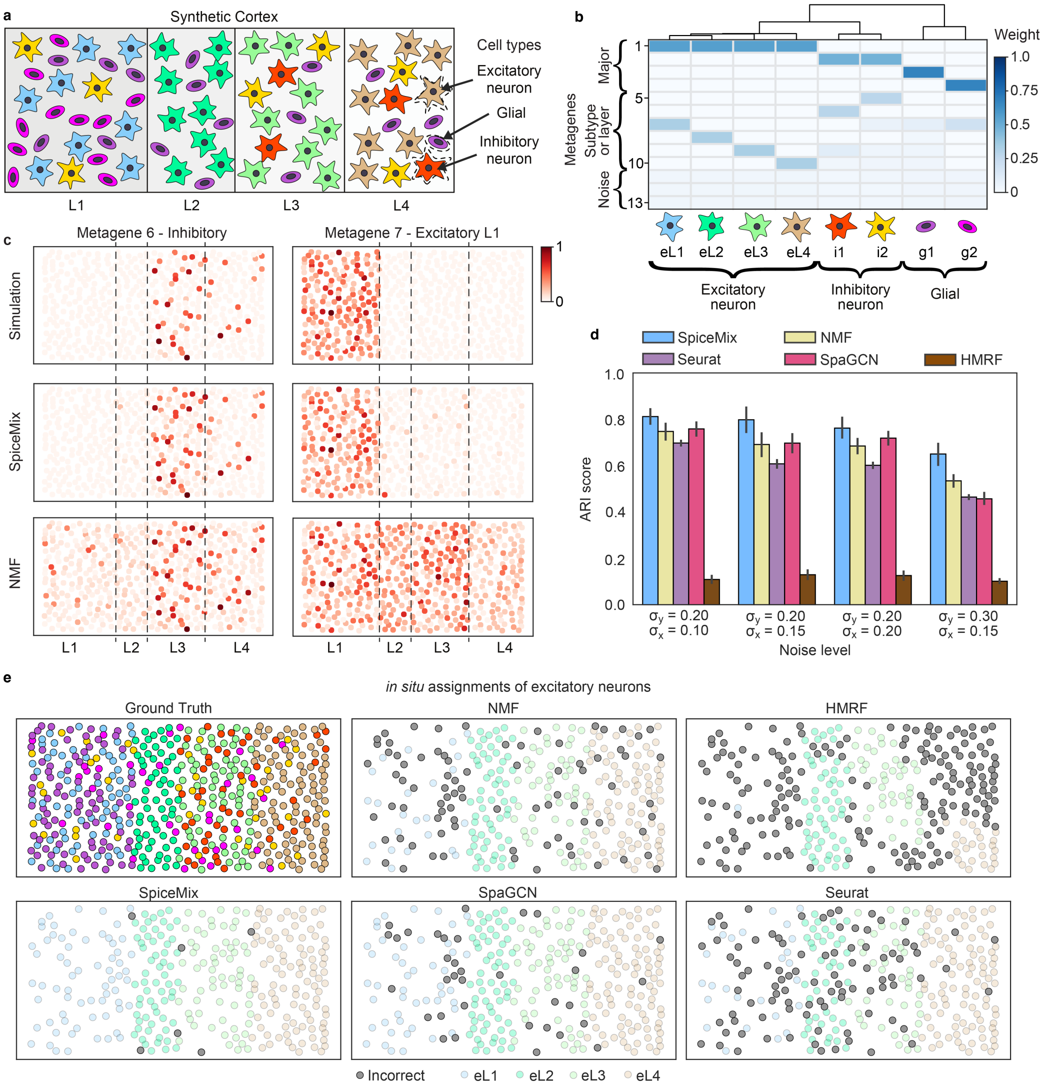

Figure 2.2: Performance evaluation based on simulated spatial transcriptome data. **a.** Illustration of the simulated spatial transcriptome data of the mouse cortex, including 3 major cell types distributed in 4 layers. Excitatory (blue, cyan, green, and brown) and inhibitory (red and yellow) neurons are star-shaped and glial cells (purple and magenta) are ovals. Subtypes are distinguished by their colors. **b.** Dendrogram showing the similarity of the expression profiles of the 8 cell types (top), their metagene profiles (middle), and their colors and shapes (bottom) used in panel **(a)**. The top 4 rows correspond to metagenes that determine major type, the next 6 rows correspond to metagenes that determine subtypes or are layer-specific, and the bottom 3 rows correspond to noise metagenes. **c.** Simulated expression of metagenes 6 and 7, from a single sample generated with $\sigma_y = 0.2$ and $\sigma_x = 0.15$, in their spatial context (top) and the inferred expression of those metagenes by SpiceMix and NMF. Expression levels of metagenes are linearly scaled to $[0, 1]$ for visualization. Visualizations in panel **(e)** are of the same simulated sample. **d.** Performance comparison of SpiceMix, NMF, HMRF, Seurat, and SpaGCN. Bar plots of the average adjusted Rand index (ARI) score, that measures the matching between the identified cell types and the true cell types, are shown. The score is averaged across n=20 replicates per scenario. Results are reported across four simulation scenarios, with varying degrees of randomness. Error bars show +/- one standard deviation. **e.** Imputed cell-type labels of each method for the excitatory neurons, shown in their spatial context. Neurons that were correctly identified are colored faintly. Neurons that were incorrectly identified are colored dark gray. The upper left panel is the ground truth cell type of all cells in the simulated sample. The colors match those of panels **(a)** and **(b)**.

dendrocyte marker gene *Mog* [103] (Fig. 2.3b (left)), in addition to a large set of marker genes for oligodendrocyte stages from [101]. Metagene 7 was distinguished from metagene 8 by its strong spatial affinity with metagenes 3 and 4 (highlighted by black arrows in Fig. 2.3c), which are expressed primarily by the excitatory neurons of deeper tissue layers (eL5, eL6a, and eL6b) (Fig. 2.3b (right)). No other method (NMF, Louvain clustering as reported by [27], and the HMRF-based method of Zhu et al. [43] as reported in [27]) could clearly distinguish these spatially-distinct cells (Fig. 2.3b (middle)). SPICEMIX also discovered spatially-variable features that led to the identification of excitatory and inhibitory neuron subtypes whose layer-specificity patterns matched those of prior scRNA-seq studies.

Together, our analysis of seqFISH+ data of the mouse cortex with SPICEMIX revealed spatially-variable features and more refined cell states. Our results demonstrate the advantages and unique capabilities of joint modelling of spatial and transcriptomic data using SPICEMIX.

2.3.4 Revealing spatial metagenes and cell types from STARmap data

Next, we applied SPICEMIX to a single-cell spatial transcriptome dataset of the mouse V1 neocortex acquired by STARmap [28], consisting of 930 cells passing quality control, all from a single field-of-view (FOV), with expression measurements for 1,020 genes. We compared primarily the results of SPICEMIX, NMF, and Wang et al. [28]. An asterisk is appended to the end of the cell labels of Wang et al. [28] when referenced.

We found that SPICEMIX identified refined, spatial subtypes (Fig. 2.4a) and improved upon the cell labels of [28] (Fig. 2.4b). The learned spatial affinities (Fig. 2.4c) enabled improved cell layer-specificity, which was particularly notable among excitatory neurons (Fig. 2.4d). The clear boundaries between excitatory layers matched layer-enrichment analysis from scRNA-seq studies (see Figure 4b in [104]), in contrast to the cell assignments reported in Figure 5d in [28], which showed significant mixing of excitatory types across boundaries. The NMF formulation of SPICEMIX helped reassign a large set of cells from the Astro-1* type of [28] to eL5, which was further refined along layer boundaries by the

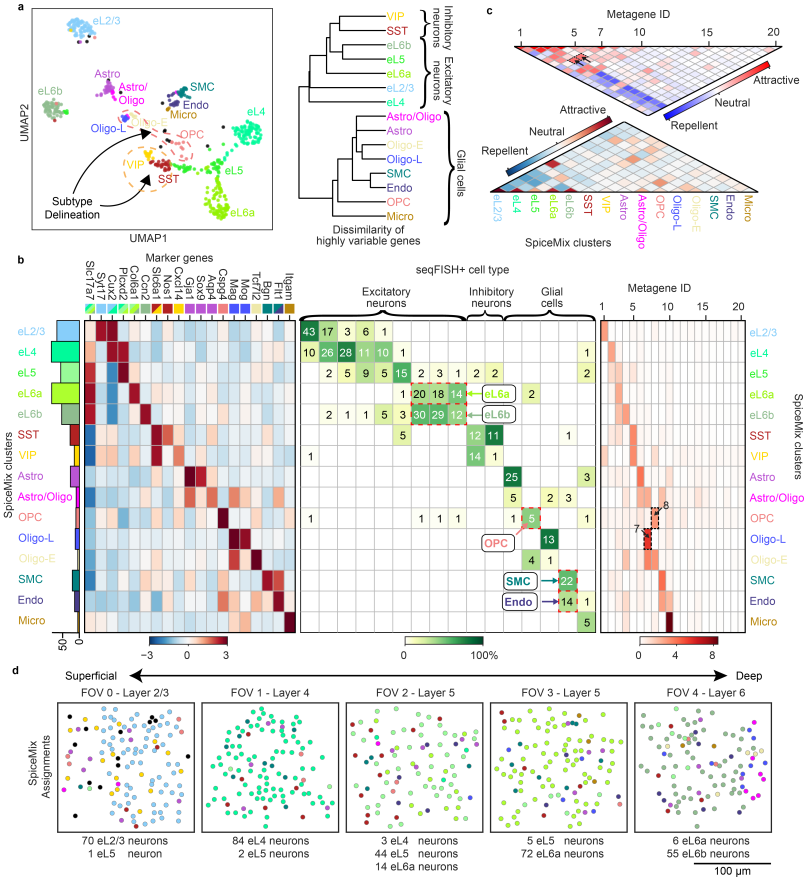

Figure 2.3: Application of SpiceMix to the seqFISH+ data from the mouse primary visual cortex [27]. Note that colors throughout the figure of cells and labels correspond to the cell-type assignments of SpiceMix. **a.** UMAP plot of the latent states of SpiceMix (left) and the dendrogram of the arithmetic average of the expression for each cell type of SpiceMix (right). It is highlighted in **(a)** (left) that SpiceMix further delineated inhibitory neurons into VIPs (yellow) and SSTs (red-brown) enclosed by the orange dashed cycle and refined oligodendrocytes and OPCs into separate subtypes: Astro/Oligo (magenta), Oligo-1 (beige), Oligo-2 (blue), and OPC (coral), enclosed within the red dashed cycle. **b.** (Top) The inferred pairwise spatial affinity of metagenes, or Σ_x^{-1}. The strong attractions between metagene 7 and metagenes 3 and 4, which helped distinguish the spatial patterns of Oligo-L cells, are highlighted by the black arrows. (Bottom) The inferred pairwise spatial affinity of SpiceMix cell types. **c.** (Left) Average z-score normalized expression of known marker genes within SpiceMix cell types, along with the number of cells belonging to each type (colored bar plot). The colored boxes on the top following the name of each marker gene correspond to their known associated cell type. (Middle) Agreement of SpiceMix cell-type assignments with those of the original analysis in [27]. (Right) Average expression of inferred metagenes within SpiceMix cell types. The expression is normalized by the standard deviation per metagene. Metagenes 7 and 8, which revealed the separation of oligodendrocyte subtypes, are highlighted by black arrows. **d.** *In situ* SpiceMix cell-type assignments for all cells in each of the five FOVs. Colors of cell types are the same as in above panels.

learned spatial affinities (Fig. 2.4b (middle), d). This reassignment was supported by the expression of known excitatory marker genes. In contrast, we found that HMRF missed sparse cell types and smoothed across layers, missing even the layer-wise structure of excitatory neurons. Further, SpiceMix achieved a refined, spatially-informed separation of three eL6 subtypes, driven by the identification of two strongly spatially-attracted metagenes: 5 and 7 (highlighted by a black arrow in Fig. 2.4c).

SpiceMix also produced a significant refinement of glial subtypes. SpiceMix identified two oligodendrocyte clusters and an OPC cluster, distinguished by their relative expression of metagenes 12, 13, and 14 (Fig. 2.4b (right)). Metagenes 12 and 13 were highly enriched in layer L6 and strongly attracted to each other (Fig. 2.4c, Fig. 2.5a). Their proportional expression by oligodendrocytes within L6 captured a maturation trajectory from OPCs to Oligo-1 cells that could not be revealed by other methods (see later section). Metagene 14 also has distinct oligodendrocyte markers (Fig. 2.4b right), but scatters from layers L2/3 to L6 (Fig. 2.5a), leading to a spatially distinct Oligo-2 type, clearly separated in the SpiceMix latent space from neighboring excitatory neurons. The expression of oligodendrocyte marker genes identified by [101] supports that the Oligo-1 and Oligo-2 clusters represent mature oligodendrocytes, distinct from the OPCs. In addition, SpiceMix distinguished astrocytes into two types (Astro-1 and Astro-2) based on metagenes 11 and 12. Although Astro-2 cells shared metagene 12 with OPCs, both their spatial location in the superficial layer and the expression of astrocyte marker genes defined them as astrocytes. In contrast, Astro-1 cells expressed metagene 11 with a scattered spatial pattern throughout all layers (Fig. 2.5a). This Astro-1/Astro-2 separation was supported by the expression of known marker genes [105], including *Gfap* (P=0.024), a marker for astrocytes in the glia limitans, and *Mfge8* (P=0.0013), a marker for a separate, diffuse astrocyte type (Fig. 2.5b). We found that NMF did not reveal these subtypes and the NMF metagenes typically exhibited unspecific spatial patterns and pairwise affinity.

These results suggest that SpiceMix is able to refine cell identity and metagene inference with distinct spatial patterns from STARmap data, further demonstrating its advantage.

2.3.5 Identifying continuous oligodendrocytes myelination stages

The expression of metagenes learned by SpiceMix from seqFISH+ and STARmap suggested the existence of continuous factors of oligodendrocyte identity. Applying Monocle2 [106] to the raw counts of cells in the STARmap dataset labeled by SpiceMix as

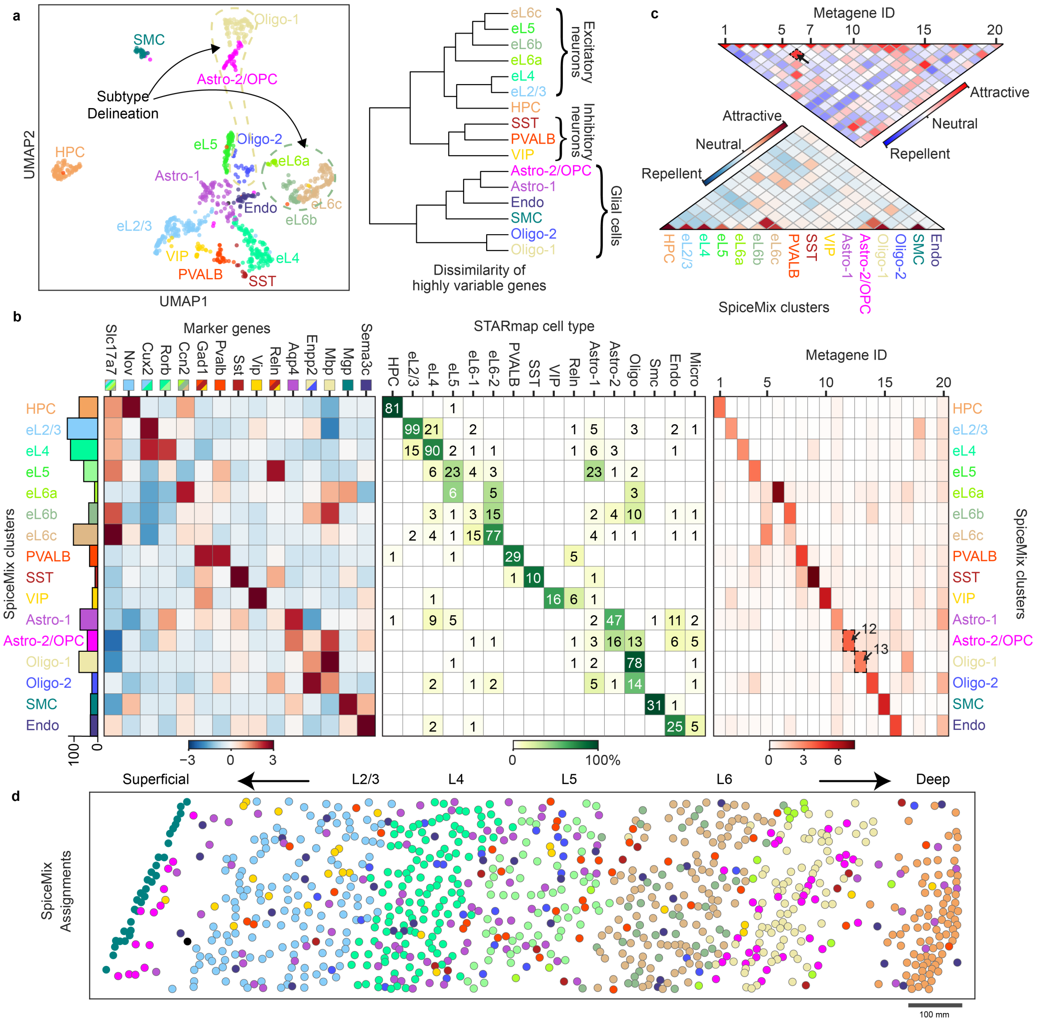

Figure 2.4: Metagenes and refined cell types discovered by SpiceMix from the STARmap data of the mouse primary visual cortex [28]. Note that colors throughout the figure of cells and labels correspond to the cell-type assignments of SpiceMix. **a.** UMAP plots of the latent states of SpiceMix and the dendrogram of the arithmetic average of the expression for each cell type of SpiceMix (right). It is highlighted in **a** (left) that SpiceMix delineated eL6 neurons into three subtypes enclosed in the green cycle and delineated oligodendrocytes and OPCs into three separate subtypes: Oligo-1 (beige), Oligo-2 (blue), and Astro-2/OPC (magenta), enclosed within the beige dashed cycle. **b.** (Top) The inferred pairwise spatial affinity of metagenes, or Σ_x^{-1}. The strong attraction between metagene 5 and metagene 7, which helped distinguish excitatory eL6 neurons, is highlighted by the black arrow. (Bottom) The inferred pairwise spatial affinity of cell types. **c.** (Left) Average z-score normalized expression of known marker genes within SpiceMix cell types, along with the number of cells belonging to each type (colored bar plot). The colored boxes on the top following the name of each marker gene correspond to their known associated cell types. (Middle) Agreement of SpiceMix cell-type assignments with those of the original analysis in [28]. (Right) Average expression of inferred metagenes within SpiceMix cell types. The expression is normalized by the standard deviation per metagene. The average proportion of metagenes 12 and 13 in oligodendrocyte cell types, which helped delineate subtypes, are highlighed by black arrows. **d.** *In situ* map of SpiceMix cell-type assignments for all cells.

oligodendrocytes showed a clear trajectory from the OPCs to the mature Oligo-1 class (Fig. 2.5c). The Oligo-2 class is likely a distinct type of mature oligodendrocytes compared to Oligo-1. Importantly, the relative expression of metagenes 12 and 13, which were

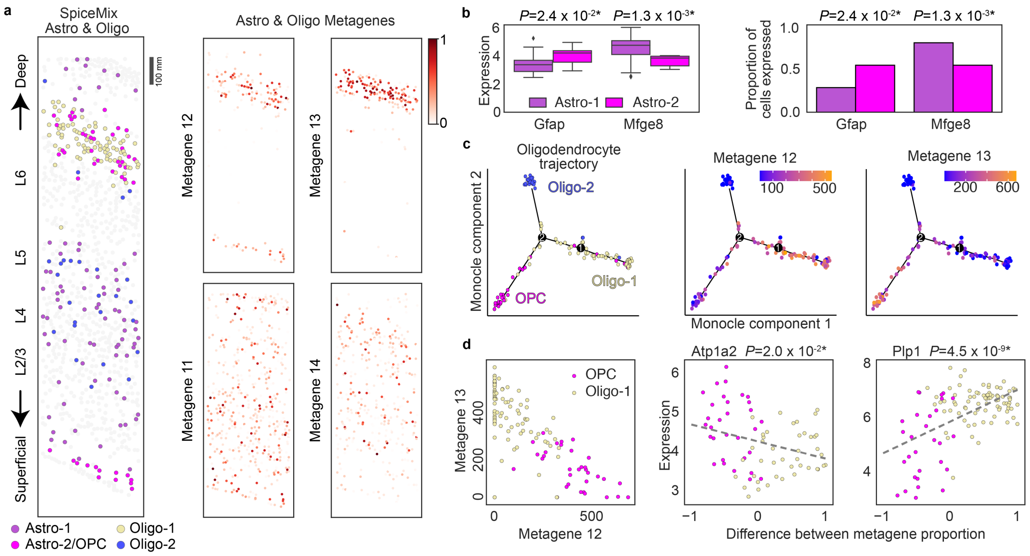

Figure 2.5: Spatial glial subtypes and the process of myelination in oligodendrocytes revealed by SpiceMix metagenes in STARmap data of the mouse primary visual cortex [28]. Note that colors throughout the figure of cells and labels correspond to the cell-type assignments of SpiceMix. **a.** (Left) *In situ* map of SpiceMix cell-type assignments for astrocyte and oligodendrocyte cells in the sample. (Middle and right) *In situ* maps of expression of both layer-specific and ubiquitous metagenes learned by SpiceMix that are relevant to astrocytes and oligodendrocytes. **b.** The log-normalized expression of astrocyte subtype marker genes in Astro-1 (n=78 cells) and Astro-2 (n=13 cells) types of SpiceMix (left), and a comparison of the percentage of cells expressing those marker genes (right). *: The two-sided Wilcoxon rank sum test $P<0.05$. **c.** Trajectory analysis of SpiceMix oligodendrocyte types using Monocle2, showing the unnormalized expression of metagenes 12 and 13 along the trajectory from OPC to Oligo-1. **d.** (Left) The expression of metagene 13 plotted against the expression of metagene 12 for oligodendrocytes of the SpiceMix Oligo-1 and OPC types. (Right) The expression of important marker genes for myelin-sheath formation in oligodendrocytes plotted against the relative expression of metagenes 12 and 13 of the same cells. The dashed lines are the fitted linear regression model. The title of each plot consists of the gene symbol and the Benjamini/Hochberg corrected two-sided Wald test with t-distribution *P*-value of having a nonzero slope, respectively. *: $P<0.05$.

highly expressed in OPC and Oligo-1 cells, respectively, strongly correlated with the inferred trajectory (Fig. 2.5c).

Using linear regression, we tested if the differences in the proportions of metagenes 12 and 13 along this trajectory corresponded to the expected change in expression of myelin sheath-related genes during myelination. The eleven genes that we tested were those from the STARmap panel attributed to myelin sheath formation, according to Gene Ontology (GO) that were expressed in at least 30% of cells. We found that the correlations of seven of the eleven genes are significant ($P<0.05$, after a two-step FDR correction for multiple testing) (Fig. 2.5d), supporting our hypothesis. One of these genes is *Atp1a2*, recently confirmed by scRNA-seq studies to be suppressed as myelination progresses [107, 108], further demonstrating the robustness of our analysis.

This result further demonstrates that the latent representation of SpiceMix is uniquely able to elucidate important biological processes underlying cell states.

2.3.6 Unveiling spatial patterns from Visium human brain data

We next sought to demonstrate the effectiveness and interpretability of SPICEMIX on a dataset of the human dorsolateral prefrontal cortex (DLPFC) acquired by the 10x Genomics Visium platform [73]. We made a direct comparison of SPICEMIX to two recent methods on this dataset: SpaGCN [44] and BayesSpace [46], which was designed for spatial-barcoding methods.

SPICEMIX achieved consistent advantages in identifying the layer structures of DLPFC (Fig. 2.6a), which consisted of six cortical layers (layer L1 to layer L6) and white matter. We focused on the 4 FOVs from sample Br8100 for this analysis. The clusters from SPICEMIX produced an ARI score between 0.54 and 0.61 (average 0.575), with consistent advantage over SpaGCN and BayesSpace (Fig. 2.6a). We observed that although SpaGCN and BayesSpace could produce layer-like patterns, these layers did not closely match the true boundaries. In contrast, SPICEMIX produced contiguous layers for all FOVs and identified clearer boundaries (Fig. 2.6b) and learned metagenes that clearly manifest the layer structure of DLPFC. Using all four FOVs as input did not significantly affect the ARI score of SpaGCN (Fig. 2.6a), and we were unable to run BayesSpace effectively on all four FOVs simultaneously. Although layer L4 could not be reliably identified by any method, the metagenes a3 and a6 learned by SPICEMIX showed differential expression among L3, L4, and L5 ($P < 10^{-300}$, highlighted in Fig. 2.6c).

The interpretability of metagenes from SPICEMIX helped unveil spatially-variable expression and spatial patterns of cell types of DLPFC. We used differentially expressed genes (DEGs) identified from [109]. The high ranks of astrocyte DEGs in metagene a1 (Fig. 2.6d) suggest that it captures astrocyte expression, along with its ubiquitous presence in all seven layers (Fig. 2.6c), consistent with a recent work [72]. Oligodendrocyte DEGs were enriched in metagenes a6 and a7, which were primarily in deep layers and the white matter, respectively (Fig. 2.6c-d). This is consistent with the spatial distributions of oligodendrocytes [110] and suggests a spatial-subtype separation. Moreover, the DEGs of excitatory neurons in superficial layers and deep layers were enriched in metagenes a3 and a6, respectively, which were present mostly in layers L1-L3 and layer L6, accordingly, reflecting the layer-like patterns of excitatory neurons (Fig. 2.6c-d). These findings confirm the unique ability of SPICEMIX to unveil spatially-variable features and cell type composition.

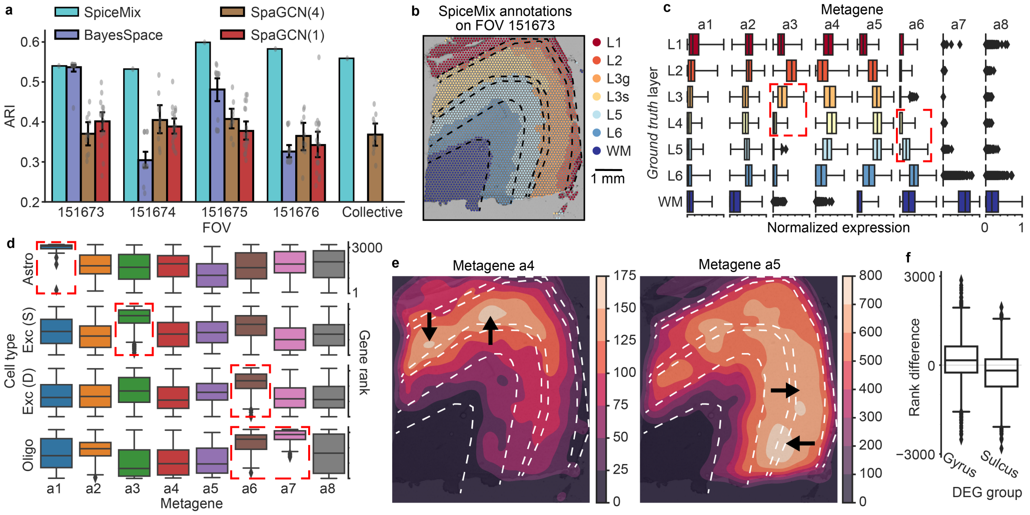

Figure 2.6: Application to the Visium dataset of human dorsolateral prefrontal cortex [73]. **a.** Comparison of the performance of SPICEMIX, BayesSpace, and SpaGCN on the 4 FOVs from sample Br8100. SPICEMIX and SpaGCN(4) were trained on 4 FOVs simultaneously and evaluated both on single FOVs and on 4 FOVs altogether. BayesSpace and SpaGCN(1) were trained and evaluated only on single FOVs. For SpaGCN and BayesSpace, gray dots represent one of n=10 runs with different random seeds. Data are presented as mean values and 95% CIs. **b.** The *in situ* layer assignments of SPICEMIX for FOV 151673. The boundaries between ground-truth layers are illustrated by dashed lines. The gyrus and sulcus subregions of L3 identified by SPICEMIX are labeled L3g and L3s, respectively. **c.** The *in situ* expression of 8 metagenes from SPICEMIX, normalized by the maximum value per metagene across FOVs. Metagenes a3 and a6 collectively distinguish L4 spots (n=7952) from L3 (n=28160) (two-sided t-test P smaller than the smallest representable value) and L5 (n=21400) (two-sided t-test $P = 6 \times 10^{-322}$; red rectangles). **d.** The rank distribution of known marker genes [109] (n=53, 406, 188, and 67 genes, respectively) of 4 cell types in the 8 metagenes. 'Exc (S)' and 'Exc (D)' denote markers of excitatory neurons of superficial and deep layers, respectively. For each row, metagenes with greater ranks are highlighted by red rectangles (one-sided highlighted-vs-rest Mann-Whitney U test $P = 2 \times 10^{-21}, 10^{-90}, 3 \times 10^{-32}, 10^{-28}$, respectively). **e.** Kernel-smoothed *in situ* expressions of metagenes a4 and a5, showing their differential expressions (highlighted by arrows) between the gyric side (right side) and the sulcal side (upper side). **f.** The distribution of the rank difference of gyro-sulcal DEGs between metagenes a4 and a5. Gyric DEGs have greater ranks in a5 than in a4 (two-sided Wilcoxon $P = 3 \times 10^{-26}$, n=1836 genes), and sulcal DEGs exhibit the opposite trend (two-sided Wilcoxon $P = 4 \times 10^{-25}$, n=1136 genes). All boxplots show the median, first, and third quartiles, and whiskers extend no further than 1.5×IQR (inter-quartile range).

2.3.7 Delineating finer anatomic structures of the human brain

SPICEMIX was able to identify finer anatomical structures and cell composition of the brain based on its learned spatially-variable metagenes from the DLPFC Visium data [73]. On the four FOVs from sample Br8100, metagenes a4 and a5 captured the gradual gyro-sulcal variability (Fig. 2.6e-f). We found that more than 50% of the genes used for SpiceMix were differentially expressed across the two regions, strongly supporting this separation. The relative ranking of DEGs within each metagene, according to its weight, was significantly associated with the respective region ($P < 10^{-24}$) (Fig. 2.6f). This shows the distinct ability of the metagenes from SpiceMix to represent gradual changes in spatial gene expression.

Applying SpiceMix to FOV 151507 from sample Br5292 (Fig. 2.7a), we found that metagenes b1-b3 defined three finer anatomical structures within layer L1 annotated in [73] (Fig. 2.7b-c)). Based on the brightness of the staining in the histology image, we classified each spot into one of four types (Fig. 2.7b (top left)): the dark stripe (yellow), the bright gap (green), the flanking cortex (blue), and ambiguous mixtures of these three regions (grey). All 7 marker genes of mural cells, which constitute the wall of blood vessels, from [105] that passed quality control were highly expressed in the dark stripe. The enrichment of 5 out of the 7 genes was significant ($P \leq 0.002$), suggesting that the dark stripe is potentially a blood vessel. Aside from the brightness, spots exhibited other varying phenotypes across the three regions, such as cell density, UMI count, and mitochondrial RNA ratio, indicating that these three regions are biologically different. We found that metagenes b1, b2, and b3 were enriched in the flanking cortex, the white gap, and the blood vessel, respectively (Fig. 2.7b-c), supporting the delineation of the three anatomical structures by SpiceMix.

Additionally, metagenes b4 and b5 defined two finer anatomical structures in the white matter region (Fig. 2.7d). Specifically, metagene b4 was mainly present in a 400μm-wide superficial layer (Fig. 2.7d (S)), whereas metagene b5 was nearly restricted to the deep part (Fig. 2.7d (D)). Spots also exhibited different phenotypes across the two structures that are supported by DEGs. Consistent with this finding, marker genes of oligodendrocytes had a higher rank in metagene b5, which was enriched in the deep part (Fig. 2.7e).

Together, these results further demonstrated the ability of SpiceMix to capture subtle but biologically important anatomical structures from spatial transcriptome data acquired by a variety of technologies.

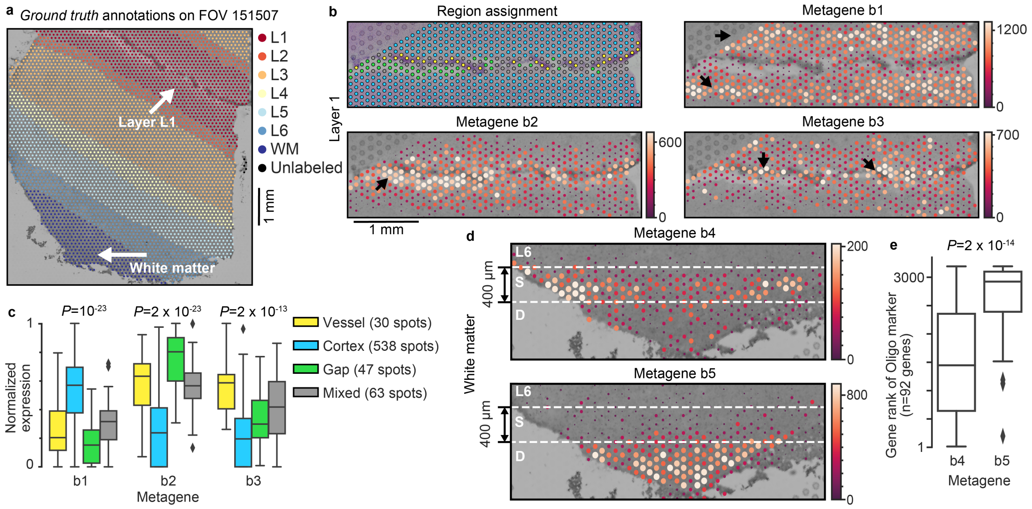

Figure 2.7: SpiceMix metagenes associated with finer anatomical structures in the human dorsolateral preforntal cortex from Visium data [73]. **a.** The *in situ* layer annotations of the *ground truth* on FOV 151507. **b.** The finer structure annotations of spots (top left) and the *in situ* inferred unnormalized expressions of metagenes b1-b3 on FOV 151507 (the other three panels). The color legend of the top left panel is in **(c)**. Based on the intensity on the histological image, a spot was assigned to a dark stripe (green), a bright gap (blue), a peripheral region (orange), or a mixture of the bright gap and dark stripe (grey). As highlighted by black arrows, metagenes b1-b3 are enriched in the peripheral region, the bright gap, and the dark stripe, respectively. **c.** The differential expressions of metagenes b1-b3 across the finer structures. One-sided one-vs-rest Mann-Whitney U test P is displayed above each column. For better visualization, the raw expression levels were divided by the maximum expression level across all spots in the 4 FOVs per metagene. **d.** The inferred *in situ* unnormalized expression of metagenes b4 and b5 on FOV 151507, implying the delineation of the superficial part (denoted by S) and the deep region (denoted by D) in white matter. **e.** The rank distribution of oligodendrocyte marker genes in metagenes b4 and b5. These genes have significantly higher ranks in metagene b5 than in b4 (one-sided Wilcoxon P is shown) All boxplots show the median and first and third quartiles, and whiskers extend to values no further than 1.5×IQR (inter-quartile range).

2.4 Discussion

We have developed SpiceMix, an unsupervised method for modeling the diverse factors of cell identity in complex tissues based on various types of spatial transcriptome data. The integrated model of SpiceMix combines the expressive power of NMF for modeling gene expression with the HMRF for modeling spatial relationships, advancing current state-of-the-art modeling for spatial transcriptomics as clearly shown in both simulation evaluation and real data applications. On single-cell spatial transcriptome data of the mouse primary visual cortex from seqFISH+ and STARmap, SpiceMix demonstrated its effectiveness in producing reliable spatially variable metagenes and biologically informative latent representations of cell identity. On the human DLPFC data acquired by Visium, SpiceMix improved the identification of annotated layers and revealed finer anatomical structures.

A significant feature of SpiceMix is the spatially variable metagene formulation, which can model the interplay of the spatial and intrinsic composition of the transcriptome and not merely the spatial patterns of individual genes [47, 50]. Crucially, as part of the model formulation, SPICEMIX considers how these metagenes are integrally related to continuous cell states, which represents a major distinction compared to other approaches [43, 44]. We note that since SPICEMIX is an unsupervised method, we have showcased its application to datasets with large, unbiased gene panels. Though for datasets with targeted panels guided largely by prior knowledge, a tool such as Tangram [54] could be utilized to extend the gene panel and thereby further increase the power of SPICEMIX.

As the field of spatial transcriptomics continues to grow and become more widely available, new technologies and datasets will open many new directions. In particular, it will be of great interest to model the dynamics of spatial patterns across diverse samples and along normal development or disease progression. Another exciting development is the generation of spatial multiomic data, which integrates transcriptome with other data types such as protein expression [111]. Understanding the relationships between different data modalities within their spatial context could lead to a more complete understanding of the *in situ* molecular underpinning of diverse cell states in complex tissues. There is also continued interest in studying cell-cell interaction and communication [112], which spatial transcriptomics can uniquely elucidate.

Enhanced computational methods that can analyze, summarize, and interpret spatial omics data will be crucial to future studies. By effectively modeling the complex mixing of latent intrinsic and spatial factors of heterogeneous cell identity in complex tissues, SPICEMIX offers a useful tool to facilitate discoveries for diverse types of spatial omics data. We note that SPICEMIX is not limited to transcriptomic data only, and its methodology may also be well-suited for multiomic data. In future work, enhancements may be made to SPICEMIX to allow for progressive changes in the learned spatial patterns. Further, the refined cell identities and learned spatial affinities of SPICEMIX may be useful for studying other aspects of tissue dynamics, including cell-cell interactions. Overall, SPICEMIX is a powerful framework for the analysis of diverse types of spatial transcriptiome and multiomic data, with the distinct advantage that it can unravel the complex mixing of latent intrinsic and spatial factors of heterogeneous cell identity in complex tissues.

Chapter 3

Ultrafast and interpretable single-cell 3D genome analysis

3.1 Introduction

The advent of high-throughput whole-genome mapping methods for the three-dimensional (3D) genome organization such as Hi-C [5] has revealed distinct features of chromatin folding in various scales within the cell nucleus, including A/B compartments [5], subcompartments [6, 7], topologically associating domains (TADs) [8, 9], and chromatin loops [6]. These multiscale 3D genome features collectively contribute to vital genome functions such as transcription [12, 13]. However, the variation of 3D genome features and their functional significance in single cells remain poorly understood [1, 14]. The recent advances of single-cell Hi-C (scHi-C) technologies have provided us with unprecedented opportunities to probe chromatin interactions at single-cell resolution, from a few cells of given cell types [15–18] to thousands of cells from complex tissues [19–21]. These new technologies and datasets have the promise to unveil the connections between genome structure and function in single cells for a wide range of biological contexts in health and disease [14].

However, the complexity of scHi-C data has created significant analysis challenges. Computational methods HiCRep/MDS [39], scHiCluster [40], LDA [17], and the more recent deep learning based methods 3DVI [41] and Higashi [113] have been developed for the embedding and imputation of the sparse scHi-C data. These existing methods, however, cannot (i) effectively infer informative embeddings for the delineation of rare cell types in complex tissues, (ii) directly identify critical chromatin organizations related to cell type-specific genome functions, and (iii) efficiently operate on large-scale datasets with limited memory resources. It remains an open question on how to develop effective computational methods that can identify rare cell types in complex tissues in an interpretable manner with high scalability, key to understanding the interplay among chromatin organization, genome functions, and cellular phenotypes.

The recent scHi-C embedding method scHiCluster [40] uses linear convolution and random walk with restart to impute the sparse contact maps and applies principal component analysis (PCA) on the imputed maps. This requires the storage of all imputed dense maps in the memory, drastically limiting its application to datasets with a large number of cells at high resolution. More recently, deep learning based scHi-C analysis methods have been proposed, including 3DVI [41] based on a deep generative model and our recent work Higashi [113] that uses a hypergraph neural network architecture [114]. Both methods suggest better embedding results with Higashi being the first scHi-C embedding approach to demonstrate that the complex neuron subtypes in human prefrontal cortex can be revealed by chromatin conformation only. However, due to the computation-intensive nature of neural networks, the scalability of both methods has much room for improvement for large-scale datasets. For 3DVI, individual variational autoencoders are trained for each genomic distance and each chromosome, leading to thousands of deep neural network models to be trained. For Higashi, since the model treats each contact of scHi-C data as individual samples, it takes a long time to fully iterate over the dataset or to train the model till convergence. Crucially, methods for improving the interpretability of the embeddings for scHi-C data are particularly lacking, limiting our understanding of 3D genome structure-function connections for a diverse set of cellular phenotypes.

Here, we develop Fast-Higashi, an interpretable and scalable framework for embedding and integrative analysis of scHi-C data. We propose a concept for single-cell 3D genome analysis, called “meta-interactions” (analogous to the definition of metagenes in scRNA-seq analysis [52]), to improve the model interpretability. Our proposed Fast-Higashi algorithm jointly produces embeddings and meta-interactions for a given scHi-C dataset. Applications to various scHi-C datasets of complex tissues demonstrate that Fast-Higashi has overall comparable or even better embeddings than existing methods but is much faster than neural-network based methods (>40x faster than 3DVI and >9x faster than Higashi), enabling ultrafast delineation of cell subtypes or rare cell types in different biological contexts. Moreover, Fast-Higashi is able to infer critical chromatin meta-interactions that define cell types with strong connections to cell type-specific gene transcription. Fast-Higashi is the fastest and most scalable method for large-scale scHi-C data analysis to date.

3.2 Methods

3.2.1 Method details

The design of Fast-Higashi is based on a tensor decomposition model, called core-PARAFAC2 [115], and is generalized to simultaneously model multiple 3-way tensors that share only a single dimension (single cells). The core-PARAFAC2 model is usually used to analyze multimodal data where observations may not be aligned along one of its modes. A concrete example in other applications is the electronic health records that contain multimodal phenotypes of multiple patients at various time points. Because a particular disease stage may begin at different time points and may have varying lengths across patients, a critical difficulty is that it is hard to align observations of different patients along the temporal dimension. Similarly, in scHi-C contact maps, TAD-like structures usually have varying sizes and boundaries in different genomic bins, obscuring the direct alignment of genomic bins. Therefore, we have developed Fast-Higashi based on core-PARAFAC2 to address this issue. In the following sections, we first introduce how Fast-Higashi performs tensor decomposition on the scHi-C datasets assuming that contact maps of only one chromosome are present. We discuss next how we generalize to multi-chromosome cases. We then derive the optimization procedure and introduce the partial random walk with restart (Partial RWR) module to address the sparseness of single-cell Hi-C dataset efficiently.

3.2.2 Problem formulation of the Fast-Higashi model

For a scHi-C dataset, let $\boldsymbol{\mathcal{C}}$ denote the set of chromosomes. We formulate a collection of scHi-C contact maps of chromosome $c \in \boldsymbol{\mathcal{C}}$ as a 3-way tensor, denoted by $X^{(c)} \in \mathbb{R}^{N_c \times L_c \times M}$, where N_c is the number of genomic loci (also denoted as genomic bins) in chromosome c, L_c is the number of features at each bin, and M is the number of cells in this dataset. In principle, L_c need not be equal to N_c because, for example, we may use different resolutions for genomic bins along the two dimensions and even include additional epigenomic features. However, for convenience, here we only consider contact maps and use the same resolution for both dimensions. We assume that $X^{(c)}$ follows a 3-way core-PARAFAC2 model which includes: (1) a 3-way tensor $B^{(c)} \in \mathbb{R}^{N_c \times L_c \times r_c}$ of r_c meta-interactions; (2) a matrix $A^{(c)} \in \mathbb{R}^{N_c \times r_c}$ of bin weights indicating importance for each bin in every meta-interaction; (3) a chromosome-specific transformation matrix $D^{(c)} \in \mathbb{R}^{R \times r_c}$; and (4) an orthogonal matrix $V \in \mathbb{R}^{M \times R}$ that contains cell embeddings and is shared across all chromosomes, where r_c and R are hyperparameters, representing the dimensions of the chromosome-specific cell embedding and shared cell embedding.

We first introduce the cell-wise form of our model. As shown in Fig. 3.1a, the ℓ-th slice of $X^{(c)}$ along the last dimension, denoted by $X^{(c)}_{\cdot,\cdot,\ell}$, is the ℓ-th single-cell contact map, and we assume that it can be approximated by the weighted sum of meta-interactions:

$$X^{(c)}_{\cdot,\cdot,\ell} = \sum_{k=1}^{r_c} \mathrm{Diag}(A^{(c)}_{\cdot,k}) \times B^{(c)}_{\cdot,\cdot,k} \times (VD^{(c)})_{\ell,k} + E^{(c)}_{\cdot,\cdot,\ell}, \tag{3.1}$$

where $E_{\cdot,\cdot,\ell} \in \mathbb{R}^{N_c \times L_c}$ is a matrix of i.i.d. Gaussian noises with zero mean and arbitrary variance. Since V is the cell embedding matrix shared across all chromosomes, right multiplying V by the chromosome-specific transformation matrix $D^{(c)}$ projects V to another space, which we term the chromosome-specific embedding space. The chromosome-specific embeddings $VD^{(c)}$ directly quantify the contribution of each meta-interaction to single-cell contact maps, i.e., the overall weight of the k-th meta-interaction in cell ℓ is equal to $(VD^{(c)})_{\ell,k}$. Additionally, we also assume that bins in a meta-interaction may have different weights, i.e., the weight of bin i in the k-th meta-interaction is $A^{(c)}_{i,k}$. Together, the weight of the k-th meta-interaction at the i-th bin in cell ℓ is equal to the product of (1) the meta-interaction weight in the chromosome-specific embedding of cell ℓ and (2) the bin weight in the bin weight matrix, i.e., $(VD^{(c)})_{\ell,k} \cdot A^{(c)}_{i,k}$.

To simplify the optimization problem, we introduce an alternative bin-wise form of this model [116]. Let $X^{(c)}_i \in \mathbb{R}^{L_c \times M}$ ($i \in [N_c]$) be the i-th slice along the first dimension of $X^{(c)}$, i.e., the features of the i-th bin across all cells, and we use similar notations for other tensors. We assume that $X^{(c)}_i$ has the following decomposition:

$$X^{(c)}_i = B^{(c)}_i \times \mathrm{Diag}(A^{(c)}_i) \times D^{(c)\top} \times V^\top + E^{(c)}_i, \tag{3.2}$$

where $E_i \in \mathbb{R}^{L_c \times M}$ is a noise matrix. Since the noise is assumed to follow i.i.d. Gaussian distributions, the optimal set of parameters is the solution to the following optimization problem:

$$\underset{\substack{\forall c,\, B^{(c)}, A^{(c)}, D^{(c)} \\ V}}{\arg\min} \sum_{c \in \boldsymbol{\mathcal{C}}} \sum_{i=1}^{N_c} \| X^{(c)}_i - B^{(c)}_i \times \mathrm{Diag}(A^{(c)}_i) \times D^{(c)\top} \times V^\top \|^2_F \tag{3.3}$$

3.2.3 Additional constraints for uniqueness

Now we introduce additional constraints to address the uniqueness issue of Eqn. 3.3 and to improve the ability of Fast-Higashi to capture critical topological patterns in single-cell Hi-C contact maps.

Without loss of generality, we show the uniqueness issue on a dataset with only one chromosome, denoted by c. Let $\left(B^{(c)}, A^{(c)}, D^{(c)}, V\right)$ be one optimal solution to Eqn. 3.3. Then for any $P^{(c)} \in \mathbb{R}^{r_c \times r_c}$ and $S^{(c)} \in \mathbb{R}^{N_c \times r_c}$,

$$\left(\left\{ B^{(c)}_i \,\mathrm{Diag}(A^{(c)}_i)(P^{(c)})^{-1} \,\mathrm{Diag}(S^{(c)}_i)^{-1} \right\}_{i=1}^{N_c}, S^{(c)}, D^{(c)} P^{(c)\top}, V \right)$$

is also an optimal solution to Eqn. 3.3, implying the non-uniqueness. To address this, we impose constraints on the Gram matrices of $B^{(c)}_i$ that,

$$B^{(c)\top}_i B^{(c)}_i \equiv B^{(c)}_0, \forall i \in [N_c] \tag{3.4}$$

where $B_0^{(c)} \in \mathbb{R}^{r_c \times r_c}$ is constant over i. This is equivalent to requiring that $B_i^{(c)}$ can be transformed to each other by left multiplying an orthogonal matrix, i.e., a rotation along the feature dimension. Note that $B_0^{(c)}$ is not determined *a priori* and is optimized during inference. Similarly, for any $Q \in \mathbb{R}^{R \times R}$, solution $\left(B^{(c)}, A^{(c)}, Q^{-1}D^{(c)}, VQ\right)$ is also equivalent to $\left(B^{(c)}, A^{(c)}, D^{(c)}, V\right)$, implying the non-uniqueness. To address this, we require V to be orthogonal. The scaling of tensors $B^{(c)}$, $A^{(c)}$, and $D^{(c)}$ also leads to non-uniqueness and is addressed in Eqn. 3.11.

These constraints enable Fast-Higashi to be less prune to noise and allow Fast-Higashi capture critical topological patterns from single-cell Hi-C contact maps. A concrete example is the TAD-like structure where the number of interactions within this region is expected to be higher but also non-uniform, in that, the near-diagonal elements usually include more interactions. The characteristics of a TAD-like structure cause the boundaries of this TAD-like structure to be the same for all bins in it but cause the location of the peak to vary across these bins. This indicates that it is impossible to directly find a pattern that fits more than one bin in this TAD-like structure. However, since we allow bin-specific rotations along the feature dimension, these rotations potentially can keep the boundary unchanged and redistribute the contacts among the features of each bin, allowing the shift of the peak. Matrices $A^{(c)}$ are designed to capture the other bin-to-bin variability in scHi-C datasets. For example, bins usually have varying accessibility, which leads to different row sums in single-cell contact maps, and even in bins from one TAD-like structure. This variability is expected to be biologically meaningful and cannot be corrected by normalization. In Fast-Higashi, the bin weight matrix $A^{(c)}$ will capture this variability. In addition, a single bin may also show cell type-specific accessibility, which is expected to be reflected as variation across meta-interactions in Fast-Higashi. In Fast-Higashi, these bin-specific and cell type-specific characteristics will be captured in the bin weight matrix $A^{(c)}$ so that (1) any two bins may have different scaling factors in one meta-interaction and (2) one bin may have different scaling factors in any two meta-interactions. Therefore, these constraints retain the ability of capturing critical structures but also reduce the parameter spaces of

Fast-Higashi, making it more robust to tolerate noise.

3.2.4 Efficient parameter inference in Fast-Higashi

Here we show key steps in the derivation of a coordinate descent optimization procedure (summarized in Algorithm 1) for the optimization problem in Eqn. 3.3. We also introduce necessary tricks for a GPU-compatible algorithm.

Reformulation of the optimization problem

To simplify the optimization, we express $B_i^{(c)}$ as $U_i^{(c)}\bar{B}^{(c)}$ where $U_i^{(c)} \in \mathbb{R}^{L_c \times r_c}$ is orthogonal, and $\bar{B}^{(c)} \in \mathbb{R}^{r_c \times r_c}$. Both $U^{(c)}$ and $\bar{B}^{(c)}$ are model parameters and are optimized during inference. The relation between $B_0^{(c)}$ in Eqn. 3.4 and $\bar{B}^{(c)}$ is $B_0^{(c)} = \bar{B}^{(c)\top}\bar{B}^{(c)}$. With this reformulation, the optimization problem in Eqn. 3.3 can be rewritten as

$$\underset{\substack{\forall c,\, U^{(c)}, \bar{B}^{(c)}, A^{(c)}, D^{(c)} \\ V}}{\arg\min} \sum_{c \in \mathcal{C}} \sum_{i=1}^{N_c} \| X_i^{(c)} - U_i^{(c)} \times \bar{B}^{(c)} \times \mathrm{Diag}(A_i^{(c)}) \times D^{(c)\top} \times V^\top \|_F^2 \tag{3.5}$$

Derivation for the optimal solution of $U_i^{(c)}$ and V^*

Now we derive the optimal value of $U_i^{(c)}$ given the rest parameters. For the sake of simplicity, let T_U be $\bar{B}^{(c)}\,\mathrm{Diag}(A_i^{(c)})(VD^{(c)})^\top$ and the optimization of $U_i^{(c)}$ can be simplified as follows:

$$U_i^{(c)*} := \underset{U_i^{(c)}}{\arg\min}\ \|X_i^{(c)} - U_i^{(c)} T_U\|_F^2 = \underset{U_i^{(c)}}{\arg\min}\ \|U_i^{(c)} T_U\|_F^2 - 2\langle X_i^{(c)}, U_i^{(c)} T_U\rangle \tag{3.6}$$

$$= \underset{U^{(c)}}{\arg\min}\ \|T_U^\top\|_F^2 - 2\langle U_i^{(c)}, X_i^{(c)} T_U^\top\rangle = \underset{U^{(c)}}{\arg\max}\ \langle U_i^{(c)}, X_i^{(c)} T_U^\top\rangle, \tag{3.7}$$

where the second to last equality is true because $\|UT\|_F = \|T\|_F$ holds for any orthogonal matrix U. Since $U_i^{(c)}$ is orthogonal, the solution to this optimization has a closed form. Specifically, let the SVD of $X_i^{(c)} T_U^\top$ be $\tilde{U}_U \tilde{\Sigma}_U \tilde{V}_U^\top$, and the optimal solution of $U_i^{(c)*}$ is $\tilde{U}_U \tilde{V}_U^\top$. Note that, the optimal value of different frontal slices of $U^{(c)}$ can be solved in parallel.

The closed form of V^* can be derived similarly. Let $T_i^{(c)} := U_i^{(c)}\bar{B}^{(c)}\,\mathrm{Diag}(A_{i,:}^{(c)})D^{(c)\top}$, and then

$$V^* := \arg\min_V \sum_{c,i} \|X_i^{(c)} - T_i^{(c)} V^\top\|_F^2 = \arg\max_V \left\langle V, \sum_{c,i} X_i^{(c)\top} T_i^{(c)} \right\rangle, \tag{3.8}$$

which implies $V^* = \tilde{U}_V \tilde{V}_V^\top$ where $\tilde{U}_V \tilde{\Sigma}_V \tilde{V}_V^\top$ is the SVD of $\sum_{c,i} X_i^{(c)\top} T_i^{(c)}$.

Derivation for the optimal solution of $\bar{B}^{(c)}$, $A^{(c)}$, and $D^{(c)}$

Next, we derive the optimization of $\bar{B}^{(c)}$, $A^{(c)}$, and $D^{(c)}$.

Since $U_i^{(c)}$ and V are orthogonal, we can simplify the optimization in Eqn. 3.5 to

$$\arg\min_{\bar{B}^{(c)}, A^{(c)}, D^{(c)}} \sum_i \|\bar{B}^{(c)} \operatorname{Diag}(A_{i,\cdot}^{(c)}) D^{(c)\top} - U_i^{(c)\top} X_i^{(c)} V\|_F^2 \tag{3.9}$$

After we stack the r_c-by-R matrices $U_i^{(c)\top} X_i^{(c)} V$ to create a 3-way tensor $Y^{(c)} \in \mathbb{R}^{N_c \times r_c \times R}$, the optimization becomes:

$$\arg\min_{\bar{B}^{(c)}, A^{(c)}, D^{(c)}} \left\| \sum_{k=1}^{r_c} A_{\cdot,k}^{(c)} \otimes \bar{B}_{\cdot,k}^{(c)} \otimes D_{\cdot,k}^{(c)} - Y^{(c)} \right\|_F^2, \tag{3.10}$$

which is exactly the PARAFAC model and $\bar{B}^{(c)}$, $A^{(c)}$, and $D^{(c)}$ can be solved by alternative least square (ALS) [86]. To guarantee the uniqueness of the solution, we include an additional constraint that controls the scaling of the three factors:

$$\left\| \bar{B}_{\cdot,k}^{(c)} \right\|_2^2 = \left\| A_{\cdot,k}^{(c)} \right\|_2^2 \quad \text{and} \quad \left\| D_{\cdot,k}^{(c)} \right\|_2^2 = 1, \quad \forall k \in [r_c] \tag{3.11}$$

Mini-batch optimization

To improve the scalability of the method, we implemented the optimization of $U^{(c)}$ in a batch-wise manner. For a typical human scHi-C dataset of 10,000 cells, if we set the resolution to 500Kb, the 3-way dense tensor of chromosome 1 that is ready for tensor operations takes up to 6GB GPU RAM, which leaves inadequate RAM for subsequent computations on GPU. To utilize the computation power of GPU, we divide $X^{(c)}$ and $U^{(c)}$ into batches along the first dimension, and update all the frontal slices of $U^{(c)}$ from this batch in parallel using GPU. To minimize the data transfer amount between CPU and GPU, we compute the $T_i^{(c)}$ for the optimization of V in Eqn. 3.8 before we remove the copy of

Algorithm 1 Optimization procedure for Fast-Higashi

```
for chromosome $c$ do
    Initialize $A^{(c)}$ to be full of one
    Initialize $\bar{B}^{(c)}$ to be the identity matrix
    Flatten the first two dimensions of $X^{(c)}$, denote its $r_c$ right singular vectors by $(VD)^{(c)}$
end for
Concatenate all $(VD)^{(c)}$ along the rank dimension, and initialize $V$ as its top $R$ left singular vectors
Initialize $D^{(c)}$ as $V^\top (VD)^{(c)}$ for every chromosome $c$
for $1 \leq t \leq T$ do
    Update the value of $U^{(c)}$ by its closed form for each chromosome $c$
    Update the value of $V$ by its closed form
    Update $\bar{B}^{(c)}, A^{(c)}, D^{(c)}$ by alternative least square (ALS) until convergence for chromosome $c$
end for
```

this batch from GPU. Since r_c is much smaller than N_c in practice, the entire tensor $T^{(c)}$ fits in the GPU. Besides, we store these 3-way tensors $X^{(c)}$ in the COO format and transfer each batch of slices into GPU in the form of sparse COO tensors, which minimizes the data transfer as well as CPU memory usage. Hence, our method is optimized for GPU with limited RAM and data transfer rate to utilize its computation power and accelerate the overall running time.

3.2.5 Initialization of the Fast-Higashi model

Here we provide efficient initialization of model parameters based on their interpretations (Algorithm 1). We initialize the matrix $A^{(c)}$ to be full of one and the square matrix $\bar{B}^{(c)}$ to be the identity matrix, for each chromosome. For chromosome c, we find the SVD of the single-cell contact maps of chromosome c and keep the top r_c right singular vectors which are the initial cell embeddings of chromosome c. To aggregate information from multiple chromosomes, we concatenate the initial cell embeddings from all chromosomes and find its SVD. We initialize the meta embedding V to be one of the orthogonal matrix and $D^{(c)}$'s to contain the rest components in the SVD.

3.2.6 Embedded partial random walk with restart (Partial RWR)

To mitigate the sparseness of the scHi-C contact maps, we sought to incorporate the random walk with restart (RWR) data imputation method [40] into the Fast-Higashi framework. However, direct utilization of RWR before the tensor decomposition process is not desirable. The RWR imputed contact maps are usually much denser than the original contact map, leading to much higher memory consumption for storing the results and lower computational efficiency for transforming data format between sparse matrices to dense tensors as well as data transferring between GPU and CPU. Our solution is to integrate the RWR process during the optimization process of tensor decomposition and compute the RWR imputation batch by batch. The challenge for this design is that, as mentioned in the above section, the batch of the tensor decomposition optimization process is defined at the frontal slice of the tensor (Eqn. 3.3), i.e., the genomic bins, while the normal RWR requires the input of a complete graph adjacency matrix. To utilize RWR in our framework, here we propose the partial random walk with restart (Partial RWR) algorithm. The procedures of this algorithm are shown in Fig. 3.1b, which consists of the following steps: For simplicity, in this section, we use $X \in \mathbb{R}^{N\times N\times M}$ to represent the tensor representation of scHi-C contact maps of one chromosome. First, we fetch a small batch of the tensor $x(i) := X_{i:i+bs} \in \mathbb{R}^{bs\times N\times M}$ along the first dimension, where bs represents the batch size. Then, based on this small batch of tensor, we calculate the local affinity matrix $a(i,\ell) \in \mathbb{R}^{bs\times bs}$ of bins within this batch for each cell ℓ based on dot-product similarity:

$$p_{j,k,\ell} = \frac{x(i)_{j,k,\ell}}{\sum_{k'} x(i)_{j,k',l}} \qquad a^*(i,\ell) = p_{\cdot,\cdot,\ell} \cdot p_{\cdot,\cdot,\ell}^\top \tag{3.12}$$

After that, the standard RWR algorithm is applied to these local affinity matrices:

$$a^t(i,\ell) = (1-\rho)a^{t-1}(i,\ell)a^*(i,\ell) + \rho\mathbb{I} \tag{3.13}$$

where $a^0(i,\ell) = \mathbb{I}$, and ρ is the restart probability in the RWR algorithm. We denote the converged results of the RWR algorithm as $a^\infty(i,\ell) \in \mathbb{R}^{bs\times bs}$ and use it as the weight to propagate the information from the original batch of the tensor $x(i, \ell)$

$$y(i, \ell) = a^{\infty}(i, \ell) \cdot x(i, \ell) \tag{3.14}$$

Finally, we use $y(i, \ell)$ as the imputed results and pass it to the tensor decomposition optimization procedure. Our analysis showed that partial RWR can approximate the imputation of standard RWR well even with small batch sizes (see later section for details. In this work, we use batch size 64 to keep the balance between accuracy and computational efficiency.

3.2.7 Benchmarking scHi-C embedding methods

In this work, we mainly compared Fast-Higashi against three existing scHi-C embedding methods: Higashi [113], scHiCluster [40], and 3DVI [41] in terms of the quality of the generated embeddings and the runtime. We kept the embedding dimensions as the recommended ones for each method. The default hyper-parameters of Fast-Higashi sets R as 64, and r_c as $0.6N_c$. But due to the orthogonal property of V, one can always set R as a large enough number, and then use only the top-k dimensions. The final dimension number k can be determined using methods developed for selecting the number of principal components for scRNA-seq analysis [117]. For methods that allow selecting the maximum genomic distance to be considered, we set it to be 100Mb for all methods. We evaluated the embeddings generated by different methods under various evaluation metrics including: (1) Modularity score between the generated embeddings and the reference cell type label, (2) Adjusted rand index (ARI) and adjusted mutual information (AMI) score between the louvain clustering results and reference cell type label. Because the embeddings from different methods may reach the best clustering results at different combinations of parameters of Louvain clustering, we did a grid search for the number of neighbors and resolution parameters of the Louvain clustering for each method. The top 5 best clustering results for each method were kept and averaged as the final results. (3) We trained a logistic regression model using 10% of the cells and predicted the cell type for the rest 90% of cells. The Micro-F1 and Macro-F1 scores between the predicted cell type and reference ones were used to quantify the performance.

For the runtime analysis, all methods require different input formats and methods including 3DVI and Higashi can choose to only generate embeddings skipping the process of imputing sparse contact maps. To make a fair comparison, the runtime of all methods was calculated without the time of data processing, including transforming the scHi-C data into the format of a hypergraph in Higashi and reformatting the sparse contact maps into bands in 3DVI. For 3DVI and Higashi, we turned off the imputation function in the program and only used them to generate embeddings. For all methods, we added multiprocessing when possible even the multiprocessing was not originally implemented in some of the methods. Specifically, we parallelized the linear convolution and the random walk with restart algorithm of scHiCluster across all cells. We also parallelized 3DVI across different chromosomes, allowing the program to make full utilization of the GPUs. All methods were tested on a Linux machine with 1 NVIDIA RTX 2080 Ti GPU card, a 16-core Intel Xeon Silver 4110 CPU, and 252GB memory. All methods were set to use GPU when supported.

3.2.8 Aggregated single-cell A/B value

We developed the aggregated single-cell A/B value to collectively quantify the chromatin conformation at multiple loci in one cell. We calculated the scA/B value of every 500Kb genomic locus in single cells following the method proposed in Tan et al. [19]. We defined the scA/B value of one gene as the average of the scA/B values of the genomic loci spanned by that gene. Although Higashi software includes an algorithm to to calculate scA/B values based on its embeddings and imputations, to avoid potential analysis bias, we instead used the more orthogonal method from Tan et al. [19]. It is worth noting that in Tan et al. [19], by using the calculated scA/B values as embeddings, the observed refined clustering results in Fast-Higashi embeddings do not exist. To summarize the behavior of a group of genes, we defined the aggregated scA/B value of these genes in one cell as the average scA/B value across all genes in that cell. To systematically assess the differential expression of a group of genes between two sets of cells, we examined the difference in the distribution of aggregated scA/B value of these genes between the two sets of cells by t-test.

3.3 Results

3.3.1 Overview of Fast-Higashi

Fig. 3.1a illustrates the overall architecture of Fast-Higashi, which is an interpretable model for scHi-C analysis. In Fast-Higashi, scHi-C contact maps from different chromosomes are represented as multiple three-way tensors. Then a tensor decomposition model is utilized and generalized to simultaneously model these 3-way tensors that share only a single dimension (single cells). The tensor decomposition model takes the tensor representation of scHi-C data as input and decomposes the tensors into multiple factor matrices (Fig. 3.1a) to jointly infer cell embeddings as well as meta-interactions. These meta-interactions manifest the aggregated patterns of chromatin interactions, which are analogous to the concept of metagenes in scRNA-seq analysis. Each meta-interaction corresponds directly to a specific dimensions of the cell embeddings, providing a direct solution to interpret the association between embedding results and 3D genome features. We derived the mini-batch optimization procedure for the tensor decomposition model such that it can efficiently model tensors with drastically different sizes and effectively scale to scHi-C datasets with a large number of cells or at high resolutions. To mitigate the sparseness of the scHi-C contact maps while keeping the advantages of mini-batch training, we proposed a partial random walk with restart algorithm (Partial RWR, Fig. 3.1b) that efficiently imputes the sparse scHi-C contact maps before passing them to the tensor decomposition model. The detailed descriptions of the tensor decomposition model, the Partial RWR module, and the optimization procedures are in the Methods section.

3.3.2 Fast-Higashi achieves accurate and fast embedding of scHi-C data

We systematically evaluated the performance of Fast-Higashi for generating embedding vectors for various scHi-C datasets. To demonstrate the effectiveness of Fast-Higashi for delineating subtle cell-to-cell variability of 3D genome features, we applied it to three recent scHi-C datasets of complex tissues at 500Kb resolution. These datasets include the Tan et al. [19] dataset, the Lee et al. [20] dataset, and the Liu et al. [21] dataset . We

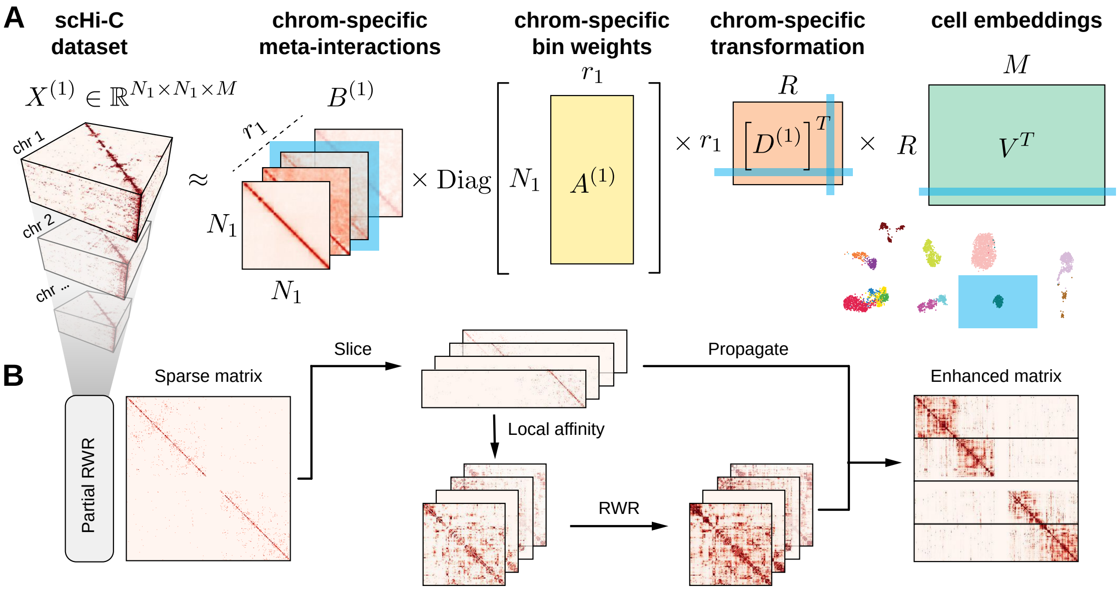

Figure 3.1: Overview of Fast-Higashi. **a.** Workflow of the Fast-Higashi algorithm. Given an input scHi-C dataset of k chromosomes, Fast-Higashi models it as k 3-way tensors. The tensor of chromosome c is denoted by $X^{(c)}$, where the first two dimensions correspond to genomic bins and the last dimension corresponds to the single cells. Fast-Higashi then decomposes the tensors $X^{(c)}$ into four factors: a set of meta-interactions ($B^{(c)}$), a genomic bin weights indicating importance for each bin ($A^{(c)}$), a cell embedding matrix V that is shared across all chromosomes, and a chromosome-specific transformation matrix $D^{(c)}$ that transforms the shared cell embeddings into chromosome-specific ones. **b.** Workflow of the partial random walk with restart (Partial RWR) algorithm. The Partial RWR is integrated into the Fast-Higashi framework. When calculating the decomposed factors for frontal slices of the tensor $X^{(c)}$, the corresponding slices would be imputed through Partial RWR first. The imputation process includes the calculation of local affinity, standard RWR algorithm, and information propagation using both sliced tensor and the RWR imputed affinity matrix.

evaluated the performance of Fast-Higashi and baselines under various evaluation metrics including: (1) the modularity score, (2) the adjusted rand index (ARI) and adjusted mutual information (AMI) scores, and (3) the Micro-F1 and Macro-F1 scores . We made direct comparisons of Fast-Higashi against three scHi-C embedding methods, including two very recently developed scHi-C embedding methods, Higashi [113] and 3DVI [41] as well as scHiCluster [40] (which has been updated recently). It has been suggested that the updated scHiCluster can distinguish neuron subtypes better on the Lee et al. dataset while the earlier version of scHiCluster cannot achieve [20, 113].

As shown in Fig. 3.2a-c, the UMAP visualizations of the Fast-Higashi embeddings on these three datasets show clear clustering patterns consistently corresponding to the annotated cell types. Notably, we observed several major advantages of the Fast-Higashi embeddings compared to other methods. On the Lee et al. dataset of the human prefrontal cortex, based on the UMAP visualization of the embedding results, Fast-Higashi can re-

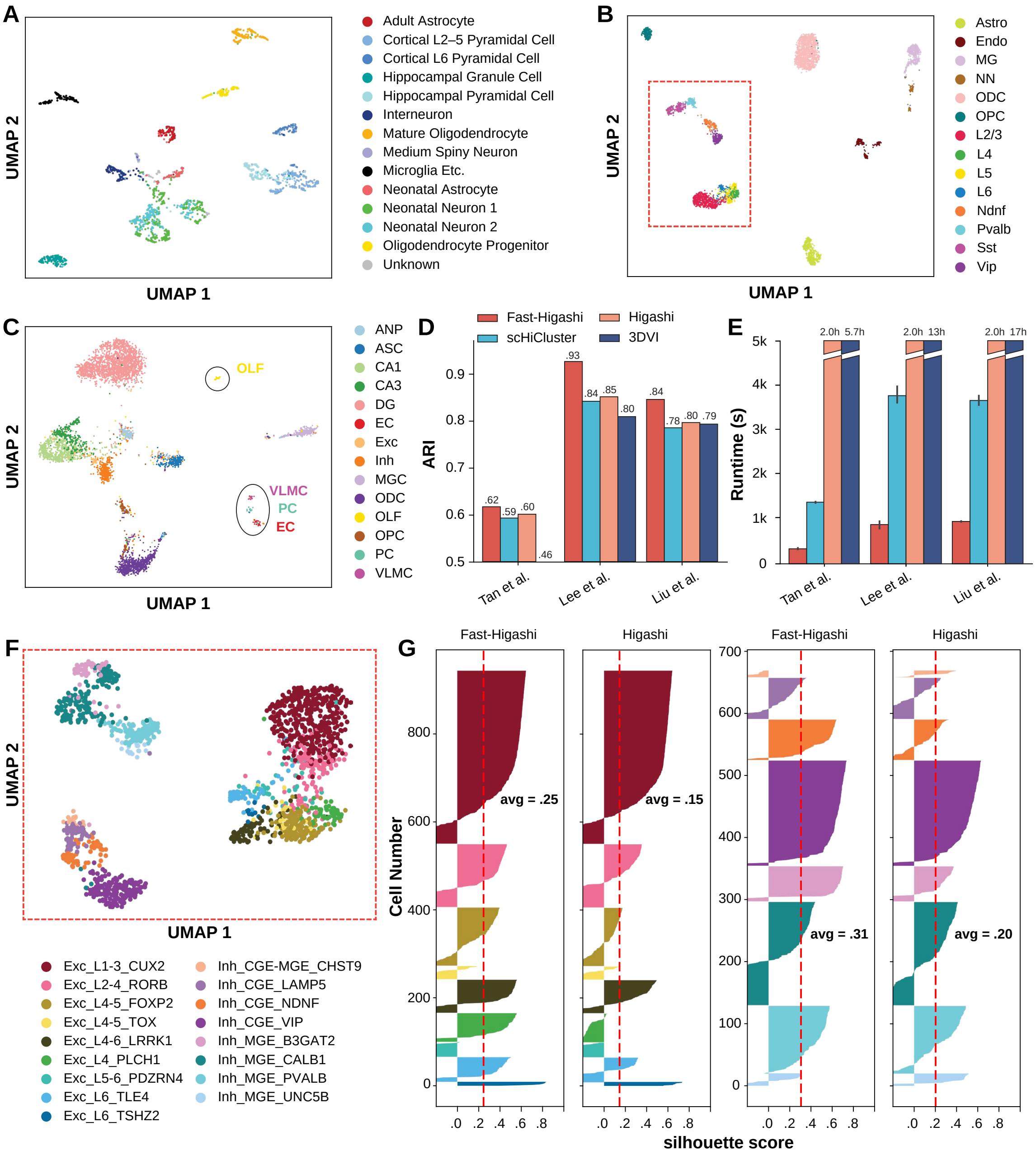

Figure 3.2: Evaluation of Fast-Higashi for generating embeddings for scHi-C data. **a.** UMAP visualization of the Fast-Higashi embeddings for the Tan et al. dataset [19]. **b.** UMAP visualization of the Fast-Higashi embeddings for the Lee et al. dataset [20]. Cells in the red box are neuron cells. **c.** UMAP visualization of the Fast-Higashi embeddings for the Liu et al. dataset [21]. **d.** Quantitative evaluation based on adjusted rand index (ARI) scores of the Louvain clustering results for each scHi-C embedding methods. **e.** Runtime of different embedding methods across different datasets. **f.** UMAP visualization of the Fast-Higashi embeddings for the neuron cells in the Lee et al. dataset (cells in the red box in **(b)**). Cell type information is from Luo et al. [118]. **g.** Quality of the embeddings for the neuron cells in the Lee et al. dataset measured as silhouette coefficients for neuron subtypes. All cell type abbreviations are consistent with the data source.

solve the differences among neuron subtypes clearly, separating all Pvalb, Sst, Vip, Ndnf, L2-3, L4, L5, and L6 neuron subtypes and showing more detailed structures within some cell types (Fig. 3.2b, marked in red box). To the best of our knowledge, this is the first time that excitatory neurons of different layers can be separated by using chromatin interaction information only. The embeddings of 3DVI separate neurons into two categories, excitatory neurons and inhibitory neurons, lacking the ability for more refined cell type delineation.

On the Liu et al. dataset of the mouse hippocampus, we again observed Fast-Higashi's clear advantage over other methods. Fast-Higashi is the only method that can separate CA3 cells from CA1 cells, and successfully identify small clusters of VLMC, PC, and EC cells, while all other methods (except Higashi) cannot (Fig. 3.2c).

All these observations are supported by our quantitative results, where Fast-Higashi consistently achieves the highest or second best scores across all metrics of all three datasets (Fig. 3.2d). We repeated the evaluation on two sci-Hi-C datasets with relatively lower coverage and/or a smaller number of cells and reached similar conclusions .

In addition, we assessed the runtime of all scHi-C embedding methods. As shown in Fig. 3.2e, Fast-Higashi is much faster than all existing scHi-C embedding methods, especially the neural-network based methods (>40x faster than 3DVI and >9x faster than Higashi on the scHi-C datasets used for benchmarking). The runtime of Higashi mostly depends on its number of training epochs and is almost constant for datasets with more than 1000 cells.

Together, these results demonstrate that Fast-Higashi achieves the state-of-the-art performance for scHi-C embeddings with an ultrafast computational efficiency.

3.3.3 Fast-Higashi enables the identification of rare cell types in complex tissues

In addition to the global evaluation on how well the Fast-Higashi embeddings correspond to the annotated cell types from the original datasets, we also sought to demonstrate that Fast-Higashi has unique capabilities to further improve the annotation of rare cell types in complex tissues.

We first visualized the Fast-Higashi embeddings of the neuron cells in the Lee et al.

[20] dataset using the UMAP projection (Fig. 3.2f). We obtained a new cell type annotation from Luo et al. [118], where the methylation profiles of the Lee et al. dataset were jointly embedded with single-cell methylation profiles from snmC-seq, snmCT-seq, and snmC2T-seq on human prefrontal cortex to annotate cell types. This joint analysis allows the characterization of neuron subtypes in the Lee et al. dataset at a much more refined resolution. Based on the UMAP visualization, we observed that the smaller clusters within the same cell type (red box in Fig. 3.2b) can in fact be delineated into more detailed cell subtypes. For instance, the two finer clusters of Sst in Fig. 3.2b correspond to the CALB1 and B3GAT2 cell subtypes in Fig. 3.2f. By comparing with the UMAP visualization of other embedding methods , we found that Fast-Higashi has the best ability to distinguish neuron subtypes, especially for the excitatory neurons. For inhibitory neurons, both Fast-Higashi and Higashi perform well and are the only two methods that can identify a smaller cluster of the UNC5B type . To further support these observations, we also evaluated each method's ability of separating neuron subtypes on the Lee et al. dataset through silhouette score analysis (Fig. 3.2g). Consistent with our observations based on the UMAP visualization, Fast-Higashi achieves the highest average silhouette score on the neuron subtypes.

We next systematically evaluated the robustness of Fast-Higashi's ability of identifying rare cell types, by simulating scHi-C dataset with different coverage. Specifically, we downsampled the contact pairs from the Lee et al. dataset to 10% to 50% of the original dataset and applied Fast-Higashi and Higashi, two models with strongest performance on these simulated datasets.

Additionally, we applied Fast-Higashi to the Tan et al. dataset of the developing mouse brain. Fast-Higashi is able to separate most of the cell types marked from the data source (Fig. 3.2a). In addition, we found two small clusters within the interneuron cell types and two separate clusters of neonatal neurons that do not correspond to the original Neonatal Neuron 1/2 labels from the dataset. With the observation on the Lee et al. dataset that the small clusters within a cell type could reflect more refined subtypes, we believe that this could also be the case for the Tan et al. dataset. Detailed results will be discussed in a later

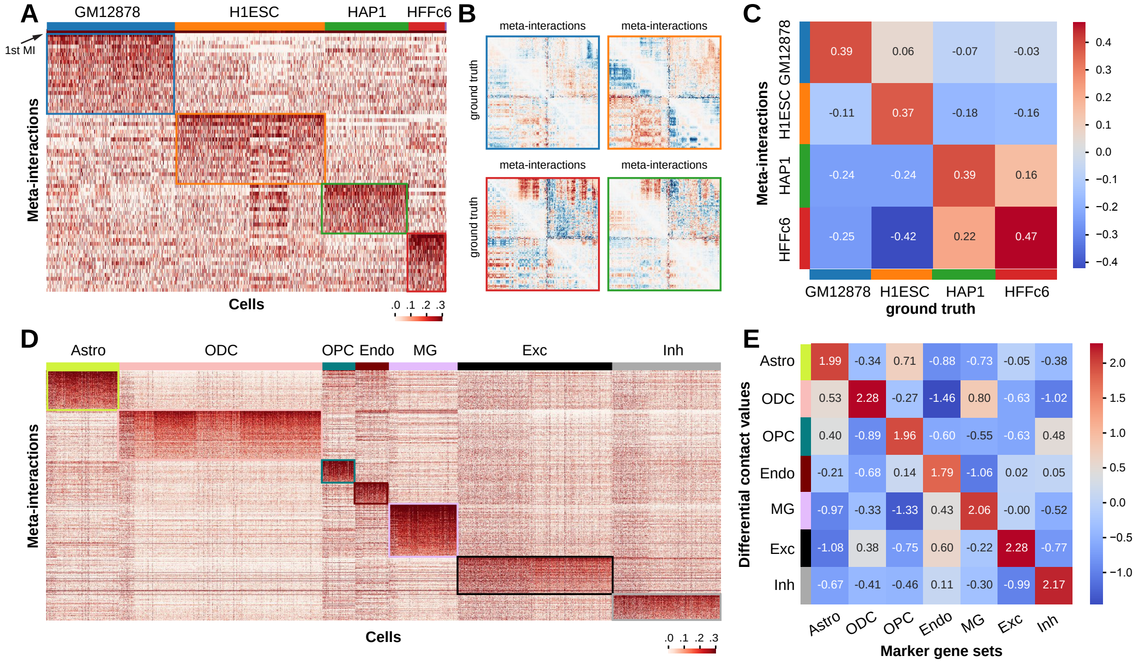

Figure 3.3: Analysis of the chromatin meta-interactions generated by Fast-Higashi. **a.** Heatmap of the single cell loadings for each meta-interaction of chromosome 1 for the Kim et al. dataset [17]. **b.** Visualization of the differential contact maps generated based on meta-interactions and those generated based on bulk Hi-C (marked with "ground truth"). Border color matches the cell type color in **(a)**. **c.** Spearman correlation between differential contact maps generated based on meta-interactions and those generated based on bulk Hi-C (marked with "ground truth"). **d.** Heatmap of the single cell loadings for each meta-interaction of the whole genome for the Lee et al. dataset [20]. **e.** Mean differential contact values of the lists of cell type marker genes averaged for each cell type. The mean differential contact values are calculated using the corresponding meta-interactions as the summation of values for each bin in the meta-interaction contact map. For each cell type, the top 200 marker genes were identified using Seurat [51].

section.

Taken together, these results confirm the unique capability of Fast-Higashi for identifying rare cell types or subtypes based on scHi-C data only.

3.3.4 Fast-Higashi effectively captures cell-type specific 3D genome structures

We then sought to demonstrate that the meta-interactions captured by Fast-Higashi reflect the cell type-specific 3D genome features and can be used to interpret the generated embeddings. As a proof-of-principle, we first analyzed the meta-interactions of chromosome 1 for the Kim et al. [17] dataset. In this section, we mainly focus on the 4 cell types with enough cell numbers, including GM12878, H1ESC, HAP1, and HFFc6. We first visualized the single cell loadings of these meta-interactions (chromosome-specific embeddings). As shown in Fig. 3.3a, each cell type has its preferred set of meta-interactions. Note that due to the utilization of SVD (singular value decomposition) for solving the meta-interaction during the optimization process, the first meta-interaction (sorted by the singular value during the SVD process) would correspond to the general contact patterns of all cells within the scHi-C data. This is consistent with the observation that for all cells, their loadings of the first meta-interaction (marked as “1st MI” in Fig. 3.3a) are large and similar across all cell types. For all other meta-interactions, they represent how the cell type-specific 3D genome features deviate from the population interaction patterns. To validate this, we aggregated the cell type-specific meta-interactions weighted by the average single cell loadings and made comparisons to the differential contact patterns calculated from the bulk Hi-C. Specifically, we first calculated a “common bulk Hi-C” as the average of the bulk Hi-C of the same four cell types. Then for each cell type we calculated the differential contact patterns as the difference between the bulk Hi-C of that cell type and the “common bulk Hi-C”. As shown in Fig. 3.3b, the differential contact patterns calculated using the bulk Hi-C share similar patterns to the aggregated cell type-specific meta-interactions. This observation is consistent with the phenomenon that the Spearman correlations between cell type specific meta-interactions and the corresponding differential contact map is the highest (Fig. 3.3c).

Next, we analyzed the meta-interactions of a scHi-C dataset on complex tissues, i.e., the Lee et al. [20] dataset. Fig. 3.3d shows the single cell loadings of the whole genome meta-interactions for this dataset. We again can observe a clear preference of meta-interaction sets for different cell types. To confirm that these cell type-specific meta-interactions manifest cell type-specific 3D genome features that are functionally relevant, we calculated the differential contact values for each bin given a specific set of meta-interactions. We first aggregated the meta-interactions for a specific cell type by the average single cell loadings, leading to one meta-interaction map of size $N \times N$ for each chromosome of one cell type. We then calculated the differential contact values by summing over the column of this meta-interaction map, representing the overall deviation of a genomic bin from its population-level pattern. By comparing to the marker genes called from scRNA-seq [119, 120], we found that there is a strong positive correlation between the differential contact values and the expressions of the marker genes (Fig. 3.3e).

These results demonstrate that the meta-interactions from Fast-Higashi effectively capture the cell type-specific 3D genome features that are relevant to cell type-specific gene regulation. The meta-interactions from Fast-Higashi can be used to associate the embedding results to a specific region of the scHi-C contact map, pointing to further investigation of differential 3D chromatin contact patterns of various cell types in complex tissues.

3.3.5 Fast-Higashi unveils single-cell 3D genome features in developing mouse brain

As discussed above, we applied Fast-Higashi to a scHi-C dataset of developing mouse brain, i.e., the Tan et al. [19] dataset, and observed local clusters of cells within the two annotated cell types in the UMAP visualization (Fig. 3.2a). We postulated that these local clusters could potentially be subtypes not captured by other scHi-C embedding methods as well as the original data source. To demonstrate that Fast-Higashi can delineate finer scale cell types and uncover developmental trajectories, we first obtained Fast-Higashi embeddings for all cortex cells and annotated the observed small clusters as Interneuron (A), Interneuron (B), Neonatal Neuron (A), and Neonatal Neuron (B) (Fig. 3.4a).

We sought to confirm our refined cell type labels of neonatal neurons and interneurons (highlighted with circles in Fig. 3.4a) by scA/B values in the gene bodies of marker genes. The marker genes were obtained from Tan et al. [19], which were calculated using Seurat [51] on the MALBAC-DT of the developing mouse brain. We quantified the A/B compartments of a set of differentially expressed genes (DEGs) by the aggregated scA/B value . Previous studies reported the existence of global correlations between the scA/B values and the gene expression level within the same cell type [68] and across different cell types in complex tissues [19, 113]. In particular, genes with higher scA/B values are more likely to be highly expressed. We found that the aggregated scA/B value of the marker genes of Pvalb and Sst is significantly higher in Interneuron (B) as compared to Interneuron (A) and the marker genes of Vip show the opposite trend (Fig. 3.4b). These results suggest

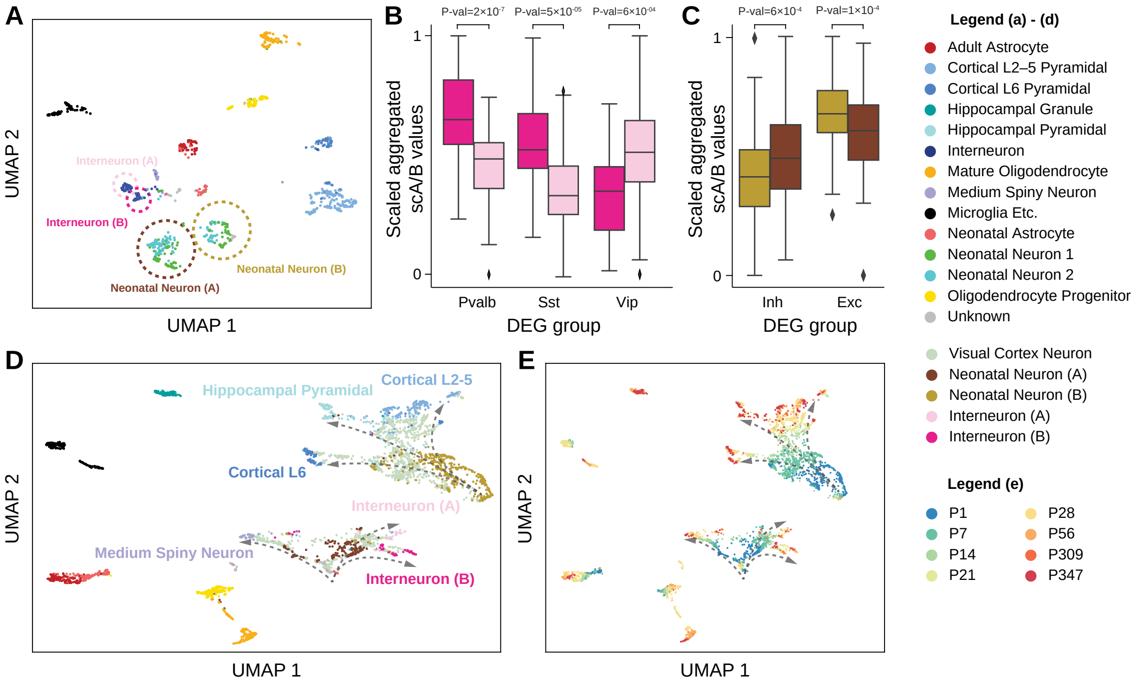

Figure 3.4: Application of Fast-Higashi to the scHi-C dataset from Tan et al. [19] of the mouse developing brain for more detailed identification of cell types and developmental trajectories. **a.** UMAP visualization of the Fast-Higashi embeddings for the cortex cells in the Tan et al. dataset. Cell subtypes identified by Fast-Higashi are highlighted with circles and texts. **b.** Distribution of the scaled aggregated single-cell A/B values for interneuron (A) and interneuron (B) subtypes identified by Fast-Higashi. For better visualization, the aggregated single-cell A/B values are linearly scaled to the range from 0 to 1 for each differentially expressed gene (DEG) group. **c.** Distribution of the scaled aggregated single-cell A/B values for Neonatal Neuron (A) and Neonatal Neuron (B) subtypes identified by Fast-Higashi. The scaling is the same as in panel **(b)**. **d-e.** UMAP visualization of the joint Fast-Higashi embeddings of visual cortex, cortex, and hippocampus scHi-C datasets in Tan et al [19]. The potential developmental trajectories from neonatal neurons to fully mature neurons are marked by dashed arrows. Note that here **(d)** is colored with cell type labels and **(e)** is colored with ages of the mouse.

that the DEGs of Pvalb and Sst neurons are expressed at a higher level in Interneuron (B) than in Interneuron (A) and that the DEGs of Vip neurons exhibit the opposite behavior, indicating that Interneuron (A) and (B) are more likely to be Vip and Pvalb/Sst, respectively. Similarly, the aggregated scA/B value of the marker genes of neonatal inhibitory neurons is higher in Neonatal Neuron (A) and the aggregated scA/B value of the marker genes of neonatal excitatory neurons is higher in Neonatal Neuron (B) (Fig. 3.4c), indicating the Neonatal Neuron (A) and (B) are neonatal inhibitory neurons and neonatal excitatory neurons, respectively. Meanwhile, the aggregated scA/B values of neonatal excitatory neurons do not show different distributions (P>0.2) between the Neonatal Neuron 1 and 2 in the original annotations from Tan et al. [19], confirming that our annotations are indeed a refinement. Collectively, we again demonstrate the advantage of Fast-Higashi in identifying finer cell types.

To further validate our refined subtype annotation, we jointly embed the cortex and hippocampus dataset of Tan et al. with another dataset of the visual cortex of developing mouse brain [19]. The cortex and hippocampus dataset consists of cells from mice at 6 ages: P1, P7, P28, P56, P309, and P347, while the visual cortex dataset includes cells of mice at ages P7, P14, P21, and P28, which covers the critical development period from P7 to P28 that was missed in the original dataset. When we applied Fast-Higashi to the union of these datasets, it recovered the complex developmental trajectories of inhibitory neurons and excitatory neurons. Specifically, in the UMAP visualization (Fig. 3.4d), a portion of the cells from the visual cortex dataset (light green) connect Neonatal Neuron (A) (Inhibitory neonatal neurons) to the 3 mature inhibitory neuronal types: Interneuron (A) (Vip), Interneuron (B) (Pvalb/Sst), and Medium Spiny Neuron. Similarly, a different set of visual cortex cells connect Neonatal Neuron (B) (Excitatory neonatal neuroins) to the multiple excitatory neuronal types: Cortical L2-5, Cortical L6, and Hippocampal Pyramidal. Since the Neonatal Neuron (A) and Neonatal Neuron (B) are primarily composed of P1/7 cells, and the 6 mature neuronal types consist of almost only P28 or older cells, placing the P14$\sim$28 cells between P1/7 cells and P28+ cells is consistent with the developmental process. Moreover, along the inferred developmental branches (Fig. 3.4e (curved arrows)), cells are indeed ordered by the mouse ages, strongly supporting the ability of Fast-Higashi in recovering trajectory from scHi-C datasets.

In summary, using the Tan et al. [19] dataset, we have demonstrated the clear advantages of Fast-Higashi in unveiling finer cell types over existing methods as well as the unique ability of Fast-Higashi to characterize the cell-to-cell variability of 3D genome features along complex biological processes.

3.4 Discussion

In this work, we developed Fast-Higashi, an ultrafast and interpretable framework for scHi-C data analysis. Our generalization from core-PARAFAC2 to Fast-Higashi not only leverages its strong scalability, but also enables joint and interpretable modeling of meta-interactions and cell embeddings. The development and incorporation of the Partial RWR algorithm further improve the performance of Fast-Higashi with negligible impact to the scalability. Evaluations of Fast-Higashi using a wide range of real scHi-C datasets have demonstrated its effectiveness and scalability for inferring informative cell embeddings, enabling the delineation of rare cell types and the reconstruction of developmental trajectories. Besides, as a proof-of-principle, we identified cell type-specific meta-interactions that are related to cell type-specific gene transcription. Together, we have demonstrated the effectiveness, scalability, and interpretability of our method Fast-Higashi.

By using its predecessor Higashi [113] as a direct baseline to compare, we demonstrated the superior scalability of Fast-Higashi to data size, robustness to data quality, and effectiveness in generating informative cell embeddings that facilitate rare cell type identification. Moreover, the unique scheme of meta-interactions allows direct analysis of cell type-specific 3D genome features that correspond to the embedding results. Even though Fast-Higashi has superior effectiveness and interpretability for scHi-C analysis, we note that Fast-Higashi is not developed to replace Higashi [113]. For instance, Fast-Higashi uses a random-walk-with-restart based method for imputing the sparse contact maps, which is efficient but also has limited imputation power. As demonstrated in Zhang et al. [113], the accurate imputation empowered by hypergraph representation learning is key to unveiling some important 3D genome features related to cell type-specific gene regulation. On the other hand, the underlying relationship between the tensor representation and the hypergraph representation of scHi-C data makes Fast-Higashi in some way a quasi-linear version of Higashi and can thus be used to initialize the Higashi model. As a proof-of-principle, we found that the Fast-Higashi initialized Higashi model can indeed achieve even better performance than any of these two methods . As one of the future directions, we plan to integrate

Fast-Higashi into the Higashi software suite, providing a more flexible and comprehensive framework for scHi-C analysis.

Fast-Higashi can be further enhanced by incorporating multimodal single-cell omics data, such as single-cell RNA-seq data and single-cell methylome data. Jointly modeling of co-assayed scHi-C data and other multimodal data has the potential to further improve cell embeddings and to establish connections between different modalities. Fast-Higashi may also be applied to study DNA-RNA interactions in single cells [121].

The continued development of scHi-C related technologies is expected to expand rapidly in the coming years. Fast-Higashi has the potential to become an essential method in the toolbox of single-cell 3D epigenomic analysis to greatly enhance the integrative investigation of 3D genome organization, genome functions, and cellular phenotypes at single-cell resolution for a wide range of biological applications.

Chapter 4

GAGE-seq concurrently profiles multiscale 3D genome organization and gene expression in single cells

4.1 Introduction

Connecting genotype to phenotype remains a challenge due to the complex principles governing genome functions. Mammalian genomes are organized within the three-dimensional (3D) space of the cell nucleus [10], featuring architectural structures across genomic scales, including chromosome territories [122], A/B compartments [6], subcompartments [6, 7], topologically associating domains (TADs) [8, 9] and subTADs [123, 124], and chromatin loops [125, 126]. These structures play critical roles for gene regulation, cellular development, and disease progression [1, 11, 12, 127–129]. Single-cell analyses provide unique insights into these processes, uncovering variability in 3D genome features in individual cells that bulk analyses might mask [1, 14, 58]. Yet, understanding how changes in multiscale 3D genome structure within a single cell influence its transcriptional program and cellular phenotypes remains a major challenge in epigenomics.

Cellular and molecular heterogeneity is pivotal in differentiation and tissue development. Advances in single-cell technologies, such as scRNA-seq and single-cell Hi-C (scHi-C), have deepened our understanding of cellular heterogeneity [130–132] and 3D genome organization [15, 16, 18, 19, 58–61]. To fully unravel the connections between 3D genome organization and transcriptional activity in individual cells, technologies that can concurrently measure both in the same cell are needed. Current computational approaches enable some integration of scHi-C and scRNA-seq [19, 113, 133], revealing connections between 3D genome organization and gene expression at cell-type level. However, such integration cannot capture the individual cell differences and cell-to-cell variation between structure and function, as it correlates data from separate cells. While imaging-based methods can provide simultaneous 3D genome organization and transcriptional activity within the same cells, they are constrained by low throughput and limited genomic coverage [68, 134–136]. These limitations underscore the need for high-throughput genomic technologies capable of co-assaying 3D genome and gene expression in the same cell.

Here, we report GAGE-seq (genome architecture and gene expression by sequencing), a highly scalable and cost-effective method for simultaneously profiling of chromatin interactions and gene expression in single cells. GAGE-seq, thanks to its combinatorial barcoding strategy, offers higher methodological throughput, as well as greater efficiency and effectiveness than recent technologies such as HiRES [137]. We applied GAGE-seq to profile 9,190 cells across diverse mammalian cell lines and tissues, including mouse brain and human bone marrow. Specifically, we developed an experimental and analytical framework for elucidating the connections between multiscale 3D genome features and cell type-specific gene expression, as well as their spatial and temporal interplay.

4.2 Methods

4.2.1 GAGE-seq data processing workflow

Demultiplexing.

DNA and RNA reads were assigned to wells based on the two rounds of barcodes. For DNA reads, only read 2 was used for demultiplexing, allowing at most 1 mismatch in each of the two rounds of barcodes. DNA reads with more than 5 mismatches in the region between the two rounds of barcodes (the 9th-23rd nt) were discarded. After demultiplexing, the first 12 nt were removed from read 1 and the first 35 nt were removed from read 2.

For RNA reads, only read 1 was used for demultiplexing, allowing at most 1 mismatch in each barcode round. RNA reads with more than 6 mismatches in the region between the two rounds of barcodes (the 19th-48th nt) or with more than 6 mismatches in the region downstream of the first round of barcode (the 57th-71th nt) were discarded. The two reference genomes were combined into a single reference genome file used for all GAGE-seq libraries. For DNA reads, BWA (0.7.17) was used for alignment. The combined reference genome was indexed using command bwa index -a bwtsw. Paired, trimmed DNA reads were aligned to the combined reference genome using command bwa mem -SP5M. For RNA reads, STAR (2.7.8a) was used for alignment. The GENCODE annotation files for human (v36) and mouse (vM25) were downloaded and concatenated. The combined reference genome was indexed using command –runMode genomeGenerate –sjdbOverhang 100 with the combined gencode annotation file. Only read 2 of RNA reads was aligned with the command STAR –outSAMunmapped Within.

Identification of contact pairs from DNA reads.

Pairtools (0.3.1.dev1) was used to identify contact pairs from paired DNA reads with command pairtools parse –walks-policy all –no-flip –min-mapq=10. After that, walk reads (i.e., DNA reads containing multiple ligation sites) were further processed. Briefly, we assumed that any pair of loci in the same DNA read forms a valid contact pair, and these contact pairs were included in the results.

Deduplication of contact pairs.

The contact pairs were deduplicated. We extract the genomic positions of the two ends of each contact pair. We define two contact pairs as directly duplicated if the two contact pairs' first ends lie within 500 nt apart and their second ends also within 500 nt. If two contact pairs are not directly duplicated, but are directly or indirectly duplicated with a third contact pair, we define the first two contact pairs as indirectly duplicated. Among each cluster (i.e., connected component) of (in)directly duplicated contact pairs, the one with the largest difference between its two ends' genomic positions was retained, and the rest were marked as duplicates.

Deduplication of RNA reads.

The RNA reads were deduplicated. Two RNA reads are defined as directly duplicated if there is at most 1 mismatch in their UMI and if their genomic positions differ by at most 5 nt. The rest of the process is similar to the deduplication of contact pairs. Only one RNA read from each duplicate cluster is retained.

4.2.2 GAGE-seq integrative analysis for mouse brain cortex.

Integration with MERFISH data.

Integration of GAGE-seq data and MERFISH data was done with Seurat (4.1.1). Only scRNA-seq profiles from the GAGE-seq data were used for this integration. In the GAGE-seq mouse brain cortex data, the following cell types of excitatory neurons were used: L2/3 IT CTX a, L2/3 IT CTX b, L2/3 IT CTX c, L4 IT CTX, L4/5 IT CTX, L5 IT CTX, L6 IT CTX, L6 CT CTX a, L6 CT CTX b, L5/6 NP CTX, and L6b CTX. In the MERFISH data, cells from L2/3 IT, L4/5 IT, L5 IT, L5/6 NP, L6 CT, L6 IT, and L6b were used. Each time, the selected cells from GAGE-seq were integrated with one slice from the MERFISH data. All genes detected and expressed in both GAGE-seq and MERFISH were used. The 'FindIntegrationAnchors' and 'IntegrateData' functions were used with default parameters, except that the number of dimensions was set to 10.

Inference of whole-transcriptome expression and 3D genome features for MERFISH cells.

The integrated single-cell expression profiles of GAGE-seq data and MERFISH data were scaled by the 'ScaleData' function from Seurat with default parameters, and the first 30 PCs were calculated by the 'RunPCA' function. A 50-nearest neighbor regressor was created to estimate whole-transcriptome expression and 3D genome features from the 30-dimensional PC space. The regressor was trained on GAGE-seq data and then applied to the MERFISH data. The Gaussian kernel was used as the weight function. For each MERFISH cell, the bandwidth was defined as the 0.3 quantile of the distances to the 50 nearest neighbors.

Integration with Paired-seq data.

The integration of GAGE-seq data with Paired-seq data [14] was done using Seurat. Only scRNA-seq profiles from the GAGE-seq data and the Paired-seq data were used for this integration. In the GAGE-seq mouse brain cortex data, we excluded three cell types: L2 IT RvPP, L2/3 IT RSP, and L5 IT RSP. In the Paired-seq data, cells from BR_NonNeu_Endothelial, HC_ExNeu_CA1, HC_ExNeu_CA23, HC_ExNeu_DG, HC_ExNeu_Subiculum, and HC_NonNeu_Ependymal were excluded. The 'SelectIntegrationFeatures', 'FindIntegrationAnchors' and 'IntegrateData' functions were used with default parameters.

Inference of accessibility for GAGE-seq cells.

The integrated single-cell expression profiles of GAGE-seq data and Paired-seq data were scaled by the 'ScaleData' function from Seurat with default parameters. The first 20 PCs were calculated by the 'RunPCA' function. To estimate whole-transcriptome expression and 3D genome features from the 40-dimensional PC space, we created a 50-nearest neighbor regressor, which was trained on Paired-seq data and then applied to the GAGE-seq data. The Gaussian kernel was used as the weight function. For each GAGE-seq cell, the bandwidth was set based on the 0.3 quantile of the distances to the 40 nearest neighbors.

4.2.3 GAGE-seq integrative analysis for bone marrow

Trajectory and pseudotime.

The pseudotime of human bone marrow cells was inferred by the 'sc.tl.diffmap' and 'sc.tl.dpt' function in Scanpy (1.9.3), jointly from the paired scRNA-seq profiles and scHi-C profiles. Specifically, cells in the HSC, MPP, MLP, and B-NK clusters were included. The first 5 PCs of the scRNA-seq profiles were used for the scRNA-based pseudotime and the first 2 PCs of the Fast-Higashi embeddings of the scHi-C profiles were used for the scHi-C-based pseudotime. The 5 scRNA-seq PCs and the 2 scHi-C PCs were then concatenated and used for the joint pseudotime. The 'sc.pp.neighbors' function was used to construct the neighbor graph with 30 (scRNA-based and joint pseudotime) or 20 (scHi-C-based pseudotime) nearest neighbors per cell. The 'sc.tl.diffmap' and 'sc.tl.dpt' function was applied with 10 diffusion components to learn a latent representation focusing on the trajectory and to infer the pseudotime for single cells. The origin of the trajectory was set based on the average expression level of HSC marker genes previously identified [138].

Unsupervised clustering of genes.

The clustering of genes was based on the expression and scA/B value. Genes expressed in at least 20 cells were included. To generate features for genes, 1) the expression levels and scA/B values were z-score normalized per gene among all cells. 2) cells were evenly divided into 10 bins based on the pseudotime, and 3) the average values of the expression and scA/B value in each bin were calculated for each gene. This process led to 20 features for each gene. The Louvain clustering algorithm was then applied to genes with 20 neighbors, a resolution of 1.5. The correlation was used as the distance metric.

4.3 Results

4.3.1 Overview of GAGE-seq

GAGE-seq is a high-throughput, effective, and robust single-cell multiomics technology that simultaneously profiles the 3D genome and transcriptome in individual cells (Fig. 4.1a). GAGE-seq leverages the highly scalable "combinatorial indexing" paradigm previously employed in sci-Hi-C [15, 17, 62, 139], as well as other single-cell methods [140–143] (Fig. 4.1a). The procedure can be summarized as follows: (i) The RNA in cross-linked and permeabilized cells or nuclei is reverse transcribed (RT) with a biotinylated poly(T) or random hexamer primer containing DNA sequences, facilitating the ligation of the first-round barcoded cDNA adaptors; (ii) Cross-linked chromatins are efficiently fragmented (the first round chromatin fragmentation) using two 4-cut restriction enzymes (RE), CviQI and MseI, both producing the same adhesive DNA end 5'-TA, enabling the identification of chromatin interactions via proximity ligation; (iii) After a second round of chromatin fragmentation to introduce adhesive DNA ends for ligating the first-round barcoded DNA adaptors, cells/nuclei are distributed to a 96-well plate, where the first-round barcodes for DNA

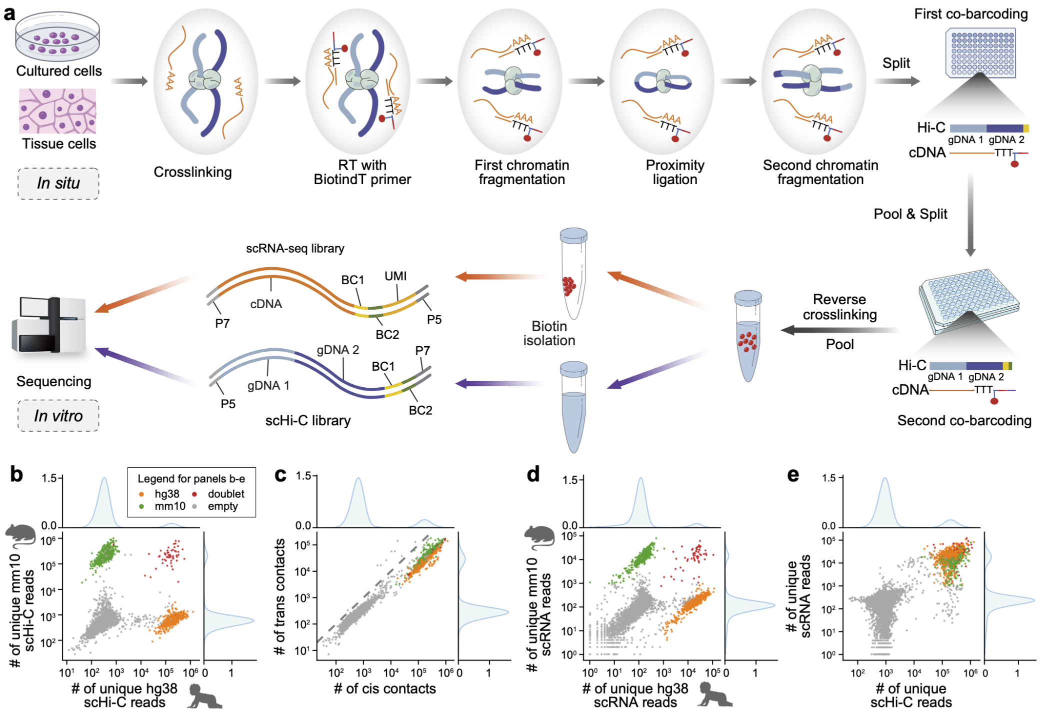

Figure 4.1: Overview and validation of GAGE-seq. **a.** Schematic representation of the GAGE-seq workflow detailing the simultaneous single-cell profiling of 3D genome architecture and gene expression. **b-e.** Validations demonstrating the specificity of GAGE-seq using mixed experiments with the human (K562) and mouse (NIH3T3). **b and d.** Scatter plots showing the collision level in the GAGE-seq scHi-C (b) and scRNA-seq (d) libraries, and histograms showing the binomial distribution of reads mapped to hg38 (top) and mm10 (right). **c.** Scatter plot showing the cis:trans ratio of scHi-C reads. **e.** Scatter plot showing the well-separation of scHi-C and scRNA reads of valid cellular indices from that of empty indices. Mouse is colored in green, human in orange, collisions in red, and empty indices in gray.

or cDNA are introduced through ligation of barcoded adaptors; (iv) Intact cells/nuclei are then pooled, diluted, and redistributed to a second 96-well plate, where the second-round barcodes for DNA or cDNA are introduced through ligation; (v) After reverse-crosslinking to release barcoded nucleic acids, all genomic DNA and cDNA are pooled, and biotinylated cDNA fragments are separated from genomic DNA with streptavidin beads; (vi) Sequencing libraries for scHi-C and scRNA-seq are separately generated and sequenced (Methods); and finally, (vii) Matched scHi-C and scRNA-seq profiles are identified according to the well-specific barcoding combinations (Fig. 4.1a). This combinatorial cellular indexing strategy can be further extended to achieve even larger throughput using additional rounds of ligation-mediated barcoding.

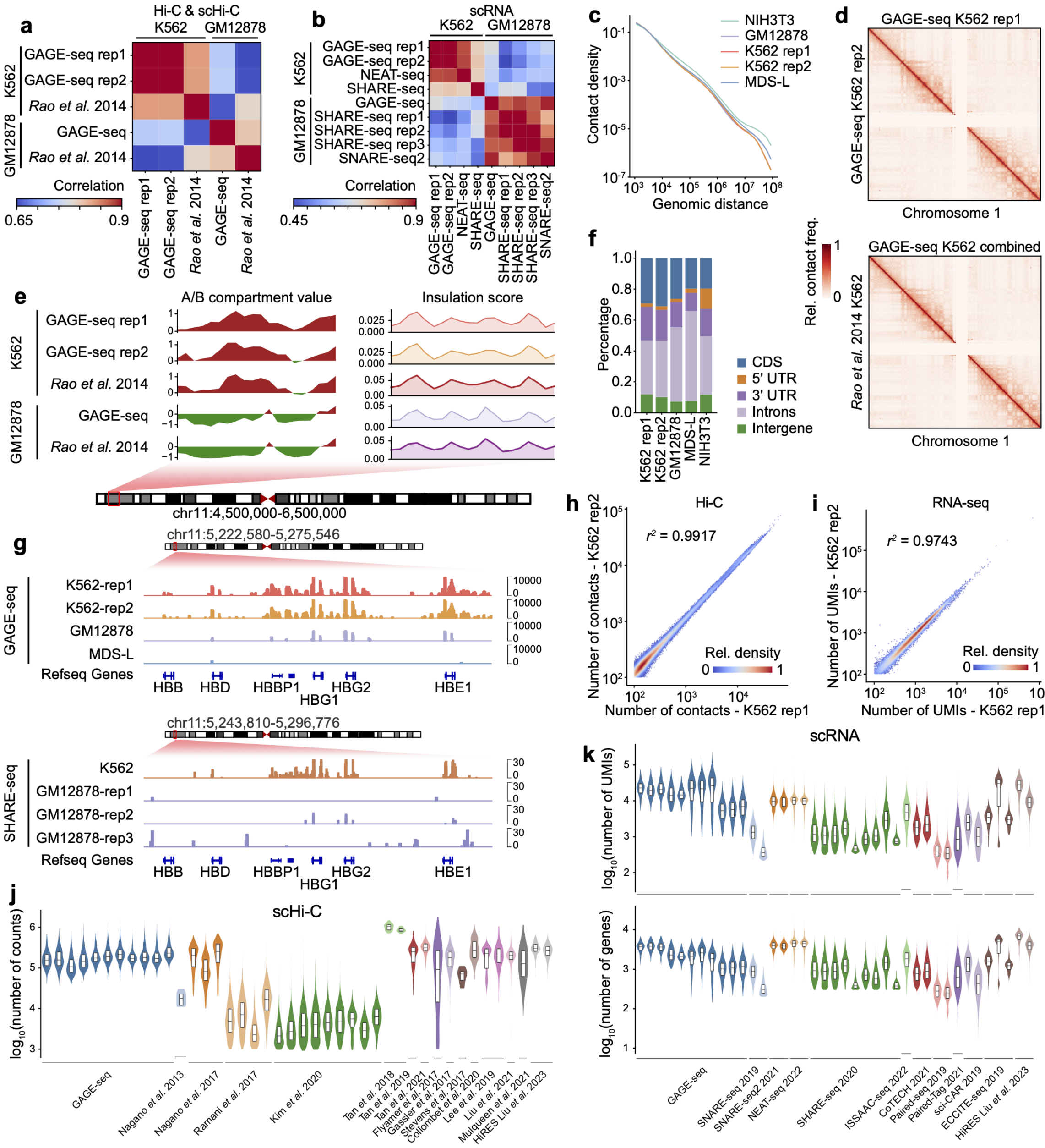

Figure 4.2: High-quality scHi-C and scRNA-seq data generated by GAGE-seq. **a.** Pearson's correlation between the aggregated scHi-C profiles from GAGE-seq replicates and the bulk in situ Hi-C data3. **b.** Comparison of aggregated scRNA-seq profiles of GAGE-seq replicates with NEAT-seq55, SHARE-seq43, and SNARE-seq256. Pearson's correlation is shown. **c.** Decay curves of chromatin contact for the GAGE-seq scHi-C libraries. **d.** Comparison of aggregated contact maps between two GAGE-seq K562 replicates (upper), and between the combined GAGE-seq K562 library and an in situ Hi-C library3 (lower). **e.** Comparison of A/B compartments and TAD-like domain calling at the human beta-globin locus between GAGE-seq (pseudo bulk) and in situ Hi-C3. **f.** RNA read distribution across gene bodies in the GAGE-seq scRNA libraries. **g.** Aggregated single-cell gene expression profiles at the GAPDH locus. Upper panel: scRNA-seq signals of GAGE-seq libraries of K562, GM12878, and MDS-L cells (hg38). Lower panel: scRNA-seq signals of SHARE-seq in GM12878 cells (hg19)43. **h.** Reproducibility between two biological replicates of GAGE-seq scHi-C libraries. **i.** Reproducibility between two biological replicates of GAGE-seq scRNA libraries. r2 statistics are shown. **j.** Comparison of GAGE-seq scHi-C library size with published scHi-C and co-assay methods **k.** Comparison of scRNA-seq library size (upper) and the number of detected genes (lower) with published co-assay methods.

4.3.2 Quality validation and benchmarking of GAGE-seq

To assess the quality and specificity of GAGE-seq data, we performed experiments using a mixture of human (K562) and mouse (NIH3T3) cell lines (Fig. 4.1b-e). Successful separation of human and mouse reads in both scHi-C and scRNA-seq data was demonstrated, identifying 683 human and 568 mouse cells out of 1,500 expected, along with 57 doublets observed in line with the expected 4.4% collision rate (Fig. 4.1b-e). Cells passing stringent quality criteria exhibited an average of 181,240 (K562, 39.2% duplicate rate) and 206,113 (NIH3T3, 38.0% duplicate rate) chromatin contacts (>1Kb intra-chromosomal) for scHi-C, as well as an average of 24,784 (K562, 35.7% duplicate rate) and 16,596 (NIH3T3, 31.2% duplicate rate) unique molecular identifiers (UMIs) from 3,699 (K562) and 2,256 (NIH3T3) genes per cell for scRNA-seq (Fig. 4.1). These robust results underscore GAGE-seq's ability to concurrently measure single-cell chromatin interactions and transcriptome with high sensitivity and accuracy. In addition, GAGE-seq's efficient fragmentation of crosslinked chromatin before proximity ligation, enabled by two four-cutters (Fig. 4.1a), allows for efficient detection of multi-way interactions, with >25% of all identified chromatin contacts in each scHi-C library.

Validating GAGE-seq in additional cell lines, GM12878 and MDS-L, further confirmed its robustness, specificity, sensitivity, and reproducibility (Fig. 4.2). Whole-genome and whole-library level analysis showed GAGE-seq's chromatin interaction and gene expression profiles strongly correlating with published datasets (Fig. 4.2a-b). Low collision rate (Fig. 4.1b), binomial distribution of scHi-C reads (Fig. 4.1b), typical chromatin contact decay curve (Fig. 4.2c), high cis-trans ratio (Fig. 4.1c), and aggregated pseudobulk and single-cell chromatin contact maps (Fig. 4.2d), as well as pseudobulk and single-cell A/B compartment scores and insulation scores (Fig. 4.2e), further confirmed the specificity of the GAGE-seq scHi-C signals. The specificity of the GAGE-seq scRNA-seq signals was demonstrated through low collision rate (4.6% in the K562/NIH3T3 library) (Fig. 4.1d), binomial distribution of RNA reads (Fig. 4.1d), and the fact that the majority of RNA reads (86%) mapped to the gene body (Fig. 4.2f), complemented by the pseudobulk and single-cell RNA signal distribution at individual gene loci (Fig. 4.2g). Notably, similar to SHARE-seq [144], GAGE-seq scRNA-seq reads were found to be 25%-50% intronic (Fig. 4.2f), indicating enriched nascent RNA. The high reproducibility across replicates was demonstrated at multiple levels (Fig. 4.2a,b,d,e,g,h,i), and its methodological resolution (library complexity) of scHi-C matched existing lower-throughput, unimodal methods, such as Dip-C [19, 61], as well as sn-m3C-seq [20, 21] (Fig. 4.2j). GAGE-seq scRNA-seq data quality was also comparable to existing methods (Fig. 4.2k). In line with previous scHi-C studies [16, 62], GAGE-seq scHi-C data revealed cell cycle stages. Compared to the recent HiRES method [137], GAGE-seq offers major advantages in throughput, efficiency, and cost-effectiveness (Fig. 4.2j-k), as well as in resolving rare cell types in complex tissues.

4.3.3 GAGE-seq reveals complex cell types in mouse cortex

To demonstrate the utility of GAGE-seq in unveiling complex cell types based on single-cell 3D genome features and gene expression within a tissue context, we turned our focus to the adult mouse brain cortex, known for its cell type diversity. Applying GAGE-seq on cells from the mouse cortex (8-9 weeks old), we generated 3,296 high-quality joint single-cell profiles of chromatin interactions and transcriptomes. On average, each cell displayed 231,136 chromatin contacts (at $\sim$50% duplication rate), with 20,160 UMIs and 1,883 genes per cell ($\sim$59% duplication rate), in line with the adult mouse whole brain data from the recently published HiRES data.

Our GAGE-seq scRNA-seq data identified 28 known cell types across three major lineages in the mouse cortex, including 15 excitatory neuron subtypes, 8 inhibitory neuron subtypes, and 5 glial cell subtypes, such as astrocytes and oligodendrocytes (Fig. 4.3a-b). These cell identities were confirmed by unique marker gene expressions (Fig. 4.3b). Notably, GAGE-seq scRNA-seq data enabled the delineation of many rare neuronal subtypes not identified by HiRES [137], such as L5 PT CTX, Sncg, and Meis2 (Fig. 4.3a-b). Reanalysis of HiRES mouse brain data with Fast-Higashi [113] further confirmed the superior performance of GAGE-seq in identifying complex cell subtypes, despite a lower sequencing depth in GAGE-seq. Although 3D genome features are known to encode cell

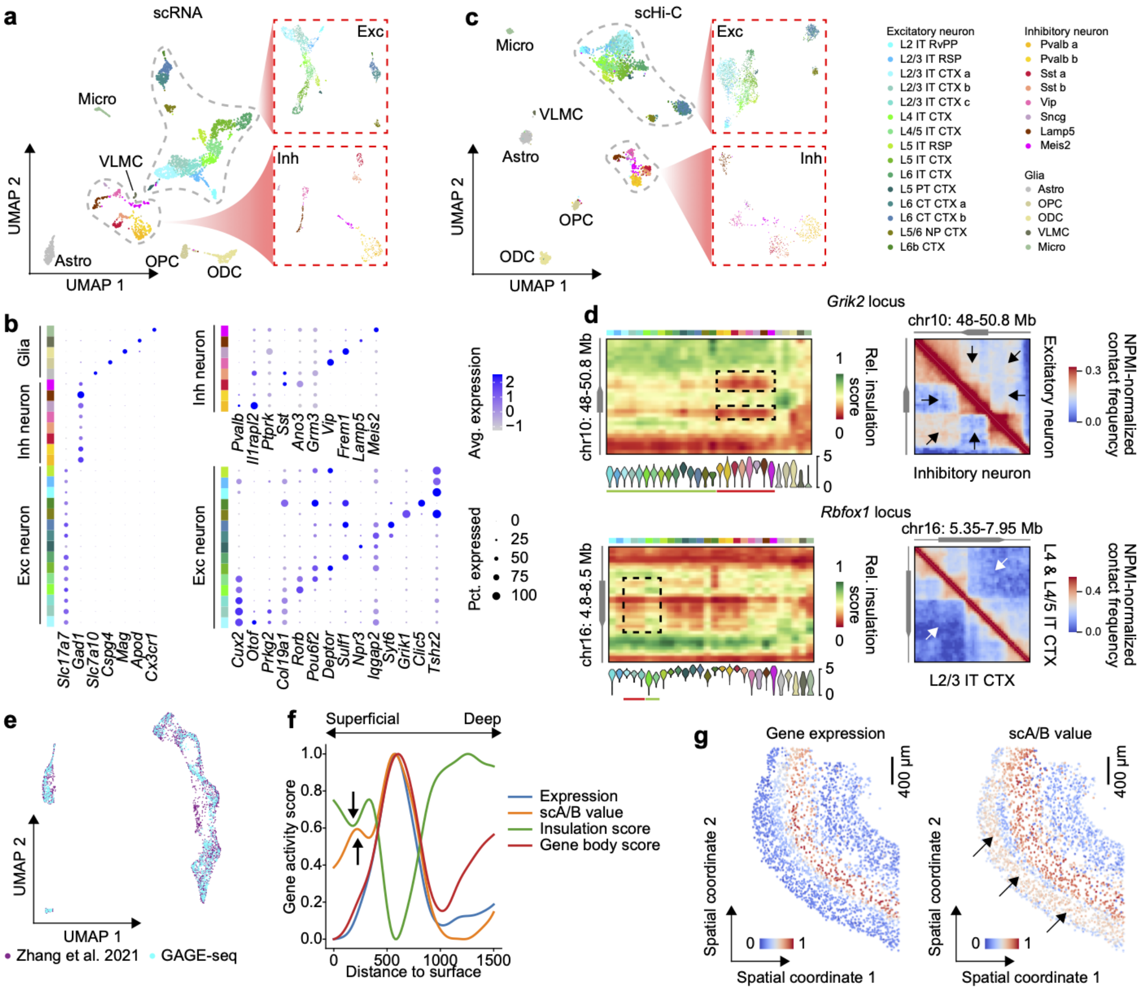

Figure 4.3: Cell types in mouse cortex characterized by GAGE-seq scHi-C and scRNA-seq. **a and c.** UMAP visualization of mouse cortex scRNA-seq (a) and scHi-C profiles (c) from GAGE-seq. Insets: UMAP visualization of excitatory neuron subtypes (top) and inhibitory neuron subtypes (bottom). **b.** Cell type-specific expression (based on scRNA-seq in GAGE-seq) of known marker genes, including glial types, neuronal types, and neuron subtypes. **d.** Visualization of cell type-specific 3D chromatin architecture and gene expression at representative gene loci. Left: aggregated single-cell insulation score (100-Kb resolution, upper) and gene expression (lower) at the Girk2 locus and the Rbfox1 locus. Right: aggregated contact maps (50-Kb resolution) of the Girk2 locus (top panel, excitatory vs inhibitory neurons) and the Rbfox1 locus (low panel, L4 & L4/5 IT CTX vs L2/3 CTX). Cell types selected in the right panels are highlighted by green lines (higher expression) or red lines (lower expression) in the corresponding left panels. **e.** UMAP visualization of the integration of GAGE-seq and a MERFISH dataset [72]. **f.** Inferred spatial patterns of gene expression and 3D genome features of L5 IT CTX marker genes. **g.** In situ plots of inferred single-cell gene expression (left) and scA/B value (right) for L5 IT CTX marker genes. Layer 3 was highlighted by black arrows in panels (f) and (g). The cell type abbreviations are based on the naming convention used in [145].

identity [137, 146], scHi-C often identified fewer cell types in complex tissues than scRNA-seq [19–21, 147]. Utilizing Fast-Higashi for scHi-C embedding, GAGE-seq distinguished all 28 transcriptome-defined cell types, including the aforementioned L5 PT CTX, Sncg, and Meis2 rare subtypes (Fig. 4.3c). The scHi-C-based delineation supports these cell types with distinct 3D genome features, with insulation scores surrounding gene bodies showing cell type-specific connection with gene expression (Fig. 4.3d; see later section with more analysis).

4.3.4 Spatial integration reveals in situ 3D genome variation

Using GAGE-seq to map the 3D genome and transcriptome of single cells, we explored the in situ variation of the 3D genome in the adult mouse cortex. We leveraged GAGE-seq scRNA-seq as a "bridge" for this analysis. Recently, the spatial transcriptomics method MERFISH successfully discerned the spatial organization of distinct cell populations in the mouse primary motor cortex [72]. We started by integrating our GAGE-seq scRNA-seq data with the MERFISH data using Seurat, allowing us to establish a connection between the two datasets (Methods).

We focused on the excitatory neuron cell types present in both GAGE-seq and MERFISH datasets. Within the integrated embedding space, cells primarily clustered by cell type, and cells from both datasets integrated cohesively, indicating high correlation between cell types identified by the two methods (Fig. 4.3e). We next characterized the in situ variation of both marker gene expression and 3D genome features of these maker gene loci in the mouse cortex. As a proof of principle, we investigated the in situ pattern of marker genes for L5 intratelencephalic (IT) CTX. The observed and inferred gene expression demonstrated a high degree of congruence, further supporting the reliability of the integration (Spearman's r=0.76, two-sided P=0). Layer 5, where L5 IT CTX cells reside, corresponded with the highest expression level, scA/B value [19], gene body score, and a low single-cell insulation score (Fig. 4.3f-g), reinforcing the overall correlation between expression and 3D genome structure. Interestingly, despite consistently low expression levels and gene body scores in more superficial layers, the scA/B value increased and the

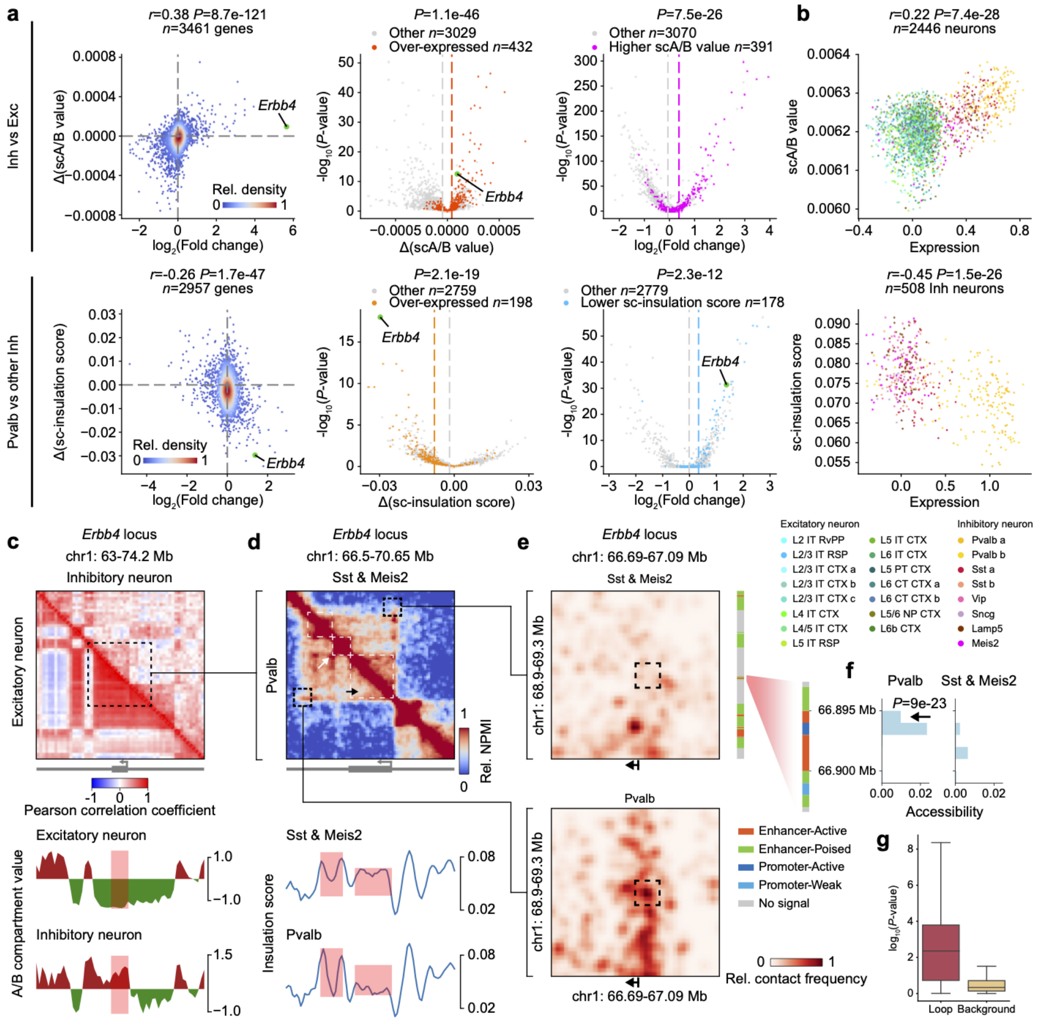

Figure 4.4: 3D genome features inform cell type-specific gene expressions in the mouse cortex. **a.** Correlations between gene expression and 3D genome features across neuron cell types. Upper row: inhibitory (n=508) vs. excitatory (n=1938). Lower row: Pvalb (n=188) vs. other inhibitory (n=320). Left column: correlation between differential expression and differential 3D genome feature (Pearson's correlation coefficients and the P-values from one-sided tests for nonzero correlations shown). Middle column: volcano plot of differential scA/B value and single-cell insulation score; Right column: volcano plot of differential expression. P-values from one-sided t-tests with unequal variance are shown in middle and right columns. **b.** Single-cell level correlation of gene expression with scA/B value (upper) or insulation score (lower) in inhibitory neurons (432 genes) and Pvalb (198 genes), respectively (Spearman's correlation coefficients and the P-values from one-sided tests for nonzero correlations shown). **c.** Comparison of A/B compartment (200-Kb resolution) of the Erbb4 locus between inhibitory and excitatory neurons. Pearson's correlation matrices of aggregated contact maps (top) and the A/B compartment scoretracks (bottom) are shown. **d.** Comparison of the pseudo-bulk contact map (50-Kb resolution) of the Erbb4 locus between Pvalb and other inhibitory subtypes. Pseudo-bulk contact maps (upper) and the insulation scores (bottom) are displayed. Two Pvalb-specific strides (white arrow) and melted TAD (black arrow) are shown in the top panel. The gene body is shown right under the contact matrices in (c) and (d), while the bottom panels highlight differential 3D genome features with light red boxes. **e.** Loop example in Pvalb (lower) and Sst and Meis2 (upper) inhibitory subtypes at 10-Kb resolution. Aggregated contact maps, regulatory element annotations52 (right), and TSS of Erbb4 (bittin arrow) are shown. **f.** Differential accessibility around the enhancer in Pvalb (left) vs. Sst and Meis2 (right), with a 1kb enhancer region highlighted (black arrow). The P-values of one-sided Mann-Whitney U tests are shown. **g.** Loop vs. non-loop contacts correlation with expression. P-values from two-sided tests for nonzero Spearman's correlation coefficients are shown (n=3,105 cells).

single-cell insulation score decreased slightly around layer 3, a cortical layer containing the L2/3 IT CTX cells that are not adjacent to the tissue boundary, suggesting potential discrepancies of expression and various 3D genome features at finer spatial resolution (highlighted by arrows in Fig. 4.3f-g).

4.3.5 Impact of 3D genome on gene expressions in single cells

We next rigorously examined the relationship between gene expression and various multi-scale 3D genome features in single cells, including A/B compartments, TAD-like domains, and chromatin loops.

Our analysis of the 3,461 genes expressed in inhibitory neurons (n=508) or excitatory neurons (n=1,938) revealed a strong correlation between cell type-specific gene expression and scA/B value, reflecting compartmentalization variations [19, 113] (Fig 4a, top panels). Inhibitory neurons, for instance, showed a much higher expression for 432 genes which corresponded to a higher scA/B value (t-test P=1.1e-46; Fig. 4.4a, top middle panel). Most of the 391 genes with a higher scA/B value in inhibitory neurons also snowed notably higher expression levels in these cells compared to excitatory neurons (t-test P=7.5e-26, Fig. 4.4a, top right panel). Overall, there is a significant correlation between differential gene expression and differential scA/B value (Pearson's r=0.38, P¡1e-100, Fig. 4.4a, top left). At the chromatin domain level, we identified a negative correlation between cell type-specific gene expression and the associated single-cell insulation score across cell types (Fig. 4.4a, bottom panels), suggesting that TAD-like domain variations around the gene body are accompanied with changes in transcriptional activity of the gene. This phenomenon, aligning with previous findings at the cell type level [113], may be attributed to domain melting noted in highly expressed long genes in mouse hippocampus and midbrain neurons47.

We subsequently examined the relationship between single-cell insulation score surrounding the gene body and the potential occurrence of domain melting within our diverse collection of cell types revealed by GAGE-seq. We focused on the four genes (Grik2, Dscam, Rbfox1, and Nrxn) known to undergo domain melting [146], profiling their scA/B value, single-cell insulation score, and single-cell gene expression. Notably, these genes manifested high expression across almost all 28 cell subtypes revealed by GAGE-seq, with the exception of Dscam and Grik2 in VLMC and Micro cells. As expected, Dscam, Rbfox1, and Nrxn3 were predominantly in the active A compartment in the majority of cell subtypes (Fig. 4.3d), while the Grik2 locus was in a weak B compartment across all the cells, despite its high expression. Aggregated single-cell insulation scores varied across the gene body, with most cell subtypes showing lower scores correlating with elevated gene expression (Fig. 4.3d). The aggregated chromatin contact maps indicate potential occurrence of domain melting around these gene bodies (Fig. 4.3d). A similar phenomenon was also detected for the Rbfox1 locus across different excitatory neurons (Fig. 4.3d, low panels).

We next further confirmed the above observed connection between multiscale 3D genome features and gene expression at single cell resolution. Higher gene expression in a cell often corresponded to a higher scA/B value and lower single-cell insulation score in the same cell (Fig. 4.4b). For instance, of the 432 genes showing a significantly elevated scA/B value in inhibitory neurons, most displayed higher expression in these neurons than in excitatory neurons (Spearman's r=0.22, P=7.4e-28, n=2446 cells; Fig. 4.4b, top panel). At the chromatin domain level, the 198 genes expressed highly in Pvalb cells exhibited notably lower single-cell insulation scores than in other inhibitory neurons (Spearman's r=0.45, P=1.5e-26, n=508 cells; Fig. 4.4b, low panel). Thus, the connection between multiscale 3D genome features and gene expression is evident at the single-cell resolution.

We then confirmed our observations on single loci. As a proof of principle, we focused on the Pvalb inhibitory subtype (including both Pvalb a and Pvalb b). We first selected genes that have 1) significantly higher scA/B values and expression in inhibitory neurons compared to excitatory neurons (Fig. 4.4a, top panels), and 2) significantly higher expression and lower single-cell insulation scores in Pvalb compared to other inhibitory neurons (Fig. 4.4a, bottom panels). This approach led us to the Erbb4 gene. The Erbb4 gene plays a pivotal role in the central nervous system and has been linked to schizophrenia [148]. As expected, we observed differential A/B compartment states correlated with cell type-specific expression of the Erbb4 gene (Fig. 4.4c), and differential single-cell insulation score that suggests domain melting in the gene locus (Fig. 4.4d, low panel). The TAD-like domain structure of the Erbb4 gene body in Sst and Meis2 cells appears to be melted in Pvalb cells (i.e., less pronounced), which is again accompanied with high gene expression in Pvalb cells (Fig. 4.4d, top panel). Additionally, it appears that the Erbb4 gene body interacts more frequently with the downstream two small TAD-like domains in Pvalb cells than in Sst and Meis2 cells (Fig. 4.4d, top panel). On a finer scale, we also observed a cell type-specific putative enhancer-promoter chromatin loop at the TSS of the Erbb4 gene in Pvalb cells (Fig. 4.4e-g). Moreover, when integrating with chromatin accessibility, the putative enhancer region exhibits differential chromatin accessibility that correlates with the cell type-specific expression of the Erbb4 gene (Fig. 4.4f).

4.3.6 Integrative analysis of GAGE-seq and chromatin accessibility

We next aimed to demonstrate how integrating GAGE-seq with chromatin accessibility data enhances the connection between CREs and target genes. For this, we integrated GAGE-seq with Paired-seq data (from the same mouse cortex region) [149]. Overall, genes with distinct contributions from 3D genome and chromatin accessibility show varied functions and integrating 3D genome and chromatin accessibility data markedly improves gene expression prediction accuracy.

Our integrative analysis of GAGE-seq and chromatin accessibility enhances the connection of CREs to their target genes. The gene expression and transcription start site (TSS)-CRE interaction frequency correlation decreases with greater genomic distance between TSS and CRE (Fig. 4.5a). Also, overlaps between Paired-seq-identified gene-CRE pairs and those identified by other approaches generally decrease with increasing genomic distance between TSS and CRE. However, refining with GAGE-seq data markedly improved this overlap, particularly for long-range (>100kb) gene-CRE pairs (Fig. 4.5b), highlighting the advantage of GAGE-seq in revealing CRE-gene pairs.

We also explored the joint regulation of gene expression by 3D genome and chromatin accessibility at individual gene loci. A strong correlation was found between Epha4 gene expression and the chromatin interaction frequency with a distal CRE, as well as between

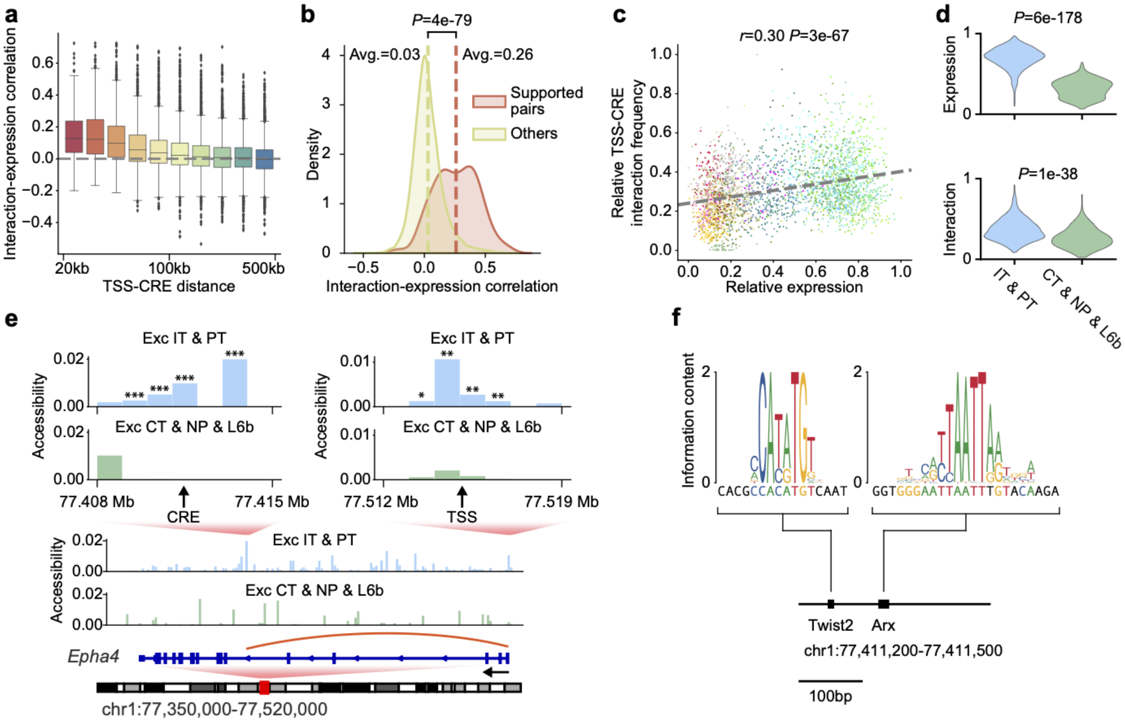

Figure 4.5: Integrative analysis of GAGE-seq and chromatin accessibility in the mouse cortex. **a.** Correlation coefficient (n=3,105 cells) between expression and TSS-CRE interaction frequency for each gene-CRE pairs from Paired-seq data , grouped by genomic distance between TSS and CRE. **b.** Comparison between gene-CRE pairs corroborated by other sources (red) and those identified only from Paired-seq data63 (yellow). The P-value of two-sided Mann-Whitney U test is shown. **c-e.** The combined effect of 3D genome and accessibility on expression at the Epha4 locus. **c.** Correlation of interaction-expression for a specific gene-CRE pair at the Epha4 gene, with dots representing single cells colored by cell type. **d.** Expression (upper) and TSS-CRE interaction frequency (lower) comparison among excitatory subtypes, revealing heightened levels in IT and PT subtypes. The P-values of one-sided Mann-Whitney U tests are shown. **e.** Accessibility comparison around the TSS and CRE (chr1: 77410959-77411960) of the Epha4 gene among excitatory subtypes, showing higher accessibility IT and PT subtypes. The P-values of two-sided Mann-Whitney U tests are shown. IT and PT subtypes are compared against CT, NP, and L6b subtypes in (d) and (e). In panel (e), *: P<1e-3; **: P<1e-5; ***: P<1e-10; the P-values in the upper left plot are (from left to right): 2e-11, 7e-20, 8e-34, 7e-52; the P-values in the upper right plot are: 6e-4, 6e-8, 7e-6, 2e-6, 1e-4. **f.** Binding sites of transcription factors Twist2 and Arx at the CRE of the Epha4 gene, depicting both the canonical motif (top) and the identified binding motif sequence (bottom) for each TF.

Epha4 gene expression and chromatin accessibility at the TSS and the distal CRE in different excitatory neuron subtypes (Fig. 4.5c-e). Motif analysis of chromatin accessibility peaks identified potential binding sites for transcription factors Twist2 (Spearman's P=1e-289) and Arx (Spearman's P=2e-132) (Fig. 4.5f). However, no significant differences were noted for A/B compartment value, insulation score, and gene body score of the Epha4 locus across neuron subtypes, indicating that fine-scale CRE-chromatin looping instead of changes in the large-scale 3D features may be responsible for the cell type-specific Epha4 expression.

4.3.7 Developmental stages of human hematopoiesis

Hematopoiesis is a classic model system with well-characterized cell type changes and their associated gene expression signatures, making it an ideal model for exploring the dynamic relationship between 3D genome structure and gene expression. We generated GAGE-seq profiles of 2,815 human bone marrow (BM) CD34+ cells after stringent quality filtering, obtaining an average of 265,336 chromatin contacts (at 50% duplication rate) and detecting on average 5,504 UMIs and 985 genes per cell (at 63% duplication rate), which is in line with the publicly available scRNA-seq datasets. To mitigate the potential impact of 3D genome's cell-cycle dynamics [16], we restricted our analysis to high-quality G0/G1 phase cells (837 cells).

Unsupervised clustering of GAGE-seq scRNA-seq data revealed six clusters (five clusters with continuous diffusion and one distinct cluster), each displaying unique gene signatures (Fig. 4.6a-b). Based on the gene expression signatures and known marker genes53, we annotated these clusters into known cell types: hematopoietic stem cell (HSC), multipotent progenitor (MPP), lymphoid-primed MPP (LMPP), multi-lymphoid progenitor (MLP), megakaryocyte-erythroid progenitor (MEP), and B lymphocyte natural killer cell progenitors (B-NK) (Fig. 4.6a-b). These clusters, representing all three major blood cell lineages, showed a lymphoid lineage preference. Our GAGE-seq scHi-C data also successfully resolved these six cell types (Fig. 4.6a-b), further demonstrating the ability of the 3D genome to encode cell type information.

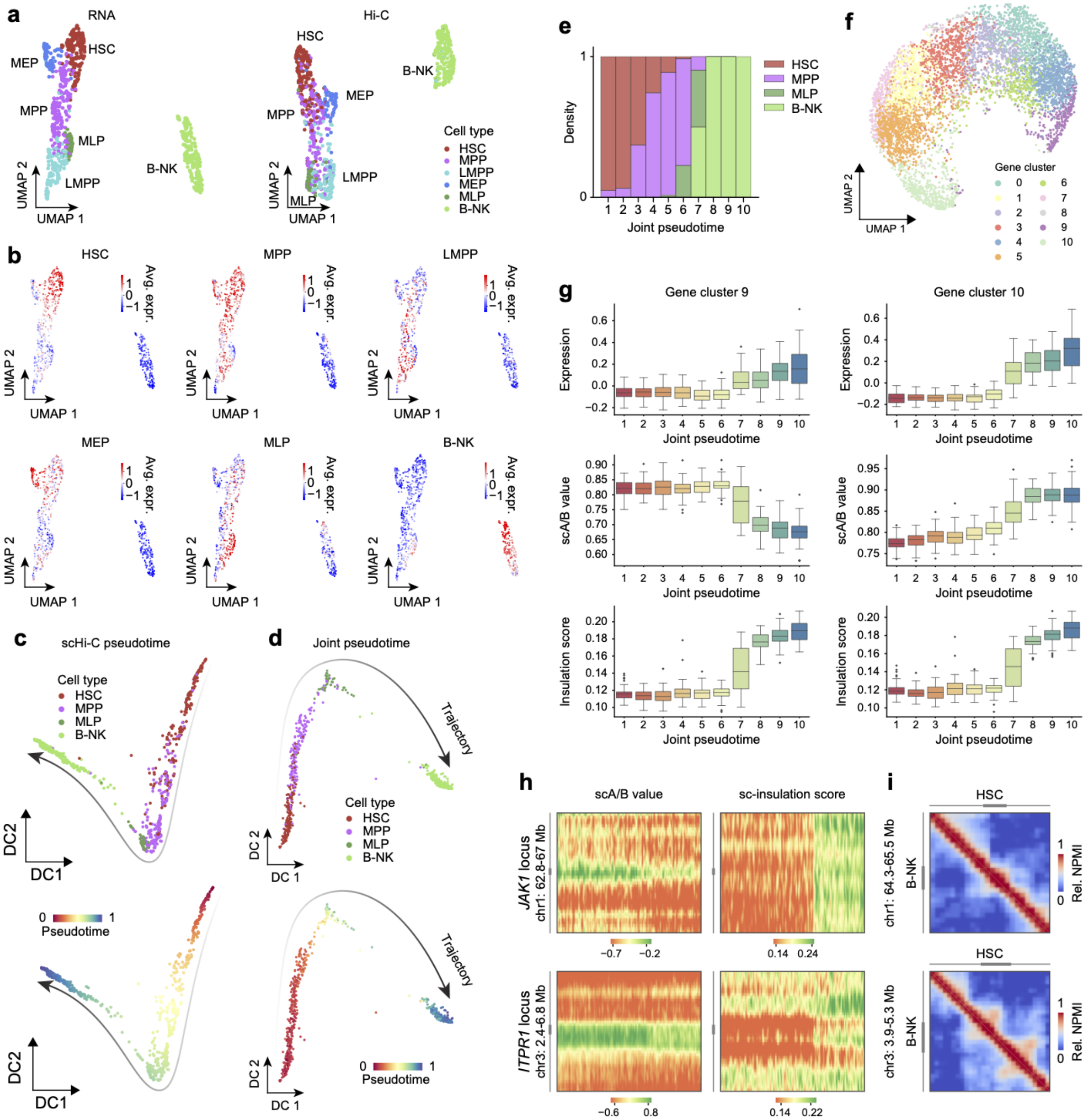

Figure 4.6: Interplay between 3D genome variation and gene expression changes in human bone marrow differentiation. **a.** UMAP visualization of GAGE-seq scRNA-seq (left) and scHi-C profiles (right) of human bone marrow CD34+ cells. **b.** Average expression of known marker genes on the UMAP plot. The 6 panels include n=124, 78, 24, 82, 126, and 356 genes for HSC, MPP, LMPP, MEP, MLP, and B-NK, respectively. **c-d.** Inferred B-NK lineage trajectory and pseudotime from scHi-C profiles (c) and jointly from scRNA-seq and scHi-C profiles (d), displayed by cell type (upper) and pseudotime (lower). **e.** Cell type compositions across 10 equally divided pseudotime bins. **f.** UMAP visualization of gene clusters determined by the temporal trend of expression and scA/B value. **g.** Temporal trends of gene expression (upper row), scA/B value (middle row), and single-cell insulation score (lower row) of gene clusters 9 (left column) and 10 (right column). **h.** scA/B (left) and single-cell insulation score (right) of the JAK1 (upper) and ITPR1 (lower) loci (at 100-Kb resolution). Each row represents a cell, ordered by the joint pseudotime from left to right. Heat maps were smoothed by a Gaussian kernel with a receptive field of 10 neighboring cells and 1 neighboring bin in each direction. **i.** Pseudo-bulk contact maps (at 50-Kb resolution) of HSC and B-NK at the JAK1 (upper) and ITPR1 (lower) loci.

Focusing on four of the six identified cell types (HSC, MPP, MLP and B-NK), which represent early B-NK lineage, we used GAGE-seq to reconstruct the developmental trajectory, demonstrating the dynamic interplay between genome structure and gene expression along this trajectory. Transcriptome and 3D genome-based pseudotime trajectories, inferred from GAGE-seq data, were highly congruent (Fig. 4.6c), suggesting that global 3D genome temporal variations overall mirror transcriptional changes and differentiation progression. Further, we created an integrated pseudotime trajectory (Fig. 4.6d, Methods), which was confirmed by the accurate alignment of the four cell types along the differentiation pseudotime and the observation that earlier-stage progenitors (e.g., HSCs) decrease while later-stage cells (e.g., B-NK) increase along the pseudotime (Fig. 4.6d-e).

4.3.8 Temporal interplays between 3D genome and gene expression

Comparisons between marker gene expression and 3D genome features in individual cell types during differentiation pseudotime suggest complex temporal interplay between both scA/B values and single-cell insulation scores with marker gene expressions.

We then performed an unsupervised clustering to further unravel relationships between gene expression and 3D genome features in the B-NK differentiation, based on all genes expressed in at least twenty single cells in the trajectory. We identified 11 distinct gene clusters (Fig. 4.6f). Notably, 5 of these 11 clusters showed a negative correlation between the changes in gene expression and scA/B value over pseudotime (Fig. 4.6g left panel). We closely examined gene cluster 9, where expression increases while scA/B value decreases. We selected two genes, JAK1 and ITPR1, which exhibit the highest similarity with the average temporal patterns of this gene cluster. Their scA/B value at the gene bodies indeed decreases over pseudotime without A/B compartment switches (Fig. 4.6h left panels). This analysis identified gene groups with varied temporal patterns, including discordant patterns in expression and scA/B value, as reported previously [19], during differentiation.

Regarding chromatin domains, a uniform temporal trend was observable in the aggregated single-cell insulation scores across all gene clusters, mirroring the pattern seen in the marker gene sets (Fig. 4.6g), indicating global 3D genome changes, manifested by widespread TAD-like domain re-organizations, in B-NK cells. For JAK1 and ITPR1, the single-cell insulation scores increased abruptly from MLP to B-NK, correlating with gene expression (Fig. 4.6h right panels), supported by aggregated contact maps (Fig. 4.6i). Additionally, we found that genes of different sizes appear to have distinct patterns with respect to single-cell insulation scores.

4.4 Discussion

Our high-throughput multiomic single-cell technology, GAGE-seq, delivers an integrative approach to co-assay 3D genome structure and gene expression in individual cells with high resolution. We show that GAGE-seq can reveal complex cell types from complex tissues not identified by other existing methods. Additionally, its data integration with spatial transcriptomic data points to great potential to reach a deeper understanding of 3D genome variation within complex tissues. Importantly, GAGE-seq also facilitates the reconstruction of differentiation trajectories based on 3D genome features, transcriptomes, or both. Our integration of GAGE-seq with single-cell chromatin accessibility data further highlights the advantage of GAGE-seq in linking CREs and their target genes. The high congruence between these modalities underscores the intimate connection between the temporal variations of the 3D genome and transcriptional rewiring during cell differentiation. Notably, GAGE-seq has revealed much more nuanced relationships between 3D genome features and gene expression during bone marrow B-NK lineage differentiation, creating a resource for future studies to disentangle causal gene regulatory changes in differentiation through the lens of 3D genome in single cells.

GAGE-seq is characterized by its efficiency, scalability, robustness, cost-effectiveness, and adaptability. We envision that GAGE-seq, along with our analytical tools, could significantly enhance the current toolkit for single-cell epigenomics. With wide-ranging applications, GAGE-seq can deepen our understanding of genome structure and function, providing insights into normal development and disease pathogenesis. Future refinements, such as enhancing barcoding strategy for higher throughput and improving detection of chromatin contacts, may allow GAGE-seq to construct high-resolution cell atlases and assess the role of pathogenic noncoding single-nucleotide variants on chromatin loops [150] in a massively parallel manner. Additionally, we anticipate a future application where GAGE-seq will be integrated with spatial labeling technologies, producing spatially-resolved scHi-C and scRNA-seq data. Such advancements will likely open up new avenues of investigation, such as exploring the role of the 3D genome in various tissue development and disease progression. Ultimately, GAGE-seq may offer the opportunity to integrate different molecular features in single cells, leading to a more comprehensive understanding of genome structure, cellular function, and their spatiotemporal variability.

Chapter 5

Hi-CFormer reveals the intricate interplay of DNA sequence, 3D genome structure, and transcriptome

5.1 Introduction

The advance of high-throughput whole-genome mapping methods for the three-dimensional (3D) genome organization such as Hi-C [5] has revealed multi-scale structures of chromatin folding within the cell nucleus, including A/B compartments [5], subcompartments [6, 7], topologically associating domains (TADs) [8, 9], and chromatin loops [6]. These structures play critical roles in gene regulation, cellular development, and disease progression [1, 11–13, 127–129]. However, the cell-to-cell variation of 3D genome structures and their functional significance remain poorly understood [1, 14]. Recent developments in single-cell Hi-C (scHi-C) technologies allow us to explore chromatin interactions with unparalleled detail, ranging from a few cells of specific types [15–18] to thousands of cells from complex tissues [19–21]. Emerging co-assayed technologies enabled the profiling and joint analysis of multiple modalities of complex tissues at the same time [151]. These emerging technologies and datasets hold the potential to reveal how genome structure relates to function in single cells across various biological settings, both in health and disease. These new technologies and datasets hold the potential to unveil the structure-function connections of the genome for a wide range of biological contexts in health and disease [14].

However, computational methods that can effectively utilize Hi-C data to reveal the roles of DNA sequence and 3D genome structure in transcriptional regulations are significantly lacking. Recently, predictive neural networks have been developed to understand the effect of coding and non-coding DNA sequences on transcriptomes, such as DeepSEA [78], Basenji2 [152], ExPecto [153], and Enformer [154]. Methodologically, these neural networks take a sequence as input and cannot directly incorporate a 2D Hi-C contact map as part of the input. Conceptually, the DNA sequence, which is the sole input to these algorithms, is shared for all cells from the same biological context, and these algorithms generate cell-specific predictions by including cell-specific model parameters. As a result, these algorithms' generalizability to unseen cells is far from ideal. Besides, the feature extraction modules (e.g., convolutional layers and transformer layers) are largely shared for all cells, obscuring the interpretation of the cell-to-cell variability in the model reasoning. Therefore, new algorithms are urgently needed to fill these important gaps.

Here, we present Hi-CFormer, a new computational method for understanding the intricate interplay of DNA sequence, 3D genome structure, and transcriptome using large language models. We formulate this goal as predicting mRNA signals from DNA sequence and 3D genome structure. On a mouse brain dataset, we show that the superior predictive performance of Hi-CFormer over sequence-only baselines. The interpretation of the trained Hi-CFormer model demonstrates its ability to capture cell-type-specific interaction among DNA sequence, 3D genome structure, and transcriptome. Hi-CFormer has the potential to shed new light on the functions of DNA sequence and 3D genome structure on transcriptional regulations.

5.2 Methods

5.2.1 Overview of Hi-CFormer

Hi-CFormer predicts mRNA signals from DNA sequence and 3D genome structure (Fig. 5.1). Hi-CFormer requires two types of input for each sample: a 409,600 bp-long DNA sequence and a Hi-C contact map for that genomic region at 1,024-bp resolution (Fig. 5.1a). Hi-

CFormer predicts the mRNA signals, i.e., the normalized number of transcripts for each 1,024-bp genomic locus. Although the DNA sequence is shared across all cells from the same biological context, Hi-CFormer learns the variability among cell types from the Hi-C information. As the entire Hi-CFormer model is shared across all cell types, i.e., there are no cell-type-specific model parameters, Hi-CFormer is able to generalize to unseen cell types.

As a proof of principle, we apply Hi-CFormer to pseudo-bulk data at the cell type level on a GAGE-seq dataset from mouse brains containing 28 cell types and 3740 highly variable genes [151]. We construct one sample centered at the transcription start site (TSS) of each highly variable for every cell type, resulting in 104,720 samples in total.

The Hi-CFormer architecture consists of four parts: (1) convolutional blocks with pooling, (2) Hi-C 1D information block, (3) 11 transformer blocks variants incorporating Hi-C 2D information (Fig. 5.1b,c), and (4) a cropping layer followed by final pointwise convolutions.

For the DNA sequence, we employ a one-hot-encoded format, specifically, A=[1, 0, 0, 0], C=[0, 1, 0, 0], G=[0, 0, 1, 0], T=[0, 0, 0, 1], N=[0, 0, 0, 0], with a length of 409,600 bp. Regarding the Hi-C contact map, we have applied feature processing partially based on prior knowledge, which results in two forms of input: Hi-C 1D signals and Hi-C 2D contact maps. The Hi-C 1D signals are derived from the Hi-C contact map at 1,024-bp resolution and include A/B compartment scores, insulation scores, and gene body scores. The Hi-C 2D contact map is directly taken by a Hi-C 2D contact map in 1024-bp resolution, so that the shape for the Hi-C 2D input is [400, 400]. For prediction, all the cell-type specific information in Hi-CFormer is from the cell-type specific Hi-C contact map, and the final predicted output is of length 240 corresponding to 245,760bp aggregated into 1024-bp bins. The convolutional blocks take only the DNA sequence as input and then reduce the genomic dimension from 409,600 bp to 400 so that each DNA position vector represents 1024 bp (although the convolutions do observe nucleotides in the adjacent pooled regions). Then the Hi-C 1D information block adds the mapped Hi-C 1D vector sequence to the

DNA position vector sequence and lets the added vector sequence contain Hi-C 1D information. The variants of transformer blocks then capture long-range interactions across the added vector sequence and also incorporating the Hi-C 2D information through the variants of the attention layer. The cropping layer trims 80 positions on each side following the setting in Enformer [154] to avoid computing the loss on the far ends because these regions are disadvantaged because they can observe regulatory elements only on one side (toward the sequence center) and not the other (the region beyond the sequence boundaries). Finally, the point-wise convolutions predict a single gene expression value for each 1024-bp genomic bin. The Hi-C-Former's architecture is similar to the state-of-the-art model Enformer [154]. However, our novel way of incorporating Hi-C information let our model can predict gene expression cell-wise or cell-type-wise, incorporate structure information for better prediction, and generalize to unseen cell/cell types.

5.2.2 Convolutional blocks with pooling

The first 7 convolutional blocks reduce the spatial dimension from 409,600 bp to 3200, and we use the same architecture with Enformer [154] about these 7 convolutional blocks and also use their pretrained parameter because their model has seen more DNA sequence structures and also for computational efficiency. Then we add one attention pooling layer(same attention pooling architecture in Enformer [154]) on our own to reduce the length from 3200 to 400. Finally, after the convolutional blocks, the shape of the input sample is [400, c]. c is the value of the hidden dimension. Enformer uses $c = 1536$ for each DNA position vector so that the pretrained convolutional blocks will transform the position vectors to 1536 hidden dimensions. But we add a transformation layer between the 7 pretrained convolutional blocks and the attention pooling in order to transform the hidden dimension to every other numbers.

5.2.3 Hi-C 1D information block

The input of Hi-C 1D data is [400, 5], we first use learnable linear transformation to map the Hi-C 1D data into c dimension then the mapped data shape is [400, c] and it is the same

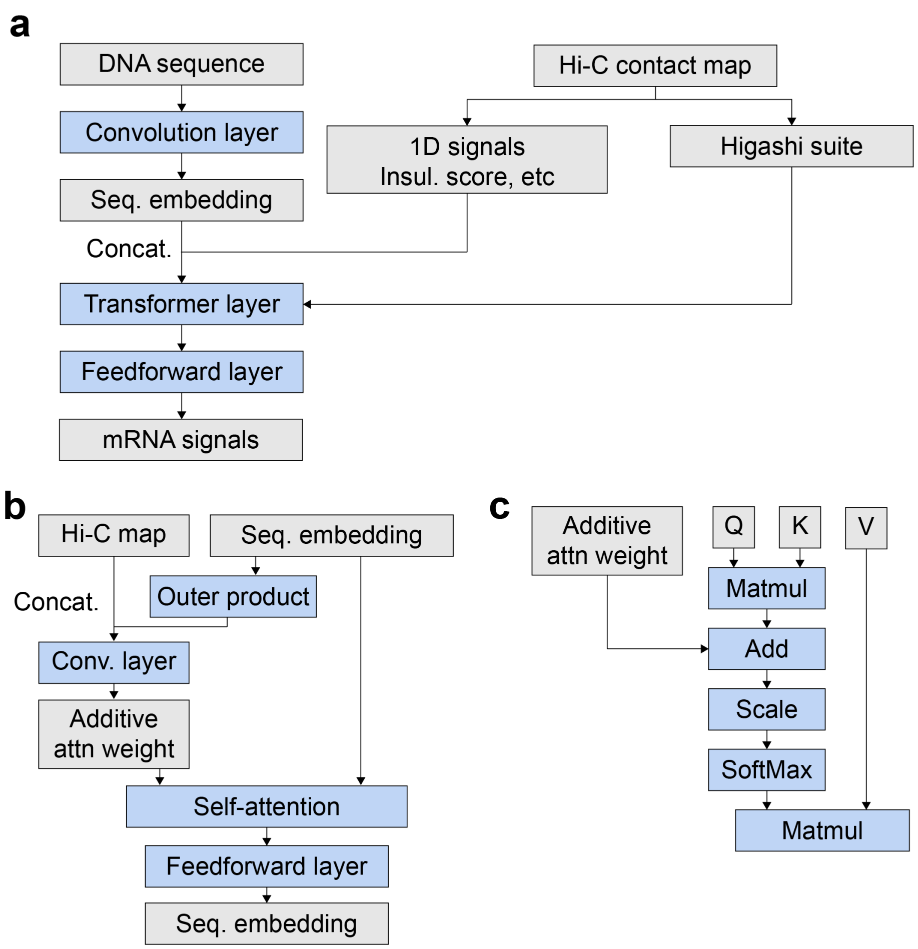

Figure 5.1: Overview of Hi-CFormer. **a.** The architecture of Hi-CFormer. **b.** The architecture of our customized variant of the transformer layer that takes an additional 2D Hi-C contact map as input. The 2D Hi-C contact map together with the outer product of sequence embedding will be fed into a convolutional neural network and be transformed into the Hi-C-based attention weight matrix. **c.** The architecture of our customized varient of the attention layer that takes an additional 2D matrix that will be added to the dot-product attention map.

with DNA position vector sequence. Then we can add the mapped Hi-C 1D data with DNA position vector sequence to get a mixed position vector sequence with the shape of [400, c], the learnable linear transformation can help adjust the space of Hi-C 1D data with the space of DNA position vectors.

5.2.4 Transformer Block Variant

The input of the Transformer Block is the vector sequence with the shape of [400, c], we use a similar transformer block architecture with Enformer, and we make some changes in order to incorporate Hi-C 2D information. The transformer block has two main ways, the first way is the same as other transformers, the input vector sequence will be fed into the multi-head attention layer. The second way is how we incorporate the Hi-C 2D information. We first calculate the dot-product of the input sequence vector in order to get some DNA sequence-wise information, then we append the dot-product matrix with the Hi-C 2D contact map and finally get a tensor with the shape of [400, 400, 2]. Then we apply three convolution layers to capture the local pattern on that matrix and finally map to a tensor of shape [400, 400, #number of attention heads], you can view this tensor as another ”attention score” calculated from Hi-C contact map and sequence vector’s dot-product for each attention head. Then we add this [400, 400, #number of attention heads] tensor to the attention score calculated by the original attention layer. After that, we can apply softmax to the new ”attention scores” to calculate the weight and all after this is the same with the traditional transformer block.

5.2.5 Positional encoding

We use the same positional encoding as Enformer [154] which is first formulated in the Transformer-XL paper.

5.2.6 Model training

The model uses the same Poisson negative log-likelihood loss function as Enformer [154] in the training stage. The training/validation/test sets were constructed as follows: the datasets we use have 3740 genes, each gene is of 409,600bp length, and we divide the total cell into 28 cell types, each gene in each cell type will be viewed as one sample, so the total sample number is 28*3740. We select 2992 genes in 22 cell types as the training set, the 748 unseen sequences in 22 seen cell types and the 2992 seen genes in 5 unseen cell types as the validation set, the 748 unseen genes in 5 unseen cell types as a test set. Therefore, our test results are showing that how we can generalize our model to unseen sequences and unseen cell types. We set the Learning rate range to be [1e-4, 5e-5, 1e-5] and the weight decay range to be [1e-4, 5e-5, 1e-5, 5e-6, 1e-6] and run all the hyperparameter combination then select the model with best validation set results as the model for analysis.

5.3 Results

5.3.1 Hi-CFormer accurately predicts expression by utilizing DNA sequence and 3D genome structure

We sought to demonstrate that Hi-CFormer can effectively utilize 3D genome structure to predict gene expression. We test our algorithm on a GAGE-seq dataset with co-assayed scHi-C and scRNA-seq from mouse brains. To train and evaluate Hi-CFormer and other models, we identified 3740 highly variable genes and calculated the pseudo-bulk mRNA signals and contact maps for each of the 28 cell types. We selected 2992 genes and 22 cell types as the training set, and we refer to them as seen genes and seen cell types, respectively. We refer to the rest of the genes and cell types as unseen genes and unseen cell types, respectively. We then evaluated the trained models on 3 datasets: 1) a dataset of 2992 seen genes and 6 unseen cell types, 2) a dataset of 748 unseen genes and 22 seen cell types, and 3) a dataset of 748 unseen genes and 6 unseen cell types. We used Pearson's correlation and the mean squared error as evaluation metrics.

Our evaluation shows the consistent and clear advantages of Hi-CFormer over baselines. The metric comparison on the training set and 3 validation sets illustrate Hi-CFormer's generalizability towards unseen cell types and unseen genes (Fig. 5.2a). Hi-CFormer also exhibits clear advantages over the sequence-only baseline on unseen cell types and unseen genes, demonstrating Hi-CFormer's effectiveness at utilizing 3D genome structure (Fig. 5.2a). We also conducted the ablation study assessing the necessity of representing 3D genome structure as both 1D scores (e.g., A/B values, insulation scores, and gene-body scores) and 2D contact maps. The Hi-CFormer models that only have access to 1D scores or 2D contact maps are denoted by Hi-CFormer (1D-only) and Hi-CFormer (2D-only), respectively. The ablation study shows that incorporating both 1D scores and 2D contact maps together yields optimal performance (Fig. 5.2a). The head-to-head comparison between Hi-CFormer and the sequence-only baseline visualizes the advantages of Hi-CFormer on individual genes (Fig. 5.2b). Specifically, Hi-CFormer outperforms the sequence-only baseline on a substantial number of genes (the light blue region in Fig. 5.2b).

A deeper dive into the cell-type-specific predictions shows Hi-CFormer's consistent advantages on 3 major cell types and all 28 finer cell types (Fig. 5.2 and Fig. [a supplementary figure]). Together, Hi-CFormer can effectively learn the intricate connection between 3D genome structure and gene expression.

The prediction performance of Hi-CFormer and the sequence-only baseline is related to gene functions. Hi-CFormer's advantages are higher in differentially expressed genes (DEGs) compared to other genes (Fig. 5.2d), implying the stronger connection between 3D genome structure and expression for DEGs. For the genes where Hi-CFormer and the sequence-only baseline have comparable performance, we found that genes with more TF binding sites in the promoter region are generally easier to predict (Fig. 5.2e), suggesting that the pre-trained CNN module is able to capture the effect of TF binding sites.

Together, evaluation on the GAGE-seq mouse brain dataset demonstrates the superior performance of Hi-CFormer, implying the important role of 3D genome structure in transcriptional regulation.

5.3.2 Hi-CFormer reveals the interplay between DNA sequence, 3D genome structure, and gene expression

We then sought to reveal the cell-type-specific interaction between DNA sequence, 3D genome structure, and transcriptome, by interpreting a trained Hi-CFormer model. Computationally, we quantify the effect of an input feature by the gradient of the predicted expression signal at the transcription start site (TSS) with respect to that input feature. For DNA sequence embedding, we take the L-2 norm of the gradient w.r.t. the embedding vector of each 1,024-bp genomic locus. The L-2 norm is a non-negative scalar and its magnitude represents the importance of that 1,024-bp genomic locus. For 3D genome structural features, including A/B values, insulation scores, and gene-body scores, the gradient can be positive and negative, with opposite implications. A positive (negative) gradient suggests that a larger value of that input feature likely leads to a higher (lower) predicted expression. A gradient close to zero implies that the input feature does not have a substantial effect on the predicted expression of that particular gene in a given cell type.

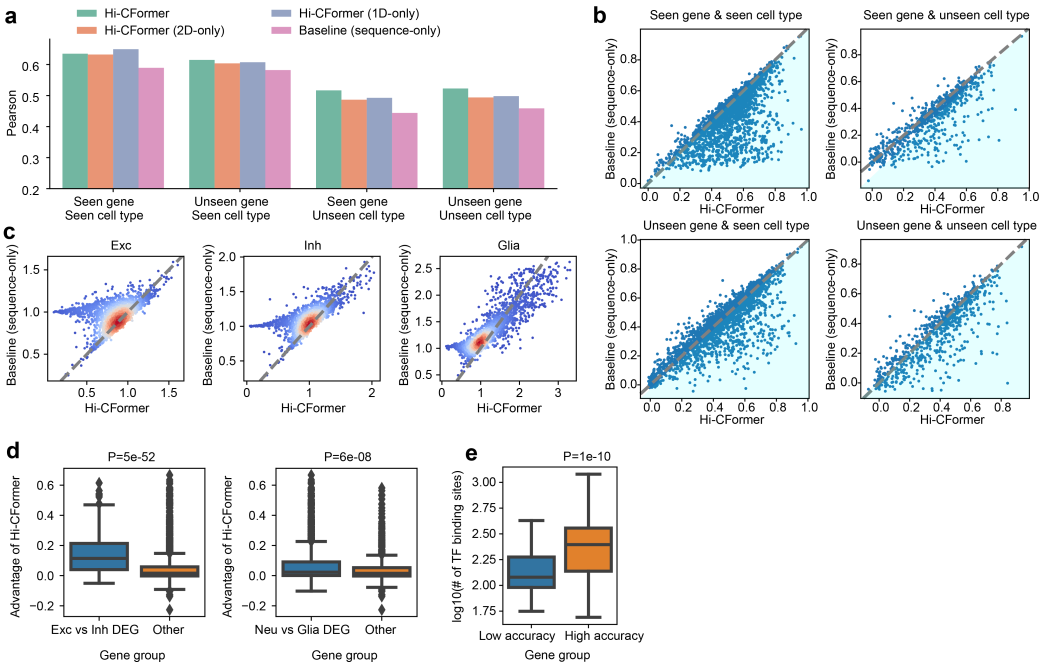

Figure 5.2: Hi-CFormer outperforms sequence-only baselines. **a.** Hi-CFormer outperforms the sequence-only baseline on unseen cell types and unseen genes, showing its superior generalizability. Pearson's correlation across genomic bins and cell types is used as the metric. **b.** Hi-CFormer shows clear and consistent advantages over the sequence-only baseline on individual genes. Pearson's correlation across genomic bins and cell types is used as the metric. **c.** Hi-CFormer shows clear advantages on major cell types. The mean squared error is used as the metric. **d.** The improvement in prediction performance of Hi-CFormer utilizing 3D genome structure information is higher in differentially expressed genes (DEGs). **e.** Genes with more transcription factor (TF) binding sites are easier to predict for both Hi-CFormer and the sequence-based baseline.

The effects of DNA sequences and 3D genome structures on predicted expression exhibit cell-type specificity. We aggregated the gradients over genomic loci and genes for each cell type. We found that DNA sequence has higher importance in excitatory neurons, whereas the gene-body score has higher importance in glial types (blue and red rectangles in Fig. 5.3a). The raw gene-body score has a higher variance in glial types than in neuronal types (Fig. 5.3b), which suggests its higher importance in glial types and supports our interpretation of Hi-CFormer. The high importance of DNA sequence in excitatory neurons is supported by two pieces of evidence. First, there are more enhancers in the upstream regions in excitatory neurons than in inhibitory neurons and glial cells (Fig. 5.3c). Second, the aforementioned enhancers contain more transcription factor (TF) bind sites in excitatory neurons than in inhibitory neurons and glial cells (Fig. 5.3d). Together, we presented cell-type-specific interplay between DNA sequence, 3D genome structure, and transcriptome.

The sequence and 3D genome structure of the gene body and flanking region have different impacts on expression. To quantify the overall impact of sequence and structural features, we aggregated the gradients over genes (Fig. 5.3e and [a supplementary figure]). The overall importance of sequence peaks at the transcription start site (TSS) and gradually decreases as the distance to TSS increases (Fig. 5.3e). In other words, the sequence of the gene body and the upstream promoter are relatively more important than those of other regions. The insulation score of the gene body is associated with negative gradients (Fig. 5.3e), consistent with its negative correlation with expression. Interestingly, the insulation score of the upstream promoter overall does not have a significant impact on predicted expression (Fig. 5.3e), suggesting the feature-to-feature variability in transcriptional regulations. A closer inspection of cell-type-specific enhancers shows that the structural features have a much higher impact on predicted expression, compared to background (Fig. 5.3f). Hence, Hi-CFormer has the potential to reveal critical transcriptional regulators.

In summary, we interpreted a trained Hi-CFormer model by taking the gradient of the predicted expression w.r.t. input features, revealing the complex and cell-type-specific interplay between DNA sequence, 3D genome structure, and transcriptome.

5.4 Discussion

In this work, we developed Hi-CFormer for analyzing the interplay between DNA sequence, 3D genome structure, and transcriptome. Our evaluation on a mouse brain dataset demonstrated the advantages of Hi-CFormer over sequence-only models for predicting cell-type-specific gene expressions. Additionally, we interpreted the reasoning of Hi-CFormer in two approaches, based on the gradients w.r.t. input features and the attention weight matrices, and revealed cell-type-specific transcriptional regulations.

The key conceptual and algorithmic innovation of Hi-CFormer is the incorporation of 3D genome structure into the input. First, the cell-type-specific 3D genome structure as part of the input enables us to train one Hi-CFormer model that is universal for all cell

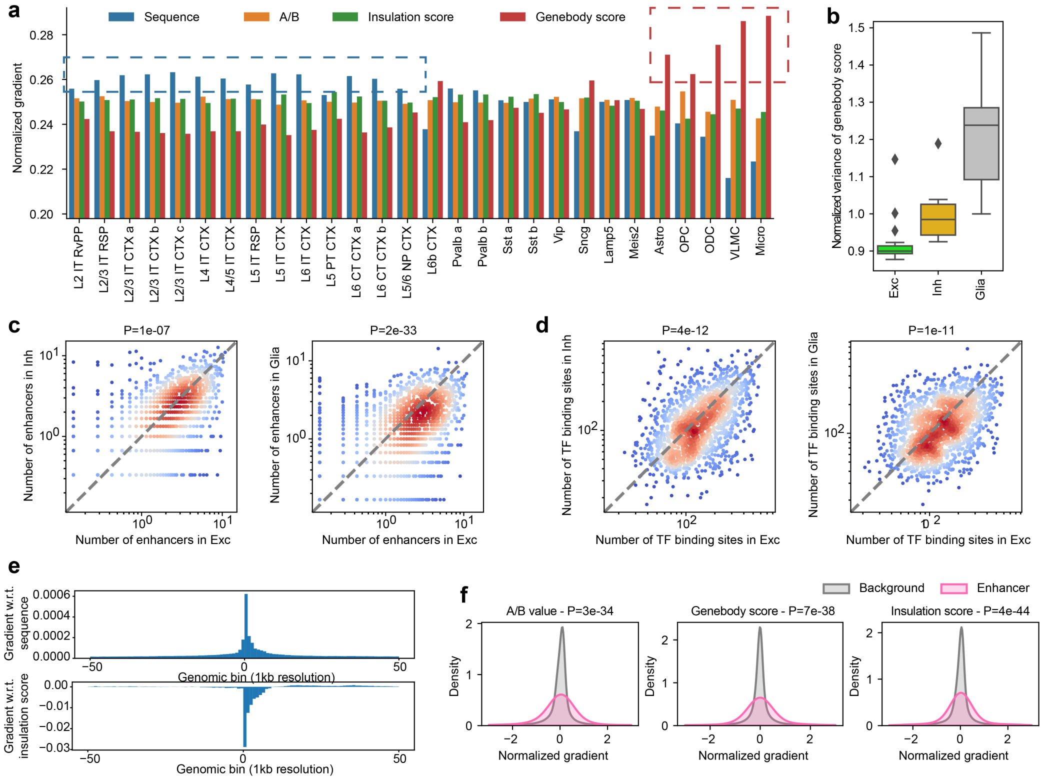

Figure 5.3: Hi-CFormer reveals cell-type-specific interplay between sequence, 3D genome structure, and transcriptome. **a.** The gradient of predicted expression with respect to input features in each cell type. The dashed blue rectangle highlights that sequence features have higher importance in excitatory neurons. The dashed red rectangle highlights that the gene-body score has higher importance in glial types. **b.** Gene-body score has a higher variance in glial types than neuronal types, supporting the importance of gene-body score in glial types. **c.** The upstream regions of genes contain more enhancers in excitatory neurons, compared to inhibitory neurons (left) and glial cells (right), supporting the importance of sequence features in excitatory neurons. **d.** The enhancers shown in panel (c) contain more transcription factor (TF) binding sites in excitatory neurons, compared to inhibitory neurons (left) and glial cells (right), supporting the importance of sequence features in excitatory neurons. **e.** The impact of sequence features (top) and multi-scale 3D genome structures (bottom) within and around the gene body on predicted expression. The importance is aggregated over cell types and genes. **f.** The structural features of cell-type-specific enhancers in the 100kb upstream regions have a higher impact on predicted expression.

types, including seen and unseen cell types. Second, the transform of Hi-C contact maps to attention weights is a novel and effective way to utilize 3D genome structure. Third, although we focused on Hi-C contact maps and the derived 1D scores, our model can include additional annotations from orthogonal datasets, such as chromatin loops and epigenetic signals.

Hi-CFormer can be further improved in several directions. First, as a data-driven method, Hi-CFormer can be improved by being trained on a larger dataset, potentially including multiple species, larger flanking regions, and single-nucleotide polymorphism. This may provide additional insights into transcriptional regulations, such as evolutionary conserveness, distal regulators, and the effect of mutations. Second, Hi-CFormer can be extended to predict other genomic tracks, such as DNA methylation and chromatin accessibility, providing a holistic picture of the intricate interplay between multiple omics.

Chapter 6

Conclusions

In this final chapter, we will first briefly summarize the methods developed in this thesis and their contributions to this field. We will then outline the future directions for studying the interplay between 3D genome structure, spatial context, chromosome accessibility, DNA sequence, and transcriptome, with methods described in this thesis as a foundation. These future directions include immediate integration of current algorithms, and future directions to move the field forward involving more advanced methodology development and broader applications.

6.1 Summary of the methods developed in the thesis

In Chapter 2, we introduce SPICEMIX, an interpretable method based on probabilistic, latent variable modeling for joint analysis of spatial information and gene expression from spatial transcriptome data. Both simulation and real data evaluations demonstrate that SPICEMIX markedly improves the inference of cell types and their spatial patterns compared with existing approaches. By applying to spatial transcriptome data of brain regions in humans and mice acquired by seqFISH+, STARmap, and Visium, we show that SPICEMIX can enhance the inference of complex cell identities, reveal interpretable spatial metagenes and uncover differentiation trajectories. SPICEMIX is a generalizable analysis framework for spatial transcriptome data to investigate the cell-type composition and spatial organization of cells in complex tissues.

In Chapter 3, we introduce Fast-Higashi, an ultrafast and interpretable method based on tensor decomposition and partial random walk with restart, enabling joint identification of cell identities and chromatin meta-interactions from sparse scHi-C data. Extensive evaluations demonstrate the advantage of Fast-Higashi over existing methods, leading to improved delineation of rare cell types and continuous developmental trajectories. Fast-Higashi can directly identify 3D genome features that define distinct cell types and help elucidate cell-type-specific connections between genome structure and function. Moreover, Fast-Higashi can generalize to incorporate other single-cell omics data. Fast-Higashi provides a highly efficient and interpretable scHi-C analysis solution that is applicable to a broad range of biological contexts.

In Chapter 4, we introduce genome architecture and gene expression by sequencing (GAGE-seq), a scalable, robust single-cell co-assay measuring 3D genome structure and transcriptome simultaneously within the same cell. Applied to mouse brain cortex and human bone marrow CD34+ cells, GAGE-seq characterized the intricate relationships between 3D genome and gene expression, showing that multiscale 3D genome features inform cell-type-specific gene expression and link regulatory elements to target genes. Integration with spatial transcriptomic data revealed in situ 3D genome variations in mouse cortex. Observations in human hematopoiesis unveiled discordant changes between 3D genome organization and gene expression, underscoring a complex, temporal interplay at the single-cell level. GAGE-seq provides a powerful, cost-effective approach for exploring genome structure and gene expression relationships at the single-cell level across diverse biological contexts.

In Chapter 5, we introduce Hi-CFormer, a transformer-based predictive model for predicting mRNA signals from DNA sequence and 3D genome structure, revealing the intricate interplay between DNA sequence, 3D genome structure, and transcriptome. Evaluation on a mouse brain dataset demonstrates the superior predictive performance of Hi-CFormer. Interpretation of the trained Hi-CFormer model reveals the cell-type-specific interplay between DNA sequence, 3D genome structure, and transcriptome.

6.2 Future Work

The integration of advanced genomic technologies such as SpiceMix, FastHigashi, GAGE-seq, and Hi-CFormer presents transformative opportunities for computational biology, particularly in drawing a holistic picture of multi-omic and multi-scale intracellular mechanisms and cell-to-cell communication. Each of these methodologies offers unique insights into the genomic and epigenomic landscapes of single cells from complex tissues, promising significant advancements in the study of various complex biological contexts. Experimentally, the multi-omic data can be efficiently obtained by GAGE-seq and other co-assayed technologies and further integrated by our developed algorithm. Computationally, the high dimensionality and high sparseness of those data can be effectively addressed by SpiceMix, FastHigashi, and Hi-CFormer, so that informative and interpretable representation can be learned and novel biological insights can be revealed. Below, we briefly discuss several directions.

6.2.1 Differential SpiceMix

The SpiceMix model introduced in Chapter 2 can be further enhanced for datasets collected from multiple biological contexts. Recently, advancements in spatial transcriptome technologies enabled the profiling of millions of cells from various conditions, such as development stages and diseases. One algorithmic improvement to SpiceMix for fully utilizing these emerging data is to incorporate differential model parameters, such as condition-specific metagenes and spatial affinities. The differential parameters will capture critical gene programs and spatial patterns related to development and the onset of disease.

6.2.2 Hierarchical SpiceMix for cellular organization

The SpiceMix model can also be adapted to understand subcellular organization and its relation to phenotypes. Nowadays, state-of-the-art spatial transcriptome technologies have reached subcellular resolutions. The SpiceMix model can be enhanced to jointly model the intracellular spatial distribution of mRNA molecules and the intercellular spatial contexts.

Specifically, we may include two layers of latent nodes: one layer of spots and the other layer of single cells. Each spot node is associated with a feature vector of observed gene expression, and we aim to learn latent representations for both spot nodes and cell nodes. Adjacent spot nodes within a cell are connected and so are adjacent cell nodes, just as in the SpiceMix model. We also learn the spatial affinities for spot nodes and cell nodes separately, as well as the connection between each cell node and the spot nodes belonging to that cell. As a result, the new probabilistic graphical model can reveal the connection between intracellular mRNA distribution, intercellular interaction, and cellular phenotype.

6.2.3 Multi-modal SpiceMix with heterogeneity nodes representing disease onsets

The SpiceMix model can be enhanced by incorporating disease-specific signals, such as the sites of protein misfolding in Alzheimer's disease. Specifically, we could incorporate a separate set of nodes that represent the sites of protein misfolding. We would also learn the spatial affinities between cell nodes and protein-misfolding nodes, aiming to reveal the cell-type-specific relation between protein-misfolding and cell identity. The enhanced model might reveal the cell-to-cell variability in the response to protein misfolding.

6.2.4 FastHigashi for co-assayed datasets

The FastHigashi algorithm introduced in Chapter 3 can be extended to co-assayed single-cell datasets, such as the GAGE-seq datasets and the sn-m3c-seq datasets. FastHigashi can effectively learn cell embeddings and meta-interactions from multiple chromosomes simultaneously. Importantly, the superior performance can be partially attributed to FastHigashi's mathematic constraints on the meta-interactions that embed our prior knowledge about chromosome structural properties. The additional co-assayed transcriptome could be incorporated as an additional matrix, which is connected to cell embeddings through a linear transformation. The additional co-assayed DNA methylation could be formulated as an additional column to the scHi-C contact maps. Together, the enhanced FastHigashi algorithm will be able to learn refined cell embeddings as well as the connection between multiple profiled modalities and cell phenotype.

6.2.5 Combination of SpiceMix and FastHigashi for spatial scHi-C data

The SpiceMix algorithm and the FastHigashi algorithm can be combined to study the spatial pattern of 3D genome structure. Specifically, we could 1) keep the formation of the spatial affinity between latent cell nodes and 2) replace the NMF connection between embedding and transcriptome in SpiceMix with the decomposition used in FastHigashi. Since high-quality spatial scHi-C datasets are not widely available yet, we could use the integrated datasets produced by our integration algorithm in Chapter 4. This combined algorithm is expected to retain the advantages of both SpiceMix and FastHigashi - to effectively learn informative cell embeddings and meta-interactions from high-dimensional and highly sparse scHi-C contact maps, while utilizing spatial information.

6.2.6 Integrative multi-omic analyses

Integrating multi-model data from technologies like GAGE-seq allows for comprehensive multi-omic analyses at the single-cell level. This approach could elucidate how changes at the genomic level (detected by GAGE-seq and other single-cell technologies) influence 3D genome structure (analyzed by FastHigashi), transcriptome (inferred by Hi-CFormer), and spatial patterns (learned by SpiceMix), and vice versa. For example, researchers could track the effects of DNA damage and repair mechanisms on gene expression and 3D genome structure in cancer cells, leading to insights into tumor evolution and metastasis. This holistic view could also help identify biomarkers for early disease detection or targets for precision therapies, enhancing personalized medicine strategies.

6.2.7 Enhanced diagnostic tools

The detailed molecular insights provided by these technologies can be leveraged to develop more precise diagnostic tools. For instance, patterns of 3D genome structure and DNA damage unique to specific diseases or stages of diseases, as identified through Hi-CFormer and GAGE-seq, could be integrated into diagnostic criteria. Additionally, changes in spatial gene expression patterns identified by SpiceMix could help in classifying subtypes of diseases more accurately. These advanced diagnostics could facilitate earlier detection and more tailored treatment plans, significantly improving patient prognosis.

Bibliography

[1] Tom Misteli. The self-organizing genome: principles of genome architecture and function. *Cell*, 183(1):28–45, 2020.

[2] Detlev Arendt, Jacob M Musser, Clare VH Baker, Aviv Bergman, Connie Cepko, Douglas H Erwin, Mihaela Pavlicev, Gerhard Schlosser, Stefanie Widder, Manfred D Laubichler, and Gunter D Wagner. The origin and evolution of cell types. *Nature Reviews Genetics*, 17(12):744–757, 2016.

[3] Xi Chen, Sarah A Teichmann, and Kerstin B Meyer. From tissues to cell types and back: Single-cell gene expression analysis of tissue architecture. *Annual Review of Biomedical Data Science*, 1:29–51, 2018.

[4] HuBMAP Consortium et al. The human body at cellular resolution: the NIH Human Biomolecular Atlas Program. *Nature*, 574(7777):187–192, 2019.

[5] Erez Lieberman-Aiden, Nynke L Van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R Lajoie, Peter J Sabo, Michael O Dorschner, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–293, 2009.

[6] Suhas SP Rao, Miriam H Huntley, Neva C Durand, Elena K Stamenova, Ivan D Bochkov, James T Robinson, Adrian L Sanborn, Ido Machol, Arina D Omer, Eric 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.

[7] Kyle Xiong and Jian Ma. Revealing hi-c subcompartments by imputing interchromosomal chromatin interactions. *Nature Communications*, 10, 2019.

[8] Jesse R Dixon, Siddarth Selvaraj, Feng Yue, Audrey Kim, Yan Li, Yin Shen, Ming

Hu, Jun S Liu, and Bing Ren. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376, 2012.

[9] Elphège P Nora, Bryan R Lajoie, Edda G Schulz, Luca Giorgetti, Ikuhiro Okamoto, Nicolas Servant, Tristan Piolot, Nynke L van Berkum, Johannes Meisig, John Sedat, et al. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*, 485(7398):381, 2012.

[10] Job Dekker, Andrew S Belmont, Mitchell Guttman, Victor O Leshyk, John T Lis, Stavros Lomvardas, Leonid A Mirny, Clodagh C O'shea, Peter J Park, Bing Ren, et al. The 4D nucleome project. *Nature*, 549(7671):219–226, 2017.

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

[12] Hui Zheng and Wei Xie. The role of 3Dgenome organization in development and cell differentiation. *Nature Reviews Molecular Cell Biology*, page 1, 2019.

[13] Rieke Kempfer and Ana Pombo. Methods for mapping 3D chromosome architecture. *Nature Reviews Genetics*, doi:10.1038/s41576-019-0195-2(4):207–226, 2019.

[14] Tianming Zhou, Ruochi Zhang, and Jian Ma. The 3D genome structure of single cells. *Annual Review of Biomedical Data Science*, 4, 2021.

[15] Vijay Ramani, Xinxian Deng, Ruolan Qiu, Kevin L Gunderson, Frank J Steemers, Christine M Disteche, William S Noble, Zhijun Duan, and Jay Shendure. Massively multiplex single-cell hi-c. *Nature Methods*, 14(3):263, 2017.

[16] Takashi Nagano, Yaniv Lubling, Csilla Várnai, Carmel Dudley, Wing Leung, Yael Baran, Netta Mendelson Cohen, Steven Wingett, Peter Fraser, and Amos Tanay. Cell-cycle dynamics of chromosomal organization at single-cell resolution. *Nature*, 547(7661):61, 2017.

[17] Hyeon-Jin Kim, Galip Gürkan Yardımcı, Giancarlo Bonora, Vijay Ramani, Jie Liu, Ruolan Qiu, Choli Lee, Jennifer Hesson, Carol B Ware, Jay 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.

[18] Guoqiang Li, Yaping Liu, Yanxiao Zhang, Naoki Kubo, Miao Yu, Rongxin Fang, Manolis Kellis, and Bing Ren. Joint profiling of DNA methylation and chromatin architecture in single cells. *Nature Methods*, 16(10):991–993, 2019.

[19] Longzhi Tan, Wenping Ma, Honggui Wu, Yinghui Zheng, Dong Xing, Ritchie Chen, Xiang Li, Nicholas Daley, Karl Deisseroth, and X Sunney Xie. Changes in genome architecture and transcriptional dynamics progress independently of sensory experience during post-natal brain development. *Cell*, 2021.

[20] Dong-Sung Lee, Chongyuan Luo, Jingtian Zhou, Sahaana Chandran, Angeline Rivkin, Anna Bartlett, Joseph R Nery, Conor Fitzpatrick, Carolyn O'Connor, Jesse R Dixon, et al. Simultaneous profiling of 3D genome structure and DNA methylation in single human cells. *Nature Methods*, pages 1–8, 2019.

[21] Hanqing Liu, Jingtian Zhou, Wei Tian, Chongyuan Luo, Anna Bartlett, Andrew Aldridge, Jacinta Lucero, Julia K Osteen, Joseph R Nery, Huaming Chen, et al. Dna methylation atlas of the mouse brain at single-cell resolution. *Nature*, 598(7879): 120–128, 2021.

[22] Je Hyuk Lee, Evan R Daugharthy, Jonathan Scheiman, Reza Kalhor, Joyce L Yang, Thomas C Ferrante, Richard Terry, Sauveur SF Jeanty, Chao Li, Ryoji Amamoto, Derek T Peters, Brian M Turczyk, Adam H Marblestone, Samuel A Inverso, Amy Bernard, Prashant Mali, Xavier Rios, John Aach, and George M Church. Highly multiplexed subcellular RNA sequencing in situ. *Science*, 343(6177):1360–1363, 2014.

[23] Kok Hao Chen, Alistair N Boettiger, Jeffrey R Moffitt, Siyuan Wang, and Xiaowei Zhuang. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science*, 348(6233):aaa6090, 2015.

[24] Sheel Shah, Eric Lubeck, Wen Zhou, and Long Cai. In situ transcription profiling of single cells reveals spatial organization of cells in the mouse hippocampus. *Neuron*, 92(2):342–357, 2016.

[25] Patrik L Ståhl, Fredrik Salmén, Sanja Vickovic, Anna Lundmark, José Fernández

Navarro, Jens Magnusson, Stefania Giacomello, Michaela Asp, Jakub O Westholm, Mikael Huss, Annelie Mollbrink, Sten Linnarsson, Simone Codeluppi, Åke Borg, Fredrik Pontén, Paul I Costea, Pelin Sahlén, Jan Mulder, Olaf Bergmann, Joakim Lundeberg, and Jonas Frisén. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353(6294):78–82, 2016.

[26] Jeffrey R Moffitt, Dhananjay Bambah-Mukku, Stephen W Eichhorn, Eric Vaughn, Karthik Shekhar, Julio D Perez, Nimrod D Rubinstein, Junjie Hao, Aviv Regev, Catherine Dulac, and Xiaowei Zhuang. Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. *Science*, 362(6416):eaau5324, 2018.

[27] Chee-Huat Linus Eng, Michael Lawson, Qian Zhu, Ruben Dries, Noushin Koulena, Yodai Takei, Jina Yun, Christopher Cronin, Christoph Karp, Guo-Cheng Yuan, and Long Cai. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. *Nature*, 568:235–239, 2019.

[28] Xiao Wang, William E Allen, Matthew A Wright, Emily L Sylwestrak, Nikolay Samusik, Sam Vesuna, Kathryn Evans, Cindy Liu, Charu Ramakrishnan, Jia Liu, Garry P Nolan, Felice-Alessio Bava, and Karl Deisseroth. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science*, 341(6400):eaat5691, 2018.

[29] Samuel G Rodriques, Robert R Stickels, Aleksandrina Goeva, Carly A Martin, Evan Murray, Charles R Vanderburg, Joshua Welch, Linlin M Chen, Fei Chen, and Evan Z Macosko. Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science*, 363(6434):1463–1467, 2019.

[30] Sanja Vickovic, Gökcen Eraslan, Fredrik Salmén, Johanna Klughammer, Linnea Stenbeck, Denis Schapiro, Tarmo Äijö, Richard Bonneau, Ludvig Bergenstråhle, José Fernandéz Navarro, Joshua Gould, Gabriel K Griffin, Åke Borg, Mostafa Ronaghi, Jonas Frisén, Joakim Lundeberg, Aviv Regev, and Patrik L Ståhl. High-definition spatial transcriptomics for in situ tissue profiling. *Nature Methods*, 16

(10):987–990, 2019.

[31] Xiaowei Zhuang. Spatially resolved single-cell genomics and transcriptomics by imaging. *Nature Methods*, 18(1):18–22, 2021.

[32] Ludvig Larsson, Jonas Frisén, and Joakim Lundeberg. Spatially resolved transcriptomics adds a new dimension to genomics. *Nature Methods*, 18(1):15–18, 2021.

[33] Robert R Stickels, Evan Murray, Pawan Kumar, Jilong Li, Jamie L Marshall, Daniela J Di Bella, Paola Arlotta, Evan Z Macosko, and Fei Chen. Highly sensitive spatial transcriptomics at near-cellular resolution with slide-seqv2. *Nature biotechnology*, 39(3):313–319, 2021.

[34] Andrew JC Russell, Jackson A Weir, Naeem M Nadaf, Matthew Shabet, Vipin Kumar, Sandeep Kambhampati, Ruth Raichur, Giovanni J Marrero, Sophia Liu, Karol S Balderrama, et al. Slide-tags enables single-nucleus barcoding for multimodal spatial genomics. *Nature*, 625(7993):101–109, 2024.

[35] Sophia Liu, J Bryan Iorgulescu, Shuqiang Li, Mehdi Borji, Irving A Barrera-Lopez, Vignesh Shanmugam, Haoxiang Lyu, Julia W Morriss, Zoe N Garcia, Evan Murray, et al. Spatial maps of t cell receptors and transcriptomes reveal distinct immune niches and interactions in the adaptive immune response. *Immunity*, 55(10):1940–1952, 2022.

[36] Wei-Ting Chen, Ashley Lu, Katleen Craessaerts, Benjamin Pavie, Carlo Sala Frigerio, Nikky Corthout, Xiaoyan Qian, Jana Laláková, Malte Kühnemund, Iryna Voytyuk, et al. Spatial transcriptomics and in situ sequencing to study alzheimer's disease. *Cell*, 182(4):976–991, 2020.

[37] Ed Lein, Lars E Borm, and Sten Linnarsson. The promise of spatial transcriptomics for neuroscience in the era of molecular cell typing. *Science*, 358(6359):64–69, 2017.

[38] Giovanni Palla, David S Fischer, Aviv Regev, and Fabian J Theis. Spatial components of molecular tissue biology. *Nature Biotechnology*, 40(3):308–318, 2022.

[39] Jie Liu, Dejun Lin, Galip Gürkan Yardımcı, and William Stafford Noble. Unsuper-

vised embedding of single-cell hi-c data. *Bioinformatics*, 34(13):i96–i104, 2018.

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

[41] Ye Zheng, Siqi Shen, and Sunduz Keles. Normalization and de-noising of single-cell Hi-C data with BandNorm and 3DVI. *bioRxiv*, 2021.

[42] Denis Schapiro, Hartland W Jackson, Swetha Raghuraman, Jana R Fischer, Vito RT Zanotelli, Daniel Schulz, Charlotte Giesen, Raúl Catena, Zsuzsanna Varga, and Bernd Bodenmiller. histoCAT: analysis of cell phenotypes and interactions in multiplex image cytometry data. *Nature Methods*, 14(9):873–876, 2017.

[43] Qian Zhu, Sheel Shah, Ruben Dries, Long Cai, and Guo-Cheng Yuan. Identification of spatially associated subpopulations by combining scRNAseq and sequential fluorescence in situ hybridization data. *Nature Biotechnology*, 36(12):1183–1190, 2018.

[44] Jian Hu, Xiangjie Li, Kyle Coleman, Amelia Schroeder, Nan Ma, David J Irwin, Edward B Lee, Russell T Shinohara, and Mingyao Li. SpaGCN: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nature Methods*, 18(11):1342–1351, 2021.

[45] Livnat Jerby-Arnon and Aviv Regev. Dialogue maps multicellular programs in tissue from single-cell or spatial transcriptomics data. *Nature Biotechnology*, pages 1–11, 2022.

[46] Edward Zhao, Matthew R Stone, Xing Ren, Jamie Guenthoer, Kimberly S Smythe, Thomas Pulliam, Stephen R Williams, Cedric R Uytingco, Sarah EB Taylor, Paul Nghiem, Jason H Bielas, and Raphael Gottardo. Spatial transcriptomics at subspot resolution with bayesspace. *Nature Biotechnology*, 39(11):1375–1384, 2021.

[47] Valentine Svensson, Sarah A Teichmann, and Oliver Stegle. SpatialDE: identifica-

tion of spatially variable genes. *Nature Methods*, 15(5):343–346, 2018.

[48] Damien Arnol, Denis Schapiro, Bernd Bodenmiller, Julio Saez-Rodriguez, and Oliver Stegle. Modeling cell-cell interactions from spatial molecular data with spatial variance component analysis. *Cell Reports*, 29(1):202–211, 2019.

[49] Mor Nitzan, Nikos Karaiskos, Nir Friedman, and Nikolaus Rajewsky. Gene expression cartography. *Nature*, 576(7785):132–137, 2019.

[50] Shiquan Sun, Jiaqiang Zhu, and Xiang Zhou. Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies. *Nature Methods*, 17(2): 193–200, 2020.

[51] Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M Mauck III, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive integration of single-cell data. *Cell*, 177(7):1888–1902, 2019.

[52] Joshua D Welch, Velina Kozareva, Ashley Ferreira, Charles Vanderburg, Carly Martin, and Evan Z Macosko. Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell*, 177(7):1873–1887, 2019.

[53] Marc Elosua-Bayes, Paula Nieto, Elisabetta Mereu, Ivo Gut, and Holger Heyn. SPOTlight: seeded NMF regression to deconvolute spatial transcriptomics spots with single-cell transcriptomes. *Nucleic Acids Research*, 49(9):e50, 2021.

[54] Tommaso Biancalani, Gabriele Scalia, Lorenzo Buffoni, Raghav Avasthi, Ziqing Lu, Aman Sanger, Neriman Tokcan, Charles R Vanderburg, Åsa Segerstolpe, Meng Zhang, Inbal Avraham-Davidi, Sanja Vickovic, Mor Nitzan, Sai Ma, Ayshwarya Subramanian, Michal Lipinski, Jason Buenrostro, Nik Bear Brown, Duccio Fanelli, Xiaowei Zhuang, Macosko Evan Z, and Aviv Regev. Deep learning and alignment of spatially resolved single-cell transcriptomes with Tangram. *Nature Methods*, 18 (11):1352–1362, 2021.

[55] Benjamin Chidester, Tianming Zhou, Shahul Alam, and Jian Ma. Spicemix enables integrative single-cell spatial modeling of cell identity. *Nature genetics*, 55(1):78–

88, 2023.

[56] Ruochi Zhang, Tianming Zhou, and Jian Ma. Ultrafast and interpretable single-cell 3d genome analysis with fast-higashi. In *International Conference on Research in Computational Molecular Biology*, pages 300–301. Springer, 2022.

[57] Tianming Zhou, Ruochi Zhang, Deyong Jia, Raymond T Doty, Adam D Munday, Daniel Gao, Li Xin, Janis L Abkowitz, Zhijun Duan, and Jian Ma. Concurrent profiling of multiscale 3d genome organization and gene expression in single mammalian cells. *Nature genetics*, 2024.

[58] Takashi Nagano, Yaniv Lubling, Tim J Stevens, Stefan Schoenfelder, Eitan Yaffe, Wendy Dean, Ernest D Laue, Amos Tanay, and Peter Fraser. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature*, 502(7469):59–64, 2013.

[59] Ilya M Flyamer, Johanna Gassler, Maxim Imakaev, Hugo B Brandão, Sergey V Ulianov, Nezar Abdennur, Sergey V Razin, Leonid A Mirny, and Kikuë Tachibana-Konwalski. Single-nucleus Hi-C reveals unique chromatin reorganization at oocyte-to-zygote transition. *Nature*, 544(7648):110–114, 2017.

[60] Tim J Stevens, David Lando, Srinjan Basu, Liam P Atkinson, Yang Cao, Steven F Lee, Martin Leeb, Kai J Wohlfahrt, Wayne Boucher, Aoife O'Shaughnessy-Kirwan, et al. 3D structures of individual mammalian genomes studied by single-cell Hi-C. *Nature*, 544(7648):59–64, 2017.

[61] Longzhi Tan, Dong Xing, Chi-Han Chang, Heng Li, and X Sunney Xie. Three-dimensional genome structures of single diploid human cells. *Science*, 361(6405): 924–928, 2018.

[62] Vijay Ramani, Xinxian Deng, Ruolan Qiu, Choli Lee, Christine M Disteche, William S Noble, Jay Shendure, and Zhijun Duan. Sci-hi-c: a single-cell hi-c method for mapping 3d genome organization in large number of single cells. *Methods*, 170: 61–68, 2020.

[63] Mary V Arrastia, Joanna W Jachowicz, Noah Ollikainen, Matthew S Curtis, Charlotte Lai, Sofia A Quinodoz, David A Selck, Mitchell Guttman, and Rustem F Ismag-

ilov. 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.

[64] Siyuan Wang, Jun-Han Su, Brian J Beliveau, Bogdan Bintu, Jeffrey R Moffitt, Chao-ting Wu, and Xiaowei Zhuang. Spatial organization of chromatin domains and compartments in single chromosomes. *Science*, 353(6299):598–602, 2016.

[65] Bogdan Bintu, Leslie J Mateo, Jun-Han Su, Nicholas A Sinnott-Armstrong, Mirae Parker, Seon Kinrot, Kei Yamaya, Alistair N Boettiger, and Xiaowei Zhuang. Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells. *Science*, 362(6413), 2018.

[66] Samuel Collombet, Noémie Ranisavljevic, Takashi Nagano, Csilla Varnai, Tarak Shisode, Wing Leung, Tristan Piolot, Rafael Galupa, Maud Borensztein, Nicolas Servant, et al. Parental-to-embryo switch of chromosome organization in early embryogenesis. *Nature*, 580(7801):142–146, 2020.

[67] Huy Q Nguyen, Shyamtanu Chattoraj, David Castillo, Son C Nguyen, Guy Nir, Antonios Lioutas, Elliot A Hershberg, Nuno MC Martins, Paul L Reginato, Mohammed Hannan, et al. 3D mapping and accelerated super-resolution imaging of the human genome using in situ sequencing. *Nature Methods*, 17(8):822–832, 2020.

[68] Jun-Han Su, Pu Zheng, Seon S Kinrot, Bogdan Bintu, and Xiaowei Zhuang. Genome-scale imaging of the 3D organization and transcriptional activity of chromatin. *Cell*, 2020.

[69] Yodai Takei, Jina Yun, Noah Ollikainen, Shiwei Zheng, Nico Pierson, Jonathan White, Sheel Shah, Julian Thomassie, Chee-Huat Linus Eng, Mitchell Guttman, et al. Global architecture of the nucleus in single cells by DNA seqFISH+ and multiplexed immunofluorescence. *bioRxiv*, 2020.

[70] Andrew C. Payne, Zachary D. Chiang, Paul L. Reginato, Sarah M. Mangiameli, Evan M. Murray, Chun-Chen Yao, Styliani Markoulaki, Andrew S. Earl, Ajay S. Labade, Rudolf Jaenisch, George M. Church, Edward S. Boyden, Jason D. Buenrostro, and Fei Chen. In situ genome sequencing resolves DNA sequence and struc-

ture in intact biological samples. *Science*, page eaay3446, 2020. doi: 10.1126/science.aay3446.

[71] Arjun Raj, Patrick Van Den Bogaard, Scott A Rifkin, Alexander Van Oudenaarden, and Sanjay Tyagi. Imaging individual mrna molecules using multiple singly labeled probes. *Nature methods*, 5(10):877–879, 2008.

[72] Meng Zhang, Stephen W Eichhorn, Brian Zingg, Zizhen Yao, Kaelan Cotter, Hongkui Zeng, Hongwei Dong, and Xiaowei Zhuang. Spatially resolved cell atlas of the mouse primary motor cortex by merfish. *Nature*, 598(7879):137–143, 2021.

[73] Kristen R. Maynard, Leonardo Collado-Torres, Lukas M. Weber, Cedric Uytingco, Brianna K. Barry, Stephen R. Williams, Joseph L. Catallini, Matthew N. Tran, Zachary Besich, Madhavi Tippani, Jennifer Chew, Yifeng Yin, Joel E Kleinman, Thomas M Hyde, Nikhil Rao, Stephanie C Hicks, Keri Martinowich, and Andrew E Jaffe. Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nature Neuroscience*, 24(3):425–436, 2021. ISSN 1097-6256, 1546-1726. doi: 10.1038/s41593-020-00787-0.

[74] David Arthur. K-means++: The advantages if careful seeding. In *Proc. Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, 2007*, pages 1027–1035, 2007.

[75] Joe H Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244, 1963.

[76] Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

[77] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

[78] Jian Zhou and Olga G Troyanskaya. Predicting effects of noncoding variants with deep learning–based sequence model. *Nature methods*, 12(10):931–934, 2015.

[79] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of

data with neural networks. *Science*, 313(5786):504–507, 2006.

[80] Tianwei Yue, Yuanxin Wang, Longxiang Zhang, Chunming Gu, Haoru Xue, Wenping Wang, Qi Lyu, and Yujie Dun. Deep learning for genomics: A concise overview. *arXiv preprint arXiv:1802.00810*, 2018.

[81] Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12): 1053–1058, 2018.

[82] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

[83] Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M Mauck, Shiwei Zheng, Andrew Butler, Maddie J Lee, Aaron J Wilk, Charlotte Darby, Michael Zager, et al. Integrated analysis of multimodal single-cell data. *Cell*, 184(13):3573–3587, 2021.

[84] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[85] Lawrence R Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[86] Rasmus Bro. Parafac. tutorial and applications. *Chemometrics and Intelligent Laboratory Systems*, 38(2):149–171, 1997.

[87] Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *nature*, 401(6755):788–791, 1999.

[88] Daniel D Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems*, pages 556–562, 2001.

[89] Jean-Philippe Brunet, Pablo Tamayo, Todd R Golub, and Jill P Mesirov. Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the National Academy of Sciences*, 101(12):4164–4169, 2004.

[90] Yongyue Zhang, Michael Brady, and Stephen Smith. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE Transactions on Medical Imaging*, 20(1):45–57, 2001.

[91] Kevin Murphy. *Machine learning: a probabilistic perspective*. MIT Press, 2012.

[92] Julian Besag. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society: Series B (Methodological)*, 48(3):259–279, 1986.

[93] LLC Gurobi Optimization. Gurobi optimizer reference manual, 2020. URL `http://www.gurobi.com`.

[94] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[95] Bosiljka Tasic, Vilas Menon, Thuc Nghi Nguyen, Tae Kyung Kim, Tim Jarsky, Zizhen Yao, Boaz Levi, Lucas T Gray, Staci A Sorensen, Tim Dolbeare, Darren Bertagnolli, Jeff Goldy, Nadiya Shapovalova, Sheana Parry, Changkyu Lee, Kimberly Smith, Amy Bernard, Linda Madisen, Susan M Sunkin, Michael Hawrylycz, Christof Koch, and Hongkui Zeng. Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nature Neuroscience*, 19(2):335–346, 2016.

[96] Ed S Lein, Michael J Hawrylycz, Nancy Ao, Mikael Ayres, Amy Bensinger, Amy Bernard, Andrew F Boe, Mark S Boguski, Kevin S Brockway, Emi J Byrnes, et al. Genome-wide atlas of gene expression in the adult mouse brain. *Nature*, 445(7124): 168–176, 2007.

[97] Tianyi Sun, Dongyuan Song, Wei Vivian Li, and Jingyi Jessica Li. scdesign2: a transparent simulator that generates high-fidelity single-cell gene expression count data with gene correlations captured. *Genome Biology*, 22(1):1–37, 2021.

[98] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[99] T Caliński and J Harabasz. A dendrite method for cluster analysis. *Commun. Stat. Simul. Comput.*, 3(1):1–27, January 1974.

[100] Rahul Satija, Jeffrey A Farrell, David Gennert, Alexander F Schier, and Aviv Regev. Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology*, 33 (5):495–502, 2015.

[101] Sueli Marques, Amit Zeisel, Simone Codeluppi, David van Bruggen, Ana Mendanha Falcão, Lin Xiao, Huiliang Li, Martin Häring, Hannah Hochgerner, Roman A Romanov, Daniel Gyllborg, Ana B Muñoz-Manchado, Gioele La Manno, Peter Lönnerberg, Elisa M Floriddia, Fatemah Rezayee, Patrik Ernfors, Ernest Arenas, Jens Hjerling-Leffler, Tibor Harkany, William D Richardson, Sten Linnarsson, and Gonçalo Castelo-Branco. Oligodendrocyte heterogeneity in the mouse juvenile and adult central nervous system. *Science*, 352(6291):1326–1329, 2016.

[102] Chuntao Zhao, Yaqi Deng, Lei Liu, Kun Yu, Liguo Zhang, Haibo Wang, Xuelian He, Jincheng Wang, Changqing Lu, Laiman N Wu, et al. Dual regulatory switch through interactions of Tcf7l2/Tcf4 with stage-specific partners propels oligodendroglial maturation. *Nature Communications*, 7(1):1–15, 2016.

[103] CHRISTOPHER Linington, MONIKA Bradl, HANS Lassmann, CHRISTOF Brunner, and KARL Vass. Augmentation of demyelination in rat acute allergic encephalomyelitis by circulating mouse monoclonal antibodies directed against a myelin/oligodendrocyte glycoprotein. *The American Journal of Pathology*, 130(3): 443–454, 1988.

[104] Bosiljka Tasic, Zizhen Yao, Lucas T Graybuck, Kimberly A Smith, Thuc Nghi Nguyen, Darren Bertagnolli, Jeff Goldy, Emma Garren, Michael N Economo, Sarada Viswanathan, et al. Shared and distinct transcriptomic cell types across neocortical areas. *Nature*, 563(7729):72–78, 2018.

[105] Amit Zeisel, Ana B Muñoz-Manchado, Simone Codeluppi, Peter Lönnerberg, Gioele La Manno, Anna Juréus, Sueli Marques, Hermany Munguba, Liqun He, Christer Betsholtz, et al. Cell types in the mouse cortex and hippocampus revealed

by single-cell RNA-seq. *Science*, 347(6226):1138–1142, 2015.

[106] Xiaojie Qiu, Qi Mao, Ying Tang, Li Wang, Raghav Chawla, Hannah A Pliner, and Cole Trapnell. Reversed graph embedding resolves complex single-cell trajectories. *Nature Methods*, 14(10):979–982, 2017.

[107] Sueli Marques, David van Bruggen, Darya Pavlovna Vanichkina, Elisa Mariagrazia Floriddia, Hermany Munguba, Leif Väremo, Stefania Giacomello, Ana Mendanha Falcão, Mandy Meijer, Åsa Kristina Björklund, et al. Transcriptional convergence of oligodendrocyte lineage progenitors during development. *Developmental Cell*, 46 (4):504–517, 2018.

[108] Rebecca M Beiter, Courtney Rivet-Noor, Andrea R Merchak, Robin Bai, David M Johanson, Erica Slogar, Katia Sol-Church, Christopher C Overall, and Alban Gaultier. Evidence for oligodendrocyte progenitor cell heterogeneity in the adult mouse brain. *Scientific Reports*, 12(1):1–15, 2022.

[109] Dataset: Allen institute for brain science (2021). allen cell types database – human multiple cortical areas [dataset]. available from: http://celltypes.brain-map.org/rnaseq, 2019.

[110] Seong-Seng Tan, Michael Kalloniatis, Hue-Trung Truong, Michele D Binder, Holly S Cate, Trevor J Kilpatrick, and Vicki E Hammond. Oligodendrocyte positioning in cerebral cortex is independent of projection neuron layering. *Glia*, 57 (9):1024–1030, 2009.

[111] Yang Liu, Mingyu Yang, Yanxiang Deng, Graham Su, Archibald Enninful, Cindy C Guo, Toma Tebaldi, Di Zhang, Dongjoo Kim, Zhiliang Bai, et al. High-spatial-resolution multi-omics sequencing via deterministic barcoding in tissue. *Cell*, 183 (6):1665–1681, 2020.

[112] Erick Armingol, Adam Officer, Olivier Harismendy, and Nathan E Lewis. Deciphering cell–cell interactions and communication from gene expression. *Nature Reviews Genetics*, 22(2):71–88, 2021.

[113] Ruochi Zhang, Tianming Zhou, and Jian Ma. Multiscale and integrative single-cell

Hi-C analysis with higashi. *Nature biotechnology*, 40(2):254–261, 2022.

[114] Ruochi Zhang, Yuesong Zou, and Jian Ma. Hyper-SAGNN: a self-attention based graph neural network for hypergraphs. In *International Conference on Learning Representations (ICLR)*, 2020.

[115] Mark H Van Benthem, Timothy J Keller, Gregory D Gillispie, and Stephanie A DeJong. Getting to the core of parafac2, a nonnegative approach. *Chemometrics and Intelligent Laboratory Systems*, 206:104127, 2020.

[116] Henk AL Kiers, Jos MF Ten Berge, and Rasmus Bro. Parafac2——part i. a direct fitting algorithm for the parafac2 model. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 13(3-4):275–294, 1999.

[117] Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M Mauck III, Shiwei Zheng, Andrew Butler, Maddie J Lee, Aaron J Wilk, Charlotte Darby, Michael Zager, et al. Integrated analysis of multimodal single-cell data. *Cell*, 184(13):3573–3587, 2021.

[118] Chongyuan Luo, Hanqing Liu, Fangming Xie, Ethan J Armand, Kimberly Siletti, Trygve E Bakken, Rongxin Fang, Wayne I Doyle, Rebecca D Hodge, Lijuan Hu, et al. Single nucleus multi-omics links human cortical cell regulatory genome diversity to disease risk variants. *bioRxiv*, 2019.

[119] Michael J Hawrylycz, Ed S Lein, Angela L Guillozet-Bongaarts, Elaine H Shen, Lydia Ng, Jeremy A Miller, Louie N Van De Lagemaat, Kimberly A Smith, Amanda Ebbert, Zackery L Riley, et al. An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature*, 489(7416):391–399, 2012.

[120] Rebecca D Hodge, Trygve E Bakken, Jeremy A Miller, Kimberly A Smith, Eliza R Barkan, Lucas T Graybuck, Jennie L Close, Brian Long, Nelson Johansen, Osnat Penn, et al. Conserved cell types with divergent features in human versus mouse cortex. *Nature*, 573(7772):61–68, 2019.

[121] Yodai Takei, Shiwei Zheng, Jina Yun, Sheel Shah, Nico Pierson, Jonathan White, Simone Schindler, Carsten H Tischbirek, Guo-Cheng Yuan, and Long Cai. Single-

cell nuclear architecture across cell types in the mouse brain. *Science*, 374(6567): 586–594, 2021.

[122] Thomas Cremer and Christoph Cremer. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nature Reviews Genetics*, 2(4):292–301, 2001.

[123] Jennifer E Phillips-Cremins, Michael EG Sauria, Amartya Sanyal, Tatiana I Gerasimova, Bryan R Lajoie, Joshua SK Bell, Chin-Tong Ong, Tracy A Hookway, Changying Guo, Yuhua Sun, et al. Architectural protein subclasses shape 3d organization of genomes during lineage commitment. *Cell*, 153(6):1281–1295, 2013.

[124] Jonathan A Beagan and Jennifer E Phillips-Cremins. On the existence and functionality of topologically associating domains. *Nature genetics*, 52(1):8–16, 2020.

[125] Tarik J Salameh, Xiaotao Wang, Fan Song, Bo Zhang, Sage M Wright, Chachrit Khunsriraksakul, and Feng Yue. A supervised learning framework for chromatin loop detection in genome-wide contact maps. *bioRxiv*, page 739698, 2019.

[126] Zhonghui Tang, Oscar Junhong Luo, Xingwang Li, Meizhen Zheng, Jacqueline Jufen Zhu, Przemyslaw Szalaj, Pawel Trzaskoma, Adriana Magalska, Jakub Wlodarczyk, Blazej Ruszczycki, et al. CTCF-mediated human 3D genome architecture reveals chromatin topology for transcription. *Cell*, 163(7):1611–1627, 2015.

[127] Jian Ma and Zhijun Duan. Replication timing becomes intertwined with 3D genome organization. *Cell*, 176(4):681–684, 2019.

[128] Malte Spielmann, Darío G Lupiáñez, and Stefan Mundlos. Structural variation in the 3d genome. *Nature Reviews Genetics*, 19(7):453–467, 2018.

[129] A Marieke Oudelaar and Douglas R Higgs. The relationship between genome structure and function. *Nature Reviews Genetics*, 22(3):154–168, 2021.

[130] Tim Stuart and Rahul Satija. Integrative single-cell analysis. *Nature Reviews Genetics*, 20(5):257–272, 2019.

[131] Junyue Cao, Diana R O'day, Hannah A Pliner, Paul D Kingsley, Mei Deng, Riza M Daza, Michael A Zager, Kimberly A Aldinger, Ronnie Blecher-Gonen, Fan Zhang,

et al. A human cell atlas of fetal gene expression. *Science*, 370(6518):eaba7721, 2020.

[132] Diego Calderon, Ronnie Blecher-Gonen, Xingfan Huang, Stefano Secchia, James Kentro, Riza M Daza, Beth Martin, Alessandro Dulja, Christoph Schaub, Cole Trapnell, et al. The continuum of drosophila embryonic development at single-cell resolution. *Science*, 377(6606):eabn5800, 2022.

[133] Chongyuan Luo, Hanqing Liu, Fangming Xie, Ethan J Armand, Kimberly Siletti, Trygve E Bakken, Rongxin Fang, Wayne I Doyle, Tim Stuart, Rebecca D Hodge, et al. Single nucleus multi-omics identifies human cortical cell regulatory genome diversity. *Cell genomics*, 2(3), 2022.

[134] Andrés M Cardozo Gizzi, Diego I Cattoni, Jean-Bernard Fiche, Sergio M Espinola, Julian Gurgo, Olivier Messina, Christophe Houbron, Yuki Ogiyama, Giorgio L Papadopoulos, Giacomo Cavalli, et al. Microscopy-based chromosome conformation capture enables simultaneous visualization of genome organization and transcription in intact organisms. *Molecular cell*, 74(1):212–222, 2019.

[135] Leslie J Mateo, Sedona E Murphy, Antonina Hafner, Isaac S Cinquini, Carly A Walker, and Alistair N Boettiger. Visualizing dna folding and rna in embryos at single-cell resolution. *Nature*, 568(7750):49–54, 2019.

[136] Yodai Takei, Jina Yun, Shiwei Zheng, Noah Ollikainen, Nico Pierson, Jonathan White, Sheel Shah, Julian Thomassie, Shengbao Suo, Chee-Huat Linus Eng, et al. Integrated spatial genomics reveals global architecture of single nuclei. *Nature*, 590 (7845):344–350, 2021.

[137] Zhiyuan Liu, Yujie Chen, Qimin Xia, Menghan Liu, Heming Xu, Yi Chi, Yujing Deng, and Dong Xing. Linking genome structures to functions by simultaneous single-cell hi-c and rna-seq. *Science*, 380(6649):1070–1076, 2023.

[138] Yawen Zhang, Xiaowei Xie, Yaojing Huang, Mengyao Liu, Qiaochuan Li, Jianming Luo, Yunyan He, Xiuxiu Yin, Shihui Ma, Wenbin Cao, et al. Temporal molecular program of human hematopoietic stem and progenitor cells after birth. *Developmen-*

tal Cell, 57(24):2745–2760, 2022.

[139] Giancarlo Bonora, Vijay Ramani, Ritambhara Singh, He Fang, Dana L Jackson, Sanjay Srivatsan, Ruolan Qiu, Choli Lee, Cole Trapnell, Jay Shendure, et al. Single-cell landscape of nuclear configuration and gene expression during stem cell differentiation and x inactivation. *Genome Biology*, 22:1–36, 2021.

[140] Jason D Buenrostro, Beijing Wu, Ulrike M Litzenburger, Dave Ruff, Michael L Gonzales, Michael P Snyder, Howard Y Chang, and William J Greenleaf. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, 523 (7561):486–490, 2015.

[141] Darren A Cusanovich, Riza Daza, Andrew Adey, Hannah A Pliner, Lena Christiansen, Kevin L Gunderson, Frank J Steemers, Cole Trapnell, and Jay Shendure. Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science*, 348(6237):910–914, 2015.

[142] Junyue Cao, Jonathan S Packer, Vijay Ramani, Darren A Cusanovich, Chau Huynh, Riza Daza, Xiaojie Qiu, Choli Lee, Scott N Furlan, Frank J Steemers, et al. Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science*, 357(6352):661–667, 2017.

[143] Alexander B Rosenberg, Charles M Roco, Richard A Muscat, Anna Kuchina, Paul Sample, Zizhen Yao, Lucas T Graybuck, David J Peeler, Sumit Mukherjee, Wei Chen, et al. Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science*, 360(6385):176–182, 2018.

[144] Sai Ma, Bing Zhang, Lindsay M LaFave, Andrew S Earl, Zachary Chiang, Yan Hu, Jiarui Ding, Alison Brack, Vinay K Kartha, Tristan Tay, et al. Chromatin potential identified by shared single-cell profiling of rna and chromatin. *Cell*, 183(4):1103–1116, 2020.

[145] Zizhen Yao, Cindy TJ van Velthoven, Thuc Nghi Nguyen, Jeff Goldy, Adriana E Sedeno-Cortes, Fahimeh Baftizadeh, Darren Bertagnolli, Tamara Casper, Megan Chiang, Kirsten Crichton, et al. A taxonomy of transcriptomic cell types across

the isocortex and hippocampal formation. *Cell*, 184(12):3222–3241, 2021.

[146] Warren Winick-Ng, Alexander Kukalev, Izabela Harabula, Luna Zea-Redondo, Dominik Szabó, Mandy Meijer, Leonid Serebreni, Yingnan Zhang, Simona Bianco, Andrea M Chiariello, et al. Cell-type specialization is encoded by specific chromatin topologies. *Nature*, 599(7886):684–691, 2021.

[147] Matthew G Heffel, Jingtian Zhou, Yi Zhang, Dong-Sung Lee, Kangcheng Hou, Oier Pastor Alonso, Kevin Abuhanna, Anthony D Schmitt, Terence Li, Maximilian Haeussler, et al. Epigenomic and chromosomal architectural reconfiguration in developing human frontal cortex and hippocampus. *bioRxiv*, pages 2022–10, 2022.

[148] Amanda J Law, Joel E Kleinman, Daniel R Weinberger, and Cynthia Shannon Weickert. Disease-associated intronic variants in the erbb4 gene are related to altered erbb4 splice-variant expression in the brain in schizophrenia. *Human molecular genetics*, 16(2):129–141, 2007.

[149] Chenxu Zhu, Yanxiao Zhang, Yang Eric Li, Jacinta Lucero, M Margarita Behrens, and Bing Ren. Joint profiling of histone modifications and transcriptome in single cells from mouse brain. *Nature methods*, 18(3):283–292, 2021.

[150] Joseph Nasser, Drew T Bergman, Charles P Fulco, Philine Guckelberger, Benjamin R Doughty, Tejal A Patwardhan, Thouis R Jones, Tung H Nguyen, Jacob C Ulirsch, Fritz Lekschas, et al. Genome-wide enhancer maps link risk variants to disease genes. *Nature*, 593(7858):238–243, 2021.

[151] Tianming Zhou, Ruochi Zhang, Deyong Jia, Raymond T Doty, Adam D Munday, Daniel Gao, Li Xin, Janis L Abkowitz, Zhijun Duan, and Jian Ma. Gage-seq concurrently profiles multiscale 3d genome organization and gene expression in single cells. *Nature Genetics*, pages 1–11, 2024.

[152] David R Kelley. Cross-species regulatory sequence activity prediction. *PLoS computational biology*, 16(7):e1008050, 2020.

[153] Jian Zhou, Chandra L Theesfeld, Kevin Yao, Kathleen M Chen, Aaron K Wong, and Olga G Troyanskaya. Deep learning sequence-based ab initio prediction of variant

effects on expression and disease risk. *Nature genetics*, 50(8):1171–1179, 2018.

[154] Žiga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R Ledsam, Agnieszka Grabska-Barwinska, Kyle R Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, and David R Kelley. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature methods*, 18(10):1196–1203, 2021.