Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/uasa20

Joint Tensor Modeling of Single Cell 3D Genome and Epigenetic Data with Muscle

Kwangmoon Park & Sündüz Keleş

To cite this article: Kwangmoon Park & Sündüz Keleş (2024) Joint Tensor Modeling of Single Cell 3D Genome and Epigenetic Data with Muscle, Journal of the American Statistical Association, 119:548, 2464-2477, DOI: 10.1080/01621459.2024.2358557

To link to this article: https://doi.org/10.1080/01621459.2024.2358557

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2024, VOL. 119, NO. 548, 2464–2477: Applications and Case Studies
https://doi.org/10.1080/01621459.2024.2358557

Joint Tensor Modeling of Single Cell 3D Genome and Epigenetic Data with Muscle

Kwangmoon Park[a] and Sündüz Keleş[a,b]

[a]Department of Statistics, University of Wisconsin, Madison, WI; [b]Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison, WI

ABSTRACT

Emerging single cell technologies that simultaneously capture long-range interactions of genomic loci together with their DNA methylation levels are advancing our understanding of three-dimensional genome structure and its interplay with the epigenome at the single cell level. While methods to analyze data from single cell high throughput chromatin conformation capture (scHi-C) experiments are maturing, methods that can jointly analyze multiple single cell modalities with scHi-C data are lacking. Here, we introduce Muscle, a semi-nonnegative joint decomposition of **Mu**ltiple **s**ingle **c**el**l** t**e**nsors, to jointly analyze 3D conformation and DNA methylation data at the single cell level. Muscle takes advantage of the inherent tensor structure of the scHi-C data, and integrates this modality with DNA methylation. We developed an alternating least squares algorithm for estimating Muscle parameters and established its optimality properties. Parameters estimated by Muscle directly align with the key components of the downstream analysis of scHi-C data in a cell type specific manner. Evaluations with data-driven experiments and simulations demonstrate the advantages of the joint modeling framework of Muscle over single modality modeling and a baseline multi modality modeling for cell type delineation and elucidating associations between modalities. Muscle is publicly available at *https://github.com/keleslab/muscle*. Supplementary materials for this article are available online, including a standardized description of the materials available for reproducing the work.

ARTICLE HISTORY
Received January 2023
Accepted May 2024

KEYWORDS
Block term tensor decomposition; Single cell 3D genome; Single cell DNA methylation; Tensor decomposition

1. Introduction

Interactions between distal genomic regions (i.e., loci) that become in close proximity of each other through chromatin loops and topologically associated domains (TADs) are key elements of gene regulatory mechanisms. High-throughput Chromatin Conformation Capture (Hi-C) sequencing technology (Lieberman-Aiden et al. 2009) captures snapshots of the long-range interactions of the genomic loci at the whole-genome level. Data from this technology consists of sequencing of millions of genomic locus-pairs that are in physical contact and is summarized by a symmetric Hi-C contact matrix, entries of which represent a measure of physical contact between the locus-pairs. Recent advancements in single cell sequencing technologies of Hi-C (scHi-C) enabled profiling interactions between distant genomic loci in individual cells (Stevens et al. 2017; Ramani et al. 2017; Tan et al. 2021; Ulianov et al. 2021) and even simultaneously with their DNA methylation status (sn-m3C-seq (Lee et al. 2019; Liu et al. 2021), scMethyl-HiC (Li et al. 2019)). These new approaches have the potential to elucidate the interplay between the epigenetic mechanisms and 3D genome structure in a wide variety of biological contexts. Computational approaches for specific scHi-C data inference tasks are appearing rapidly (e.g., scHiCluster (Zhou et al. 2019), scHiC Topics (Kim et al. 2020), Higashi (Zhang, Zhou, and Ma 2022a), BandNorm and scVI-3D (Zheng, Shen, and Keleş 2022), and Fast-Higashi (Zhang, Zhou, and Ma 2022b), scHiCTools (Li et al. 2021)). However, computational tools for integrating scHi-C with other data modalities such as transcriptomics, epigenomics, and epigenetics are lagging behind. Notably, the only method that can integrate scHi-C with scRNA-seq is scGAD (Shen, Zheng, and Keleş 2022). However, scGAD's common feature-based integration approach does not capitalize on the simultaneous profiling of 3D conformation and DNA methylation status of cells as enabled by sn-m3C-seq (Lee et al. 2019; Liu et al. 2021) and scMethyl-HiC (Li et al. 2019). In contrast, Higashi (Zhang, Zhou, and Ma 2022a) facilitates joint analysis of scHi-C and DNA methylation data; however, the inference implemented is limited to cell type clustering, and its practical utility is hindered by its computational requirements, which led to development of Fast-Higashi (Zhang, Zhou, and Ma 2022b). Fast-Higashi improved scalability of Higashi significantly; however, its current framework and implementation has not yet been leveraged to handle multiple modalities jointly.

In addition to the lack of integrative modeling approaches for scHi-C and DNA methylation, another key shortcoming of exiting scHi-C analysis methods, including scHiCluster, scHiC Topics, Higashi, scVI-3D, and Fast-Higashi, is a lack of alignment between the parameters estimated by these methods and the key parameters of interest in the scHi-C analysis. While these

CONTACT Sündüz Keleş *keles@stat.wisc.edu* Department of Statistics, University of Wisconsin, Madison, WI; Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison, WI.

Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JASA*.

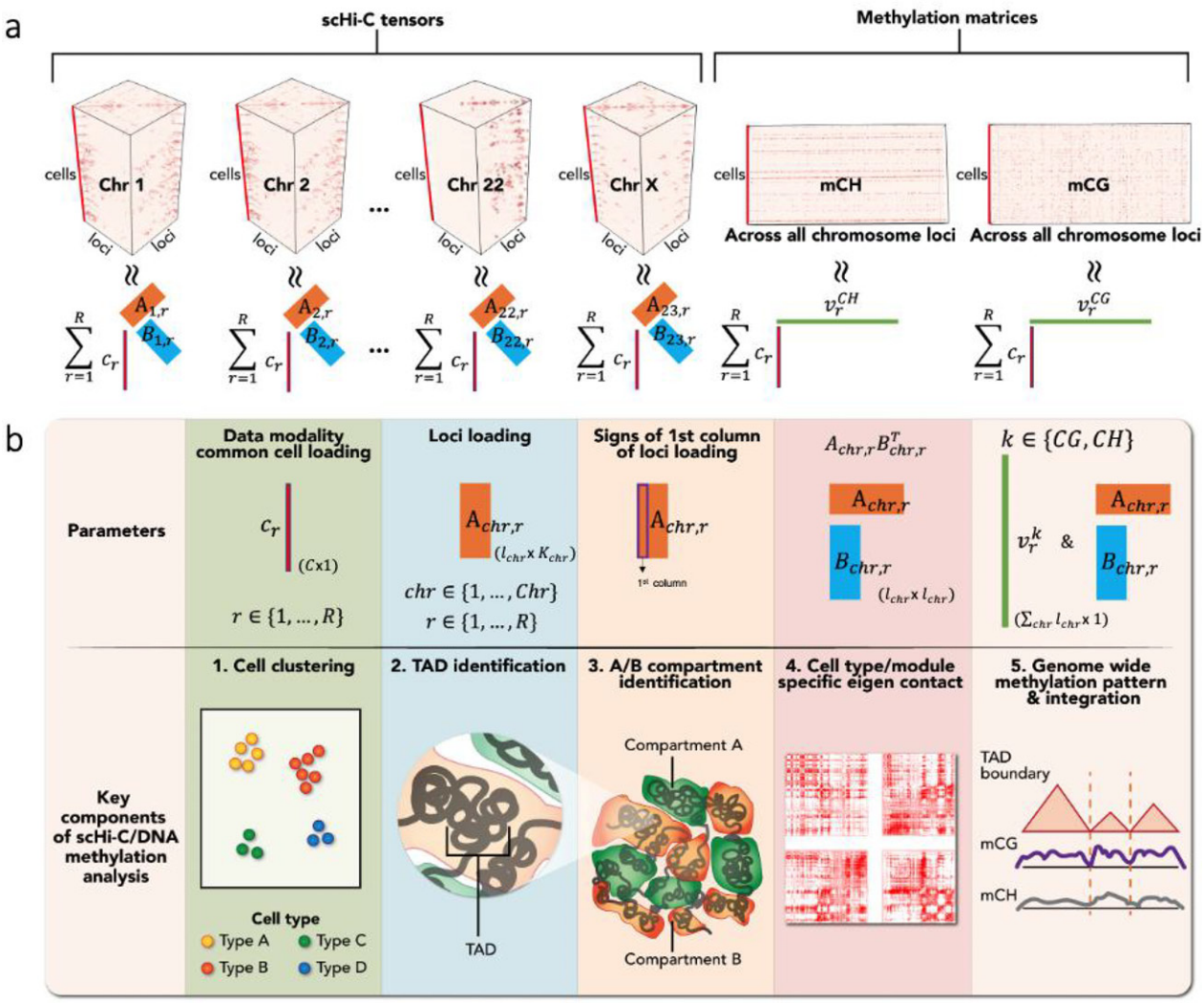

Figure 1. Overview of Muscle multiple single cell tensor model. (a) Each chromosome-specific scHi-C tensor with size $I_{chr} \times I_{chr} \times C$ is a summation of R "rank-1" modules $\left\{(\mathbf{A}_{chr,r}\mathbf{B}_{chr,r}^T) \circ \mathbf{c}_r \middle| r \in 1, \ldots, R\right\}$, where I_{chr} and C denote the # of loci for chromosome *chr* and the # of cells, respectively. Each module contains three factor loadings. The data modality common cell loading $\mathbf{c}_r \geq 0$ encodes which cell type the module corresponds to and provides a "label/name tag" for the module. Each of the chromosome-specific loci loadings $\mathbf{A}_{chr,r}$, $\mathbf{B}_{chr,r}$ encompass structural chromatin characteristics of a specific cell type, and the eigen contact $\mathbf{A}_{chr,r}\mathbf{B}_{chr,r}^T$ is the resulting interaction pattern (i.e., eigen contact matrix) of the cell type. Both mCG, mCH methylation matrices with size $\left(\sum_{chr} I_{chr}\right) \times C$ are also summation of "rank-1" modules $\left\{\mathbf{v}_r^k \circ \mathbf{c}_r \middle| r \in 1, \cdots, R\right\}$. $\mathbf{v}_r^k$, $k \in \{CG, CH\}$, encodes the methylation profile along the genome for the cell type inferred from the cell loading $\mathbf{c}_r$. The sizes of the tensors and the matrices are determined by I_{chr} and C, where the former varies by the bin size (resolution) of the analysis, and the latter depends on the size of the dataset. For the Kim et al. (2020) dataset, we have $C \approx 10,000$ and I_{chr} ranges between 49 and 250 at 1Mb resolution. (b) Muscle parameters align with the downstream analysis of interest. 1) The cell loading vectors $\{\mathbf{c}_r | r \in 1, \ldots, R\}$ enable cell clustering and identification of modules corresponding to each cell type. 2) Low dimensional projection of loci loading $\mathbf{A}_{chr,r}$ reveals loci clustering structure and TADs. 3) The first column vector of loci loading $\mathbf{A}_{chr,r}$ encodes A/B compartments which are large-scale genome territories. 4) Direct visualization of eigen contact matrix $\mathbf{A}_{chr,r}\mathbf{B}_{chr,r}^T$ reveals contact pattern of the cell type that the *r*th module corresponds to. 5) The methylation loci loading vector $\mathbf{v}_r^k$ aligns with the eigen contact matrix $\mathbf{A}_{chr,r}\mathbf{B}_{chr,r}^T$ or scHi-C loci loading $\mathbf{A}_{chr,r}$ to yield associations between DNA methylation and 3D genome structure of the cell type identified by cell loading vector $\mathbf{c}_r$. The equation within a parenthesis at the bottom corner of each object, for example, $(C \times 1)$ for c_r, denotes the dimensionality of it.

methods are able to learn latent representations of individual cells for downstream cell type clustering, inferring chromosome organization characteristics such as topologically associating domains (TADs) (Pombo and Dillon 2015), A/B compartments (Lieberman-Aiden et al. 2009) from their output requires, often complex, additional downstream processing of the estimated parameters or denoised data after aggregating contact matrices of the cells within each inferred cell type. From a strictly statistical perspective, the model parameters are estimated in isolation and without consideration of the intended post-processing procedures, which might lead to unreliable and sub-optimal inference.

Here, we develop a new statistical model named Muscle for **Mu**ltiple **s**ingle **c**el**l** **te**nsors to better align the estimated model parameters with the key inference parameters of scHi-C analysis and to enable integration of scHi-C with other data modalities. Muscle's multiple tensor framework encodes parameters such as cell-specific loadings shared by all data modalities, loci loadings specific to data modalities, scHi-C eigen contact matrices with one-to-one alignments to the cell types, A/B compartment structures, loci groupings or TADs, cell type specific methylation profiles, all of which are critical for 3D genome and methylation analysis (Figure 1). A key advantage of Muscle is that it can be deployed with only scHi-C data as well as with multiple

single cell data modalities. Application of Muscle to multiple scHi-C datasets with gold standard (Ramani et al. 2017; Lee et al. 2019; Li et al. 2019; Kim et al. 2020; Tan et al. 2021) demonstrates that Muscle performs as well or even better than existing methods for cell clustering and, more critically, can infer chromatin conformation structures in a cell type specific manner. Simulation studies comparing Muscle to a baseline method reveal consistently better performance by Muscle and supports the robustness of Muscle to a wide range of signal-to-noise levels. Muscle in the joint analysis for the sn-m3C-seq data (Lee et al. 2019; Liu et al. 2021) successfully identifies cell type specific associations between DNA methylation and 3D genome structure including TAD boundaries and compartment territories. Collectively, Muscle represents a significant modeling advancement in the joint analysis of scHi-C data with other data modalities.

2. Muscle Model

2.1. Muscle Model Representation

Muscle is a multiple tensor framework for single-cell multi-modal omics data. Here, we focus on scHi-C and DNA methylation and illustrate how Muscle parameters provide direct intuitive integrative inference of these single cell data modalities. Figure 1(a) provides an overview of Muscle, which starts out with a tensor view of multi-modal scHi-C data and single cell DNA methylation data (top row). For scHi-C data modality, each set of *cis*-interaction (i.e., only intra-chromosomal interactions) contact matrices of a single chromosome is viewed as an order three tensor, with dimensions # of loci on the chromosome (denoted as I_{chr}), # of loci on the chromosome, and # of cells (denoted as C). Here, a slice along the cell mode (or dimension) corresponds to a chromosome-specific scHi-C contact matrix for a single cell. For the human genome, this results in 23 tensors with the common cell mode but differing numbers of loci. For the single cell DNA methylation data, we form a mCG (mCH) matrix (i.e., order two tensor) with dimensions # of CpGs (non-CpGs) and # of cells, containing the CpG (non-CpG) site methylation level. The cell mode is shared between the scHi-C and methylation tensors because of the multi-modality (i.e., scHi-C and methylation read outs are taken simultaneously from a single cell) of the data.

After forming the entire order two and three tensors, Muscle parameterizes each of their mean tensors (bottom row of Figure 1(a)) following a semi-nonnegative Block Term Decomposition (BTD) form. Specifically, each of the scHi-C tensors is modeled as a summation of R "rank-1" modules, $\left\{(\mathbf{A}_{chr,r}\mathbf{B}_{chr,r}^T) \circ \mathbf{c}_r \middle| r \in 1,\ldots,R\right\}$. Note that we abuse the term "rank-1" to denote a rank-$(K_{chr}, K_{chr}, 1)$ tensor for simplicity in this article, where K_{chr} is defined as the block rank in BTD (De Lathauwer 2008). Each rank-1 module captures a latent contact pattern of the data. The two chromosome-specific loci loadings $\mathbf{A}_{chr,r}, \mathbf{B}_{chr,r} \in \mathbb{R}^{I_{chr} \times K_{chr}}$ harbor physical interaction information of the module and a nonnegative data modality common cell loading vector (i.e., loadings shared by all data modalities) $\mathbf{c}_r \in \mathbb{R}_+^C$ captures cell type information of the module. The methylation matrices are in the form of a semi-nonnegative matrix factorization, which is similarly a summation of R rank-1 modules $\left\{\mathbf{v}_r^k \circ \mathbf{c}_r \middle| r \in 1,\ldots,R\right\}$, where $k \in \{CG, CH\}$. Likewise, the data modality common cell loadings $\mathbf{c}_r$, shared with scHi-C, encodes modules specific to cell types, and the methylation loci loadings $\mathbf{v}_r^k$ identify cell type specific methylation patterns.

It is worthwhile to note that Muscle's Block term decomposition (BTD) is similar to but more flexible than widely used CANDECOMP/PARAFAC (CP) decomposition (De Silva and Lim 2008; Kolda and Bader 2009; Wang and Li 2020). This framework expresses a tensor $\mathcal{M}$ as a summation of outer products of single vectors, for example, $\mathcal{M} = \sum_{r=1}^{R} \lambda_r \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r$, where $\mathbf{a}_r, \mathbf{b}_r, \mathbf{c}_r$ are vectors with unit Euclidean norm, and λ_r is a real value that is analogous to an eigenvalue in matrix case. Our explorations of the scHi-C datasets relieved that the cell type specific average contact matrices tend to have higher rank than 1. Therefore, it is more appropriate to model the estimand as rank K $(\mathbf{A}_r\mathbf{B}_r^T)$ rather than a rank 1 matrix $(\mathbf{a}_r \circ \mathbf{b}_r)$, which underlines the better applicability of BTD on scHi-C data than CP decomposition. We also note that the BTD is a special case of the Tucker decomposition (Tucker 1966; De Lathauwer, De Moor, and Vandewalle 2000; Zhang and Xia 2018) framework, which extends the singular value matrix into a core tensor $\mathcal{C}$ and expresses $\mathcal{M}$ as a multiplication of the loading matrices $\mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3$ onto the core tensor $\mathcal{C}$. Specifically, BTD is a special case that has a core tensor with block diagonal structure (Rontogiannis, Kofidis, and Giampouras 2021).

Muscle formulation has two unique components that allow it to leverage multiple data modalities and enable direct inference for key parameters of interest. First, each cell loading vector $\mathbf{c}_r$ that is common to all data modalities learns the cell type information jointly across all chromosomes and data modalities, and, hence, is critical for the integrative analysis. Second, the non-negativity constraint on each cell loading vector $\mathbf{c}_r$ facilitates interpretation of each rank-1 module. For instance, if the cell loading vector of rth module has large values for a subset of the cells, that is, from the same cell type, the matrix $\mathbf{A}_{chr,r}\mathbf{B}_{chr,r}^T$ encodes the contact pattern for these groups of cells. In addition, the module's loci loadings $\mathbf{A}_{chr,r}$, $\mathbf{v}_r^k$ convey the cell type specific characteristics of genomic loci including A/B compartment structure and methylation profiles. We remark that in a PARAFAC2 (Kiers, Ten Berge, and Bro 1999) decomposition-based model, which was used by Fast-Higashi (Zhang, Zhou, and Ma 2022b), similar interpretation is hindered by the sign indeterminacy issues of singular vectors, which are the "cell embeddings" of the Fast-Higashi. Specifically, aligning of modules with cell types can not be achieved if a cell embedding vector of the module has both large negative and positive values for different cell types. We discuss practical implications of this limitation in more detail in Section S4. In contrast, Muscle's formulation achieves unification between model parameters and the key parameters needed for downstream inference. In sum, Muscle enables intuitive and direct interpretation of the model results as depicted in Figure 1(b). Each of Muscle's model parameters or a combination thereof aligns with key inference parameters of 3D chromatin organization along with the DNA methylation profile.

2.2. Statistical Framework of Muscle

In this section, we introduce the statistical framework of Muscle and a brief overview of parameter estimation of Muscle in the next section. This exposition uses the following key definitions and notations. A set of sequential numbers, for example, $\{1, \ldots, K\}$, are denoted as $[K]$. For a third order tensor $\mathcal{Y} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$, $\|\mathcal{Y}\|_F$ denotes the Frobenius norm, while for a vector $\mathbf{v}$, $\|v\|$ refers to the Euclidean norm of the vector. Finally, $\circ$ denotes outer product.

For a single cell $c \in [C]$, we have Chr number of chromosomes and a symmetric Hi-C contact matrix of size $l_{chr} \times l_{chr}$ for each $chr \in [Chr]$, where each (i, j)th entry of a contact matrix quantifies the observed physical interaction (e.g., contact) level between genomic loci i and j. For chromosome chr, the contact matrices stacked along cells have the same size $l_{chr} \times l_{chr}$. Hence, the data can be viewed as a (l_{chr}, l_{chr}, C)-dimensional tensor for each chromosome. We denote each pre-processed (e.g., log transformed) scHi-C tensor as $\mathcal{Y}_{chr} \in \mathbb{R}^{l_{chr} \times l_{chr} \times C}$, $chr \in [Chr]$. Details about data pre-processing are in Section S3.

The scHi-C tensors $\{\mathcal{Y}_{chr} \in \mathbb{R}^{l_{chr} \times l_{chr} \times C} | chr \in [Chr]\}$ and methylation matrices $\mathbf{Y}^{CG}, \mathbf{Y}^{CH} \in \mathbb{R}^{\sum_{chr} l_{chr} \times C}$, binned at the desired resolution (e.g., 500 Kb or 1 Mb), are modeled as

$$\begin{aligned}
\mathcal{Y}_{chr} &= \mathcal{M}_{chr} + \mathcal{E}_{chr}, \quad \epsilon_{i,j,c,chr} \overset{\text{iid}}{\sim} N(0, \sigma_1^2), \\
&\quad \forall chr \in [Chr], \qquad (1) \\
\mathbf{Y}^k &= \mathbf{M}^k + \mathbf{E}^k, \quad \epsilon^k_{l,c} \overset{\text{iid}}{\sim} N(0, \sigma_2^2), \quad \text{for} \\
&\quad k \in \{CG, CH\}, \\
\text{s.t. } \mathcal{M}_{chr} &= \sum_{r=1}^{R} (\mathbf{A}_{chr,r} \mathbf{B}^T_{chr,r}) \circ \mathbf{c}_r, \\
\mathbf{M}^k &= \sum_{r=1}^{R} \mathbf{v}^k_r \circ \mathbf{c}_r, \quad \text{for } k \in \{CG, CH\}, \qquad (2) \\
\mathbf{c}_r &\geq 0, \quad \|\mathbf{c}_r\| = 1, \quad \mathbf{B}^T_{chr,r} \mathbf{B}_{chr,r} = \mathbf{I}, \\
\mathbf{A}_{chr,r} &= \mathbf{B}_{chr,r} \mathbf{D}_{chr,r}, \qquad (3) \\
\text{and } \frac{\sigma_1^2}{\sigma_2^2} &= \frac{N_h}{N_m}, \quad \forall r \in [R], \quad \forall chr \in [Chr], \qquad (4)
\end{aligned}$$

where $\mathbf{c_r} \geq 0$ indicates that all the entries are nonnegative, and all the chromosome specific signal and noise tensors are with size $\mathcal{M}_{chr}, \mathcal{E}_{chr} \in \mathbb{R}^{l_{chr} \times l_{chr} \times C}$, and mCG, mCH methylation signal and error matrices $\mathbf{M}^{CG}, \mathbf{M}^{CH}, \mathbf{E}^{CG}, \mathbf{E}^{CH} \in \mathbb{R}^{\sum_{chr} l_{chr} \times C}$. Furthermore, all the error terms $\mathcal{E}_{chr}, \mathbf{E}^k$ are entry-wise independent of each other. Also note that $\mathbf{D}_{chr,r} = \text{diag}(\lambda_{chr,r,k})_{k=1}^{K_{chr}} \in \mathbb{R}^{K_{chr} \times K_{chr}}$ and $\lambda_{chr,r,k} > 0$ so that $\mathbf{A}_{chr,r} \in \mathbb{R}^{l_{chr} \times K_{chr}}$ and $\mathbf{B}_{chr,r} \in \mathbb{R}^{l_{chr} \times K_{chr}}$ are equivalent up to multiplication of diagonal matrix absorbing the magnitude of the module. Here, the total size of scHi-C tensors is defined as $N_h = C \times \sum_{chr} l^2_{chr}$, and, similarly, the size of a methylation matrix is defined as $N_m = C \times \sum_{chr} l_{chr}$. These size terms are leveraged to model the proportion of the variances of the two sources of data (4). The signal tensor $\mathcal{M}_{chr}$ is in the form of block term decomposition (De Lathauwer 2008) and, the mean methylation matrices $\mathbf{M}^{CG}, \mathbf{M}^{CH}$ have the form of a semi-nonnegative matrix factorization. A key component of this integrative framework is that the nonnegative cell loading vectors $\mathbf{c}_r \in \mathbb{R}^C_+$, $r \in [R]$, are shared between the models (2), enabling the cell loadings to be learned by leveraging both data modalities.

2.3. Muscle Model Estimation

We introduce the estimation problem and algorithm overview of Muscle. Given the scHi-C tensors $\{\mathcal{Y}_{chr} \in \mathbb{R}^{l_{chr} \times l_{chr} \times C} | chr \in [Chr]\}$ and methylation matrices $\mathbf{Y}^{CG}, \mathbf{Y}^{CH} \in \mathbb{R}^{\sum_{chr} l_{chr} \times C}$, Muscle solves the Maximum Likelihood Estimation equivalent problem:

$$\min_{\substack{\mathbf{A}_{chr,r}, \mathbf{B}_{chr,r}, \mathbf{c}_r \\ \mathbf{v}^{CG}_r, \mathbf{v}^{CH}_r}} \left\{ \frac{1}{N_h} \sum_{chr=1}^{Chr} \left\| \mathcal{Y}_{chr} - \sum_{r=1}^{R} (\mathbf{A}_{chr,r} \mathbf{B}^T_{chr,r}) \circ \mathbf{c}_r \right\|_F^2 + \frac{1}{N_m} \sum_{k \in \{CG,CH\}} \left\| \mathbf{Y}^k - \sum_{r=1}^{R} \mathbf{v}^k_r \circ \mathbf{c}_r \right\|_F^2 \right\}. \qquad (5)$$

To solve this non-convex problem, we derive an Alternating Least Squares (ALS) algorithm (Algorithm 1 of Section S1). The ALS algoritm iteratively obtains loci loadings $\mathbf{A}_{chr,r}$, $\mathbf{B}_{chr,r}$, and $\mathbf{v}^k_r$ given the cell loadings $\mathbf{c}_r$, shared by both data modalities across all the chromosomes, and updates $\mathbf{c}_r$ by pooling all the loci loading information across the modalities and chromosomes. We derive the optimality properties of the ALS in Section S2.

3. Benchmarking with Datasets with Gold Standard

3.1. Muscle is Widely Applicable for Cell Type Identification with Even Single Modality scHi-C Data

We start out by exploring Muscle's applicability with single modality scHi-C data by evaluating its cell clustering performance with multiple 3D genome datasets (Ramani et al. 2017; Lee et al. 2019; Li et al. 2019; Kim et al. 2020; Tan et al. 2021). Details on parameter settings are provided in the Section S3. Muscle enables cell clustering through the estimated cell loadings $\{\mathbf{c}_r \in \mathbb{R}^C_+ | r \in [R]\}$. Figure 2(a), (b) display the scatterplots of the first two UMAP coordinates of cell embeddings from Muscle and other scHi-C analysis methods (Zhou et al. 2019; Kim et al. 2020; Li et al. 2021; Zheng, Shen, and Keleş 2022; Zhang, Zhou, and Ma 2022a) for the Li et al. (2019) and Kim et al. (2020) datasets, which have the least and the most numbers of cells, respectively. Similar displays for rest of the datasets are available in Figure S2. Figure 2(a) highlights that Muscle and Higashi (Zhang, Zhou, and Ma 2022a) are the only models that can separate Serum 1 cells from the others. Similarly, in Figure 2(b), the separation of the four major cell types (GM12878, H1Esc, HFF, HAP1) is more evident for Muscle, scHiC Topics (Kim et al. 2020), and scVI-3D (Zheng, Shen, and Keleş 2022) compared to the others. Next, we systematically evaluated the "cell type identification by clustering" performances of the methods. We used both the learned embeddings of the methods (e.g., cell loadings $\mathbf{c}_r$, $r \in [R]$ for Muscle) and their low dimensional projections with UMAP and tSNE for cell clustering with k-means and employed the Adjusted Rand Index (ARI) and Average Silhouette Score metrics for evaluation based on gold standard cell labels (Figure 2(c)). Figure 2(d) summarizes the

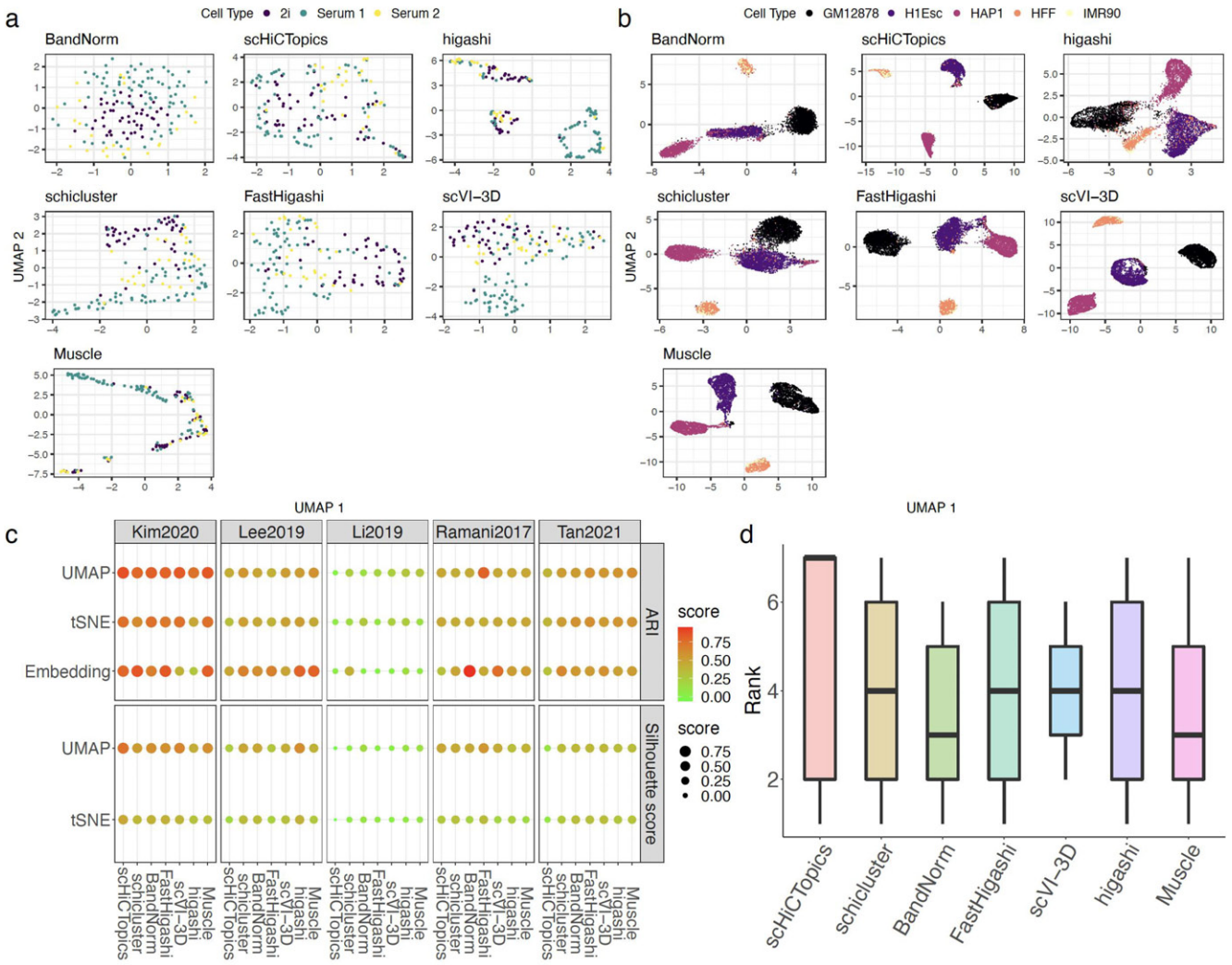

Figure 2. Computational evaluation and benchmarking of Muscle cell clustering with scHi-C data. (a) and (b) UMAP coordinates of the cells from Li et al. (2019) and Kim et al. (2020) scHi-C datasets. Muscle UMAP coordinates are obtained from estimated $\{\mathbf{c}_r \in \mathbb{R}_+^C | r \in [R]\}$. Cells are colored based on known cell type labels. (c) Evaluation of the cell clustering by different methods. Larger and redder circles correspond to larger scores. (d) Summary of the method rankings for each dataset and evaluation metric displayed in panel (c).

overall ranking of the clustering performances for each of the combinations in Figure 2(c), and yields that Muscle along with BandNorm shows the best overall ranking for cell clustering solely based on scHi-C data. This establishes Muscle's applicability with scHi-C data even in the single data modality setting. In addition to the large-scale benchmarking experiments, we further explored the practical implications of the differences in the formulations of Muscle and PARAFAC2-based Fast-Higashi in Section S4.

3.2. *Integrative Framework of Muscle Improves Cell Type Clustering in the Multi-Modal Setting*

After establishing Muscle's on par performance with existing methods in the single modality setting, we turned our attention to the integrative framework. We used the Lee et al. (2019) and Liu et al. (2021) sn-m3C-seq datasets, which simultaneously profiled 3D genome and DNA methylation in 14 human brain prefrontal cortex cell types and 10 mouse hippocampal cell types, respectively. In the integrative analysis, the cell loadings $\{\mathbf{c}_r \in \mathbb{R}_+^C | r \in [R]\}$ are learnt using both data modalities. Figure 3(a), (b) provide a direct comparison of matrix factorization via singular value decomposition (SVD) using only DNA methylation components (only mCG or only mCH; top middle, top right panel of Figure 3(a), (c)) and Muscle using only the scHi-C (top left panel of Figure 3(a), (b)) with the integrative Muscle (bottom right panel of Figure 3(a), (b)). Visual inspection of Figure 3(a), (b) reveal how Muscle leverages different data modalities. Specifically, for Lee et al. (2019) data, the integrative model (bottom right panel of Figure 3(a)) provides complete separation of the inhibitory neuronal cell types (Ndnf, Pvalb, Sst, Vip cells within red dashed line boxes), while the results for scHi-C only and mCG only modalities lack such a separation (top left, top middle). While the OPC and ODC cells (cells within golden solid line boxes) are not separated in mCG and mCH only modalities (top middle, top right), these cells can be separated in the integrative analysis (bottom right). Moreover, when a more refined set of cell labels from Luo et al. (2022) is employed, Muscle result aligns with the new labels for the neuronal cells, while revealing more refined

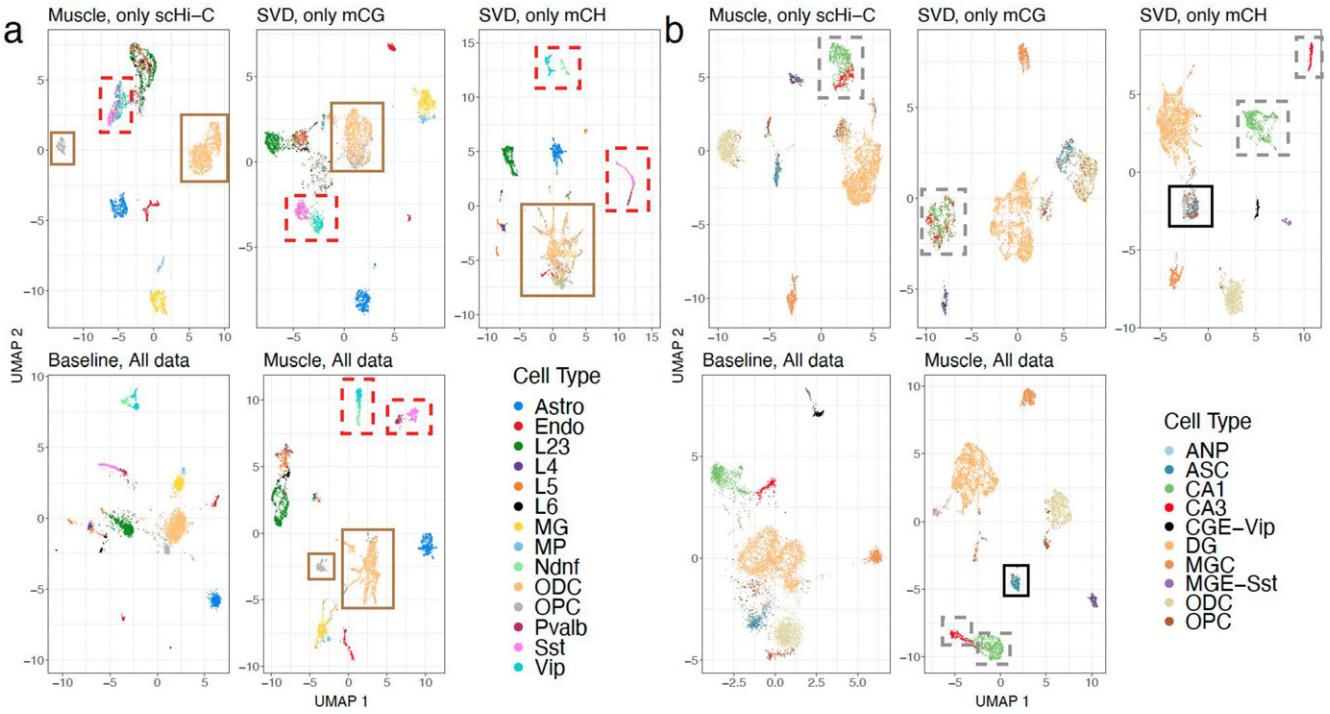

Figure 3. Graphical evaluation and benchmarking of Muscle cell clustering with the multi-modal set up. (a) and (b) (Top-left) UMAP coordinates of the cells from Lee et al. (2019) (or Liu et al. (2021)) sn-m3C-seq data based only on scHi-C modality. Muscle UMAP coordinates are obtained from estimated cell loading vectors $\{\boldsymbol{c}_r \in \mathbb{R}^C_+ | r \in [R]\}$. Cells are colored based on known cell type labels. (Top-middle) UMAP coordinates of the cells from Lee et al. (2019) (or Liu et al. 2021) sn-m3C-seq data based only on the mCG methylation modality. The UMAP coordinates are obtained from estimated cell loading vectors $\{\boldsymbol{c}_r \in \mathbb{R}^C_+ | r \in [R]\}$ of SVD. (Top-right) UMAP coordinates of the cells based only on mCH methylation modality. (Bottom-left) UMAP coordinates of the cells based on both scHi-C and DNA methylation modalities, obtained from baseline method. The UMAP coordinates are obtained from concatenation of the scVI-3D (Zheng, Shen, and Keleş 2022) embedding of scHi-C and SVD loadings of the DNA methylation datasets after normalization. (Bottom-right) UMAP coordinates of the cells based on both scHi-C and DNA methylation modalities, obtained from Muscle.

cell labels on the non-neuronal cells (Figure S4). Furthermore, when Muscle only integrates mCH and scHi-C modalities, it still achieves the separation of the OPC, ODC, and Endo cells. This illustrates that while the mCH modality completely mixes these cell types, leveraging scHi-C modality (Figure S5) aids in their separation. We also compared the Muscle cell clustering results in the multi-modal setting (Figure 3(a) bottom right) with a baseline multi-modal approach (Figure 3(a) bottom left), where we leveraged scVI-3D (Zheng, Shen, and Keleş 2022) for the analysis of the scHi-C modality and concatenated the resulting cell embeddings with the ones obtained from SVD on mCG, mCH matrices after normalization. Overall, the cell type separation from this baseline approach (Figure 3(a) bottom left panel) is inferior to that of Muscle depicted in Figure 3(a), especially for the ODC and OPC cells and the Pvalb and Sst cells. The overall performances reveal the marked improvement by the Muscle multi modal setting (left two columns of Table 1).

Analysis of a more recent sn-m3C-seq dataset from mouse hippocampus (Liu et al. 2021) provided insights similar to those of the above analysis. Specifically, results from the analysis of this dataset revealed that the integrative Muscle enabled complete separation of the cell types CA1 and CA3 (cells within gray dashed line boxes in Figure 3(b) bottom right). These cell types appeared to be less separated in the scHi-C only and mCG only analysis (Figure 3(b) top left and middle). In this case, Muscle leveraged the cell type separation information of CA1 and CA3 cells from the mCH modality (Figure 3(b) top right). Similarly, while delineation of the cell type ASC (cells within black solid line box) from only the mCH modality exhibited ambiguity (Figure 3(b) top right), the integrative Muscle model achieved good separation of this cell type from the others (Figure 3(b) bottom right) by leveraging the cell type separation information from the scHi-C modality (Figure 3(b) top left). The overall performances of these settings are summarized in the right two columns of Table 1.

Table 1. Numerical evaluation and benchmarking of Muscle cell clustering with the multi-modal set up.

	Lee et al.		Liu et al.	
Modality	ARI	KNN	ARI	KNN
All (Baseline)	0.67	0.93	0.61	0.85
All (Muscle)	**0.81**	**0.93**	**0.87**	**0.93**
mCG	0.58	0.82	0.57	0.86
mCH	0.59	0.83	0.83	0.92
scHi-C	0.80	0.85	0.67	0.89

NOTE: For each of the Lee et al. (2019) and Liu et al. (2021) data, two metrics are investigated for the evaluation. Left : The ARI scores from *k*-means clustering of the cells with the learned embeddings under single (scHi-C only, mCG only, mCH only) and multi-modal settings (Muscle, baseline). Right : K Nearest Neighborhood (KNN) classification accuracy with cell loadings as features and gold standard cell type labels as classes. KNN results are averaged over 20 sets of training-test data splits where test data harbored 10% of the randomly selected cells. The number of neighbors was set as $K = 20$. The bold value for each column denotes the largest (ARI or KNN) score for the column.

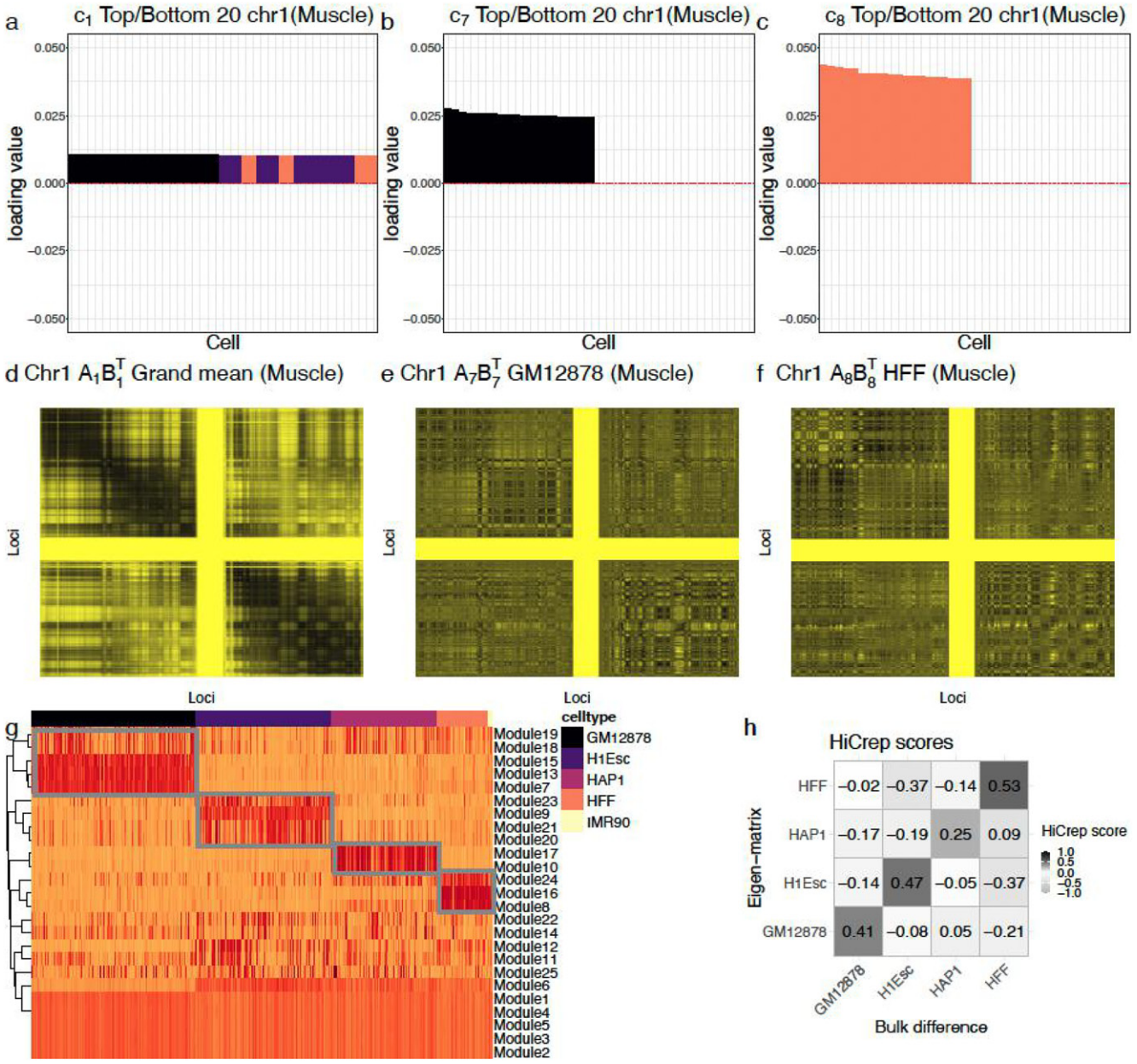

Figure 4. Cell type specific eigen contact captures by Muscle. (a)–(c) Muscle cell loading vectors c_1, c_7, c_8 depicted as barplots, respectively. The cell loading vectors are constrained to be nonnegative and yield the representative cell type of each module. Bars are colored based on the true cell type labels of the cells in panel (g). Top and bottom 20 cell loadings are displayed for brevity. (d)–(f) Visualization of Muscle eigen contact matrices ($\mathbf{A}_{1,r}\mathbf{B}_{1,r}^T$ terms) for $r = 1, 7, 8$, which capture grand average, GM12878 and HFF contact patterns. In panels (d)–(f), darker entries indicate higher interactions between the loci. (g) Heatmap of the entire set of Muscle cell loading vectors $\{\mathbf{c}_r \in \mathbb{R}_+^C | r \in [R]\}$. Each row displays the estimated cell loadings $\mathbf{c}_r$ of the module and each column corresponds to a cell. (h) HiCrep score comparison of Muscle eigen contact matrices $\mathbf{A}_{1,r}\mathbf{B}_{1,r}^T$ of modules $r = 7$ (GM12878), $r = 8$ (HFF), $r = 9$ (H1Esc), and $r = 10$ (HAP1) against the gold standard cell type specific bulk contact matrices. The y-axis denotes inferred cell type specific eigen contact matrices and the x-axis denotes the cell type specific bulk contact matrices.

3.3. *Muscle Yields Cell Type Specific Modules that Delineate Cell Type Specific Contact Matrices*

Next, we explored the inference readily available from Muscle for downstream scHi-C analysis with the Kim et al. (2020) dataset that harbored five human cell lines (GM12878, H1Esc, HAP1, HFF, and IMR90). Muscle's rank-1 modules, $(\mathbf{A}_{chr,r}\mathbf{B}_{chr,r}^T) \circ \mathbf{c}_r$, capture parsimonious representations of the cell types. The magnitudes of the Muscle cell loading vectors, $\mathbf{c}_r \geq 0$, delineate cells corresponding to the same cell type/state, therefore, linking modules to specific cell groups. Consequently, the corresponding eigen contact, $\mathbf{A}_{chr,r}\mathbf{B}_{chr,r}^T$, for module r describe the denoised specific contact matrix for the corresponding cell type. For example, Muscle cell loading vector c_1 has constant values across the cells (Figure 4(a)). Hence, this module can be interpreted as a grand mean pattern of the entire cell types, and the eigen contact matrix $\mathbf{A}_{1,1}\mathbf{B}_{1,1}^T$ (Figure 4(d)) corresponds to the grand mean contact matrix. We further note that c_7 has exclusively large values for the GM12878 cells (Figure 4(b)); hence, the eigen contact matrix $\mathbf{A}_{1,7}\mathbf{B}_{1,7}^T$ (Figure 4(e)) corresponds to the mean contact pattern of chr 1 in GM12878 cells adjusting for the grand mean pattern captured by the first module. Similarly, $\mathbf{A}_{1,8}\mathbf{B}_{1,8}^T$ (Figure 4(f)) displays the HFF specific contact pattern

because the cell loading vector c_8 of this module is specific to cell type HFF, that is, with large positive entries for HFF cells (Figure 4(c)). Investigation of this type of module identification by cell loading vectors for neuronal cells of Lee et al. (2019) are reported in Figure S6.

To validate that the eigen contact matrices, $\mathbf{A}_{1,r}\mathbf{B}_{1,r}^T$, are indeed cell type specific, we calculated HiCRep scores (Yang et al. 2017), a modified version of Spearman correlation to compare two Hi-C contact matrices, between cell type specific contact matrices ($\mathbf{A}_{1,r}\mathbf{B}_{1,r}^T$, where different r values correspond to different cell types) and cell type specific contact matrices generated from the gold standard cell type specific bulk data (Kim et al. 2020). The heatmap in Figure 4(g) displays the entire cell loading vectors $\mathbf{c}_r$ and clearly demarcates modules specific to each cell type. While several cell types have multiple modules, we chose $r \in \{7, 8, 9, 10\}$, each of which had the largest size $\|\mathbf{A}_{1,r}\|_F$ (i.e., parameters with the largest sizes) among the modules corresponding to the cell types GM12878, H1Esc, HFF, and HAP1, respectively. Figure 4(h) demonstrates that the similarity score is the highest when the eigen contact matrix $\mathbf{A}_{1,r}\mathbf{B}_{1,r}^T$ of a cell type is compared against its own gold standard data (i.e., scores along the diagonal). Collectively, these results further proffer the main advantages of Muscle's tensor decomposition framework which targets the key parameters. Further exemplary exploration of Muscle's eigen contact maps to investigate cell type specific gene regulation or the associations between additional experimental variables and contact patterns of locus-pairs are illustrated in Section S5.

3.4. Muscle Yields Cell Type Specific TADs and A/B Compartments

Topologically associating domains (TADs) constitute large genomic regions with larger numbers of interactions between loci within the region compared to interactions of loci with the loci outside the region. TADs are highly cell type specific since they recapitulate cell type specific regulation (Yu and Ren 2017). Muscle parameters $\mathbf{A}_{chr,r}$ reveal TADs for module r, and the cell type of the module is delineated by the positive entries of $\mathbf{c}_r$. We note that for elucidating TADs, the loci loading $\mathbf{A}_{chr,r}$ is used instead of $\mathbf{B}_{chr,r}$, which are identical up to module magnitude multiplication. $\mathbf{A}_{chr,r}$ is more appropriate for inference since it absorbs the magnitude of the module r (see (3)).

We next investigated the TADs for the Kim et al. (2020) data analyzed in the previous section. Figure S9(a) displays the UMAP of loci loadings, $\mathbf{A}_{1,7}$, of module 7, which corresponds to cell type GM12878 (as depicted in Figure 4(b)), concatenated with $\mathbf{A}_{1,1}$ which captures the grand average pattern across all the cells (as depicted in Figure 4(a)), for chr 1. Labeling these loci according to the gold standard TADs identified from GM12878 bulk data (Figure S9(b)) reveals that the loci loadings of Muscle organize the loci within a chromosome in a way consistent with their TAD structures. Next, we formally evaluated the performance of TAD calling from estimated Muscle loci loadings by regarding the known TADs from a cell type's bulk contact map as the gold standard. TADs from Muscle are identified by a k-means approach on the aggregated loci loadings matrices ($A_{chr,1}$ and $A_{chr,r}$) and compared with the gold standard based on precision and recall metrics. For both the gold standard TADs and the Muscle inferred TADs, we defined the TAD closest to the first diagonal entry of a contact map as TAD #1 and defined the TAD closest to the last diagonal entry as TAD #K. We carried out the analysis throughout chromosomes and the four major cell types for the Kim et al. (2020) dataset (Figure 5(a)). This analysis yielded precision and recall values around 80%–82% across cell types. Further comparison of Muscle's loci loading clustering based TAD interference to a pseudo-bulkification (aggregation over cells from a cell type) based TAD inference on the Muscle's denoised tensor $\hat{\mathcal{M}}_{chr}$ is provided in Figure S10(a).

A/B compartments, which constitute genome territories with high (A) or low (B) gene expression compared to other territories, is another class of genome compartmentalization that can be inferred from scHi-C data (Lieberman-Aiden et al. 2009). In addition to identifying TADs, the first column vector of $\mathbf{A}_{chr,r}$ for each $chr \in [Chr]$ and $r \in [R]$ provides the A/B compartment structures in a cell type specific manner. This is because the loci loadings $\mathbf{A}_{chr,r}, \mathbf{B}_{chr,r}$ are obtained from an eigen decomposition of the projected scHi-C tensor $\mathcal{Y}_{chr}$ onto the subspace spanned by $\mathbf{c}_r$ (Algorithm 1), and hence, the first column of $\mathbf{A}_{chr,r}$ is formed by the multiplication of the largest eigenvalue and the corresponding eigenvector. Consequently, the first column of $\mathbf{A}_{chr,r}$ captures the major contact pattern of the module r, which would represent the largest scale genome territory, that is, A/B compartments. Figure 5(c) displays the exact same loci clustering of $\mathbf{A}_{1,7}$ as in Figure S9(a) where the labels of loci are obtained by the sign of the first column of $\mathbf{A}_{1,7}$. The loci with blue colors in Figure 5(c) are inferred to be in A compartments, while the loci with red colors are in the B compartments. We validated this labeling by comparing it to the labeling from the gold standard GM12878 bulk Hi-C data's A/B compartment structure (Figure 5(d)). We further evaluated the A/B compartment inference for each cell type (as identified by modules $r \in \{7, 8, 9, 10\}$) by comparing inferred compartmentalizations (averaged over all the 23 chromosomes) with those from the cell type specific bulk data (Figure 5(b)). This evaluation yielded that, on average, Muscle identified 75% of the gold standard A and B compartment loci correctly. We further evaluated the correlation between Muscle loci loading vector and the PC1 of the cell type specific bulk contact map, which are used for inferring A/B compartments (Figure S10(c)). Overall, the estimation targets of Muscle directly match the parameters of interest in scHi-C data analysis and the estimated parameters readily reveal TAD and A/B compartment structures of the cell types without additional downstream analysis. As a remark, the pseudo-bulk analysis of Muscle for TAD and A/B compartments aligns quite well with those from loci loading matrices (Figure S10(a) and (b)), supporting that each rank-1 module of Muscle captures general 3D genome characteristics. However, a more refined result for the TAD and A/B compartment can be obtained from the pseudo-bulkification of the denoised tensor, which is a collective of the Muscle parameters (Figure S11).

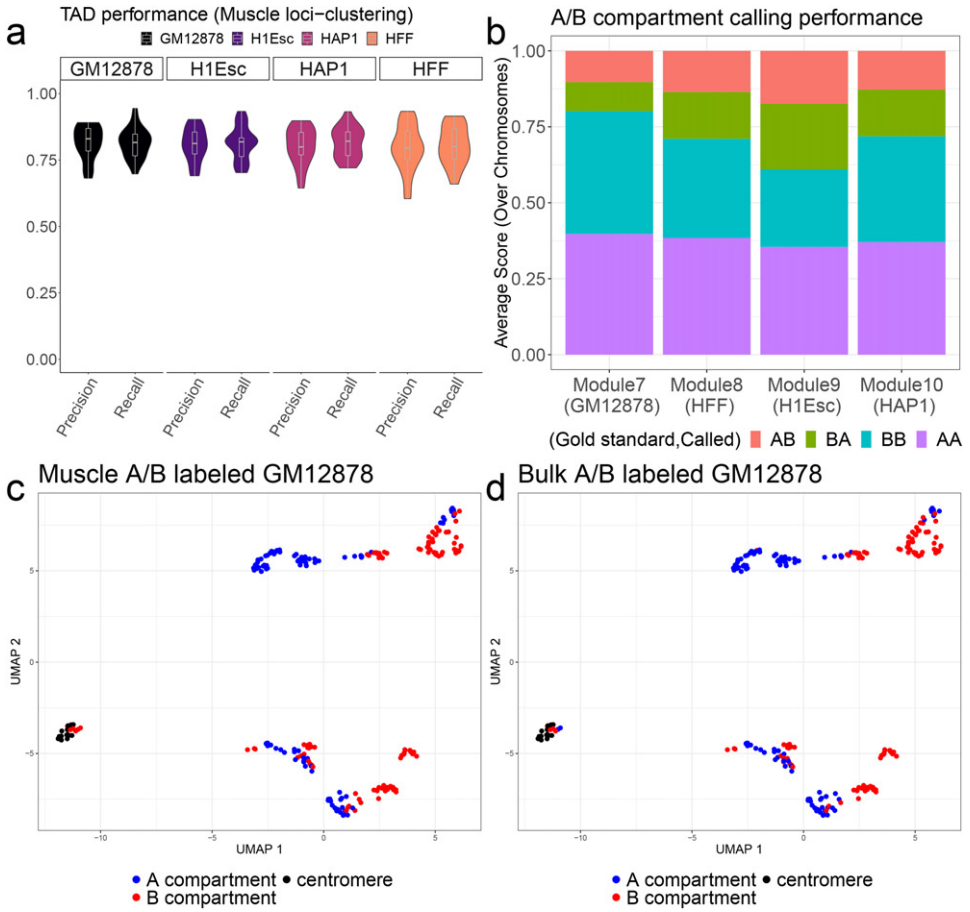

Figure 5. TAD and A/B compartment identification from estimated Muscle parameters. (a) Evaluation of TAD inference of Muscle based on the gold standard TADs from cell type specific external bulk Hi-C data. Recall is evaluated by fixing each gold standard TAD and calculating the proportion of loci contained in the corresponding TAD identified by Muscle loci loading clustering. For precision, each Muscle inferred TAD is fixed and the proportion of loci contained in the corresponding gold standard TAD is calculated. Within each panel, distribution of precision and recall values across all the chromosomes within a cell type are displayed. (b) Evaluation of A/B compartment territory inference of Muscle based on the gold standard A/B compartments from cell type specific bulk Hi-C data. Each barplot represents a cell type (module) and displays the mean proportion of correctly inferred A (or B) compartments averaged across the chromosomes. The first element of a label (e.g., AB) is for gold standard compartment and the other element is for inferred compartment. The regions corresponding to the centromere are excluded from the analysis. (c) UMAP coordinates of chr 1 loci from Figure S9(a) colored with respect to the signs of the first column of $\mathbf{A}_{1,7}$, with "+" depicted in blue and corresponding to A compartment and "-" depicted in red and corresponding to B compartment. (d) UMAP coordinates of chr 1 loci from Figure S9(a) colored with respect to the gold standard A/B compartment results from bulk GM12878 Hi-C data.

3.5. *Muscle Unveils Cell Type Specific Associations between Chromatin Conformation Structures Inferred from scHi-C Data Modality and DNA Methylation Modality*

DNA methylation in both the CpG and non-CpG sites is generally negatively correlated with the gene expression levels in mammalian neurons (Lister et al. 2013; Luo, Hajkova, and Ecker 2018). Using Muscle integrative analysis results of Lee et al. (2019) sn-m3C-seq data, we explored whether the A/B compartment structure of the loci inferred from scHi-C modality loci loadings $\mathbf{A}_{chr,r}[\ ,1]$ associated with the Muscle denoised methylation level for each cell type corresponding to the module r. Specifically, we investigated the association between genome compartmentalization and DNA methylation pattern in an excitatory neuronal cell type (L5) and in an inhibitory neuronal cell type (Vip). These two cell types are well-separated in the integrative Muscle analysis (bottom right panel of Figure 3(a)). We observed that there exist association between fitted methylation (aggregated over a cell type) and A/B compartmentalization (Figure 6(a) and (d)) with A compartment having lower DNA methylation level in general. It also revealed that the association is cell type specific. In Vip cells, loci in B compartment territory have significantly higher methylation levels on CpG sites than those of the loci in the A compartment ($p = 9.7 \times 10^{-6}$), while the association in L5 cells was not as significant as that of Vip cells ($p = 0.07$). While the non-CpG methylation over the A and B compartment showed same directionality for the association as CpG methylation, the association strength does not seem clear across the two cell types (with $p = 0.03$ and $p = 0.34$). Evaluation of association between pseudo-bulk methylation and A/B compartmentalization from PC1 of pseudo-bulk Hi-C contact map confirmed this observation (Figure S13). These reinstate that association of loci methylation levels with genome territorial structures is cell and methylation site type specific.

We next exploited the integrative analysis results from the point of CCCTC-binding Factor (CTCF) DNA binding protein. The activities of CTCF are inhibited by DNA methylation around the CTCF binding sites (Wang et al. 2012). In particular, DNA methylation plays a significant role in disruption of

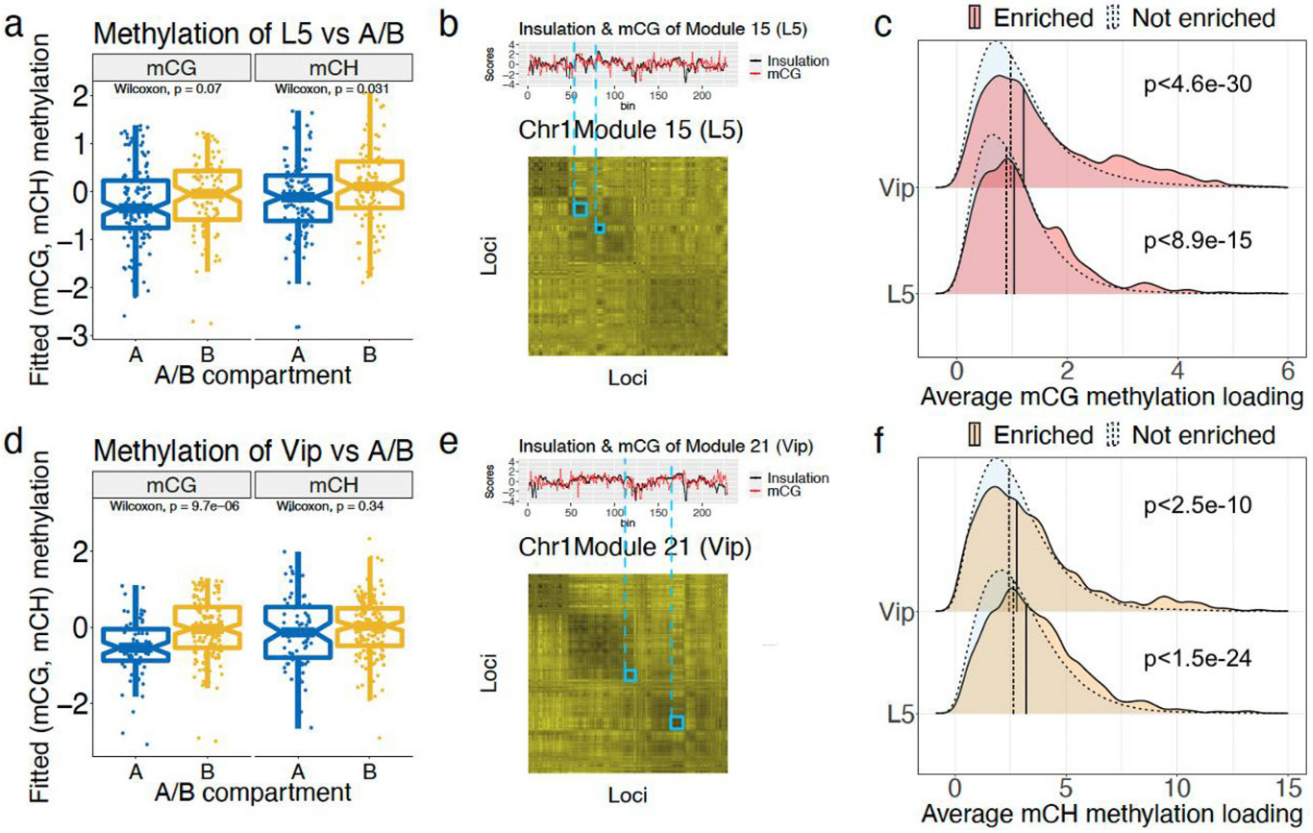

Figure 6. Association analysis of methylation patterns with the broader 3D chromatin structures. (a) and (d) Distributions of fitted (and scaled) methylation levels averaged over cell type L5(a) and Vip(d) stratified with respect to CpG (mCG) and non-CpG (mCH) and within A/B compartments for L5 (panel a, identified from $\mathbf{A}_{1,15}[\,,1]$) and Vip (panel d, identified from $\mathbf{A}_{1,21}[\,,1]$) cells. Differences in methylation levels are evaluated with a Wilcoxon rank-sum test. (b) and (e) Top row of the panel: the dotted red line displays the inferred methylation pattern along chr 1 loci, scaled as $\mathbf{v}_1^{CG}/\left\|\mathbf{v}_1^{CG}\right\| + \mathbf{v}_r^{CG}/\left\|\mathbf{v}_r^{CG}\right\|$. The solid black line represents insulation scores obtained from scaled eigen contact matrices, $\mathbf{A}_{1,1}\mathbf{B}_{1,1}^T/\left\|\mathbf{A}_{1,1}\mathbf{B}_{1,1}^T\right\|_F + \mathbf{A}_{1,r}\mathbf{B}_{1,r}^T/\left\|\mathbf{A}_{1,r}\mathbf{B}_{1,r}^T\right\|_F$. Bottom rows of the panels display scaled eigen contact matrices. In (b) and (e) methylation loadings of loci are used to highlight the major patterns captured by Muscle, while (a) and (d) showcase the entire summarization by the Muscle fit. (c) and (f) Comparison of average absolute mCG methylation loadings of loci for locus-pair grouped as "Enriched" (with absolute eigen contact values, i.e., differential interactions, in the top 0.5%) and "Not-enriched" (with absolute eigen contact values below the top 0.5%) for Vip (top row) and L5 cell type (bottom row). Vertical lines mark the median of the distributions. The results for other cell types are available in Figure S8.

CTCF binding around key tumor suppressing genes in cancer (Rodriguez et al. 2010). CTCF is also a key player in folding of chromatin into domains. Specifically, TAD boundaries, where cohesin and CTCF form a DNA binding complex to hold the DNA loops together, are enriched for CTCF binding sites (Rao et al. 2014; Pombo and Dillon 2015). Furthermore, previous studies on hierarchical structure of the genome, that is, meta-TADs sizes of which are on average 10Mb (Esposito et al. 2020), found that the CTCF binding activity is enriched at the boundaries of meta-TADs, while the enrichment is maximized at the scale of TADs (Fraser et al. 2015; Zhan et al. 2017). As a result, we expect that the meta-TAD boundary regions have more CTCF binding, and hence less DNA methylation that would hinder CTCF binding activity. Figure 6(b), (e) display methylation patterns from the methylation loci loading parameter $\mathbf{v}_r^{CG}$ and the insulation scores (Gong et al. 2018), which quantify how *unlikely* a locus is to be a (meta-)TAD boundary, from the eigen contact matrix. These comparisons reveal that insulation score patterns align with the large-scale methylation patterns. Considering the concept of meta-TADs at this 1Mb resolution, it further corroborates that genomic loci that are likely to be at meta-TAD boundaries (i.e., with a drop in insulation scores marked with the blue dashed lines) have low methylation levels.

Finally, we asked whether an exploratory analysis of module-specific Muscle eigen contact maps and methylation loci loadings can yield associations between methylation levels and differential interactions. Recalling that an eigen-contact map captures a cell type's differential interactions compared to the grand mean and the methylation loading summarizes the cell type's differential methylation pattern compared to the mean, we queried how the differential methylation patterns of locus-pairs with differential interactions for a cell type varied. Figure 6(c) and (f) highlight that the loci enriched for differential interactions have significantly higher levels of differential methylation for both the Vip and L5 cells and the methylation sites (CpG sites, $p < 8.9e-15$ for L5, $p < 4.6e-30$ for Vip, and non-CpG sites, $p < 1.5e-24$ for L5, $p < 2.5e-10$ for Vip with one-sided Wilcoxon rank sum test). These results, obtained directly from the estimated Muscle parameters, indicate that the parts of the genome with differential contacts also exhibit higher levels of differential methylation.

4. Simulation Studies

Datasets with known cell types enabled us to illustrate the superior performance of the joint analysis with Muscle against both the single modality analysis and a baseline integrative approach. We further studied advantages of the tensor decomposition framework of Muscle over the matricization-based baseline method with simulation experiments. In these experiments, we ensured that the data generation process does not conform with Muscle's model (given in (1)–(4)) to quantify Muscle's robustness against model misspecification.

4.1. Data Generation

The scHi-C tensor $\mathcal{Y} \in \mathbb{R}^{40\times 40\times 120}$ and the DNA methylation matrix $Y \in \mathbb{R}^{40\times 120}$, for 40 genomic loci and 120 cells across three cell types (with 40 cells from each cell type), were simulated from the following Negative binomial models:

$$\mathcal{Y}_{ijc} \overset{\text{iid}}{\sim} NB\left(\mathcal{M}_{ijc}, \text{size} = \frac{\mathcal{M}_{ijc}}{\phi_1 - 1}\right), \quad \text{for all } i \in [40], j \in [40], c \in [120]$$

$$\mathbf{Y}_{lc} \overset{\text{iid}}{\sim} NB\left(\mathbf{M}_{lc}, \text{size} = \frac{\mathbf{M}_{lc}}{\phi_2 - 1}\right), \quad \text{for all } l \in [40], c \in [120]$$

$$\mathcal{M} = \sum_{r=1}^{3} (A_r B_r^T) \circ \mathbf{c}_r, \quad \mathbf{A}_r, \mathbf{B}_r \in \mathbb{R}^{40\times 2},$$

$$\mathbf{M} = \sum_{r=1}^{3} \mathbf{v}_r \circ \mathbf{c}_r, \quad \mathbf{c}_r \geq 0, \quad \|\mathbf{c}_r\| = 1,$$

$$\mathbf{v}_r \in \mathbb{R}^{40}, \quad \frac{\phi_1}{\phi_2} = \frac{N_h}{N_m} = 40.$$

We generated three cell types by setting the entries 1 to 40 of c_1, 41 to 80 of c_2, and 81 to 120 of c_3 to 3.1 and all the other entries of the cell loading vectors to 1 before size normalization. For each module r, $r \in [3]$, the first column of loci loading matrix $\mathbf{A}_r \in \mathbb{R}^{40\times 2}$ is set to represent the A/B compartment structure (checker board-like pattern) of a contact matrix and the other column is set to represent a single TAD structure (square box-like pattern) along the diagonals of a contact matrix. The scHi-C loci loading matrix $\mathbf{B}_r \in \mathbb{R}^{40\times 2}$ is a column-wise normalization of $\mathbf{A}_r$ so that it becomes equivalent to $\mathbf{A}_r$ up to module size magnitude. This formulation, in turn, generates each eigen contact of the contact matrix $\mathbf{A}_r\mathbf{B}_r^T$ as in Figure S14(a)–(c). The methylation loadings $\mathbf{v}_r$ are randomly generated from a Poisson distribution with rate parameter $\lambda = 0.23$. The constructed methylation modules $\mathbf{v}_r \circ \mathbf{c}_r$ are displayed in Figure S14(j)–(l). The distributional assumptions on $\mathcal{Y}$ and Y result in

$$\mathbb{E}[\mathcal{Y}_{ijc}] = \mathcal{M}_{ijc}, \quad Var(\mathcal{Y}_{ijc}) = \mathcal{M}_{ijc} + \frac{\mathcal{M}_{ijc}^2}{\mathcal{M}_{ijc}}(\phi_1 - 1) = \mathcal{M}_{ijc}\phi_1$$

$$\mathbb{E}[\mathbf{Y}_{lc}] = \mathbf{M}_{lc}, \quad Var(\mathbf{Y}_{lc}) = \mathbf{M}_{lc} + \frac{\mathbf{M}_{lc}^2}{\mathbf{M}_{lc}}(\phi_2 - 1) = \mathbf{M}_{lc}\phi_2.$$

This data generation set up further ensures that $\mathcal{M}_{ijc} \approx \mathbf{M}_{lc}$ $\forall i, j, l, c$ (Figure S15(e)). Consequently, the proportion of the variances between two data modalities approximately satisfies $\text{var}(\mathcal{Y}_{ijc})/\text{var}(\mathbf{Y}_{lc}) \approx \phi_1/\phi_2$, and allows us to vary the proportion of the variances of the two sources of data by modulating ϕ_1, ϕ_2. Under this data generation scheme, the resulting scHi-C and DNA methylation data exhibit general characteristics of the observed scHi-C and DNA methylation datasets (examples are provided in Figure S15(a)–(d)).

4.2. Simulation Results

We varied the noise level of the methylation data as $\phi_2 \in \{1.1, 1.2, \ldots, 3\}$ and that of the scHi-C data ϕ_1 is automatically determined based on the proportional variance construction. For each of the ϕ_2 values, we generated 100 simulation replicates and quantified the performances of Muscle and the matricization-based baseline method. For each Muscle fit, the proportion of the variance was set to 40. The rank R was set $R = 4$ for both methods.

Comparison of the cell clustering of the methods revealed that Muscle results in significantly higher ARI scores than the baseline method, with a median difference of 0.1 across all the noise levels, ϕ_2 (Figure 7, $p < 2e-16$ with t-test, when adjusted for the noise level).

We next investigated how well each method recovers the true means $\mathcal{M}$ and $\mathbf{M}$. Specifically, we evaluated the Spearman correlation between the true $\mathcal{M}$ and estimated $\hat{\mathcal{M}}$ for the scHi-C modality and the Spearman correlation between the true $\mathbf{M}$

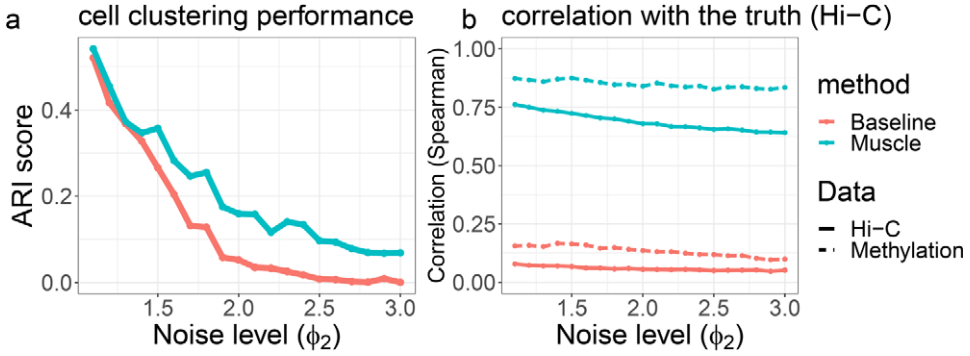

Figure 7. Evaluation of multi-modality analysis by Muscle and the matricization-based baseline method with simulation studies. (a) ARI scores of the methods across all the noise levels ϕ_2 averaged over 100 replicates. (b) Solid lines: Spearman correlations of the methods between the true scHi-C mean tensor $\mathcal{M}$ and the estimated mean tensor $\hat{\mathcal{M}}$ across noise levels ϕ_2 averaged over 100 replicates. Dashed lines: Spearman correlations of the methods between the true mean DNA methylation matrix M and the estimated mean methylation matrix $\hat{M}$ across noise levels ϕ_2 averaged over 100 replicates.

and estimated $\hat{\mathbf{M}}$ for the DNA methylation modality for both of the methods. The solid lines in Figure 7(b) illustrates that the estimates from the baseline model have almost zero correlations with the true mean scHi-C contact matrices, whereas Muscle estimates have markedly higher correlation values ($\approx$ 0.7) across all the noise levels ϕ_2. This is also evident visually from Figure S14. While the baseline method results in markedly noisy eigen contacts (Figure S14(g)–(i)), the Muscle eigen contacts reasonably capture the cell type specific modules (Figure S14(d)–(f)). Likewise, the dashed lines in Figure 7(b) also illustrates that the baseline model results in low correlations with the true mean DNA methylation matrix ($\approx$ 0.15), while Muscle achieves markedly higher correlation values ($\approx$ 0.9) across all the noise levels ϕ_2. This can also be visualized in Figure S14(j)–(r).

More detailed results on the cell clustering and the recovery of the mean scHi-C tensor $\mathcal{M}$ and mean methylation matrix $\mathbf{M}$ are provided in Figure S16. These specifically summarize the results based on the setting with the noise level $\phi_2 = 1.1$. The UMAP plots of the cell loadings depicted in Figure S16(a)–(b) show that Muscle exhibits more apparent cell clustering than the baseline method (clustering ARIs of 0.3 and 0.5 for the baseline method and Muscle, respectively). Figure S16(c)–(d) directly compare the true and the estimated mean methylation matrices and indicate that Muscle's recovery of the mean methylation matrix better aligns with the true $\mathbf{M}$ compared to that of the baseline method. In addition, association between the true mean scHi-C $\mathcal{M}$ and the Muscle estimate is more evident compared to the estimate from the baseline method (Figure S16(e)–(f)).

In order to further explore the advantages of Muscle model parameterization, we carried out a comparison between cell type specific contact matrices estimated by the pseudo-bulkified denoised data from the baseline method and the Muscle eigen contact matrix. For the baseline method (Figure S16(a)), we leveraged the cell labels from k-means clustering to generate cell type specific pseudo-bulk differential contact matrices as estimators of the cell type specific contact patterns (Figure S17(a)). In contrast, Muscle first identified which cell type each rank-1 module belonged to based on the cell loading vector (c_r) (Figure S17(d)), and reported the corresponding eigen contact matrix ($A_r B_r^T$) as the estimator of the cell type specific contact patterns (Figure S17(c)). It is evident that the estimator derived from Muscle is markedly less noisy and more similar to the true contact pattern of cell type C (depicted in Figure S17(b)) than the one from the pseudo-bulkified denoised data of the baseline method. Furthermore, evaluations across 100 simulation replicates yielded that Muscle eigen contact matrix correlated markedly better with the true contact matrix with a correlation of 0.69, while that of the baseline method yielded a correlation of 0.22.

In addition, we also compared loop calling on the estimated contact matrices. Specifically, we compared the loops identified from cell type C's (a) true mean contact matrix; (b) Muscle eigen contact matrix (Module 1+Module 4); (c) The pseudo-bulkified contact matrix of the baseline method. The loops were identified by Fit-HiC (Ay, Bailey, and Noble 2014) for all the contact matrices. This comparison revealed that 65% of the true significant loops (from true mean contact matrix) were identified based on Muscle eigen contact matrix (q-value$<$0.1) and 71% of the loops identified from Muscle eigen contact matrix were contained in the true significant loops. In contrast, 52% of the true significant loops were identified based on the pseudo-bulkified contact matrix of the baseline method and 64% of the loops identified based on the baseline method were contained in the true set. Overall, these results highlight the power of Muscle estimated parameters for downstream analysis.

5. Discussion

We presented Muscle as a joint tensor decomposition framework for integrative analysis of scHi-C and DNA methylation data. Computational experiments with labeled real data and simulated data demonstrated that Muscle's integrative framework can leverage multiple single cell data modalities to enhance cell type identification. Furthermore, Muscle performed on par or better than existing approaches when presented with single modality scHi-C data, and natural baseline approaches based on concatenation for integrative analysis. In addition to on par performance in cell type identification, Muscle exhibits clear advantages over the baseline approaches for the integrative analysis in terms of statistical interpretation, algorithmic optimality, and information transfer between modalities (Section S6). Notably, as a key advantage, we showcased how Muscle's parameterization encodes key parameters of interest (cell type specific contact matrices, TADs, A/B compartments).

In applications of Muscle, we observed that the Muscle cell loading parameter that is shared across multiple modalities plays a critical role in integrative inference. This parameter can be sensitive to the level of variability between the data modalities, necessitating appropriate modeling of the variance terms to balance the contribution of different modalities during integration. We used a proportional variance assumption for the scHi-C and methylation modalities and were able to capitalize on the discriminative abilities of the individual modalities for cell types (Figure 3). A more flexible variance modeling approach might be beneficial for integration of additional data modalities. Correcting for batch effects can have critical implications for the single cell data analysis (Korsunsky et al. 2019; Zheng, Shen, and Keleş 2022). Muscle, in general, relies on excluding rank-1 modules, cell loadings of which significantly associate with the batch labels, and empirical observations suggest that the band debiasing step of Muscle described in Section S3 corrects for mild batch effects as in Zheng, Shen, and Keleş (2022) (Figure S3). While Muscle relied on the Gaussian distribution assumption on the transformed count data, it can be relaxed with tensor models incorporating count distributions (Hong, Kolda, and Duersch 2020). Another important point in tensor analysis is the selection of the tensor rank. Rank determination in tensors is an NP-hard problem (Håstad 1990) and Fast-Higashi relies on relatively high rank. In Muscle, we employed a heuristic approach to penalize over-fitting; however, a direct regularization, for example, group LASSO (Yuan and Lin 2006), on rank-1 components could also be employed. In addition, generalizing the block term decomposition rank K_{chr} to be specific to each module across $r \in [R]$ with a more explicit guidance on

the choice with a theoretical guarantee could be an interesting extension for the Muscle model.

Lastly, while Muscle provides integration, inference, and interpretation advantages compared to alternative methods, its current implementation is relatively slow compared to some of the fast scHi-C analysis methods and warrants further advancement. Specifically, for Lee et al. (2019) scHi-C data at 1Mb resolution, an unoptimized implementation of Muscle required 18 hr (23 cores CPU), while Higashi took 49 hr (10 cores CPU), scHiC Topics took 36 hr (1 core CPU), scVI-3D took 4 hr (23 cores GPU), Fast-Higashi took 1 hr (23 cores CPU), scHiCluster took 30 min (23 cores CPU), and BandNorm took 15 min (23 cores CPU). The speed bottleneck of Muscle is mainly due to additional decomposition steps for estimating loci loadings, which encode key downstream parameters of interests, and warrants further computational developments.

Supplementary Materials

The supplementary materials include 1. Details of the Muscle algorithm, 2. Proof of Lemmas for the Muscle algorithm updates, 3. Details on data processing, 4. Comparison of Muscle tensor formulation against PARAFAC2 formulation, 5. Additional downstream analysis, 6. Discussion of additional advantages of Muscle over the baseline multi-modal methods, 7. Additional figures from Figure S1 to S17, 8. Author Contributions Checklist (ACC) Form.

Acknowledgments

We thank Siqi Shen from the University of Wisconsin-Madison for his assistance with the raw data pre-processing. We thank Siqi Shen and Dr. Ye Zheng (Fred Hutchinson Cancer Center) for sharing detailed results of their published work. We also thank Dr. Hanbaek Lyu, Chanwoo Lee, and Coleman Breen from the University of Wisconsin-Madison for insightful discussions.

Disclosure Statement

The authors report there are no competing interests to declare.

Funding

This work was supported by grants from the National Human Genome Research Institute (ID: HG003747, HG012881, HG011371) and Chan Zuckerberg Initiative (ID: DI-0000000113).

ORCID

Kwangmoon Park http://orcid.org/0000-0002-0987-923X
Sündüz Keleş http://orcid.org/0000-0001-9048-0922

References

Ay, F., Bailey, T. L., and Noble, T. L. (2014), "Statistical Confidence Estimation for Hi-C Data Reveals Regulatory Chromatin Contacts," *Genome Research*, 24, 999–1011. [2475]

De Lathauwer, L. (2008), "Decompositions of a Higher-Order Tensor in Block Terms-Part II: Definitions and Uniqueness," *SIAM Journal on Matrix Analysis and Applications*, 30, 1033–1066. [2466,2467]

De Lathauwer, L., De Moor, B., and Vandewalle, J. (2000), "A Multilinear Singular Value Decomposition," *SIAM Journal on Matrix Analysis and Applications*, 21, 1253–1278. [2466]

De Silva, V., and Lim, L.-H. (2008), "Tensor Rank and the Ill-Posedness of the Best Low-Rank Approximation Problem," *SIAM Journal on Matrix Analysis and Applications*, 30, 1084–1127. [2466]

Esposito, A., Chiariello, A. M., Conte, M., Fiorillo, L., Musella, F., Sciarretta, R., and Bianco, S. (2020), "Higher-Order Chromosome Structures Investigated by Polymer Physics in Cellular Morphogenesis and Differentiation," *Journal of Molecular Biology*, 432, 701–711. [2473]

Fraser, J., Ferrai, C., Chiariello, A. M., Schueler, M., Rito, T., Laudanno, G., Barbieri, M., Moore, B. L., Kraemer, D. C., Aitken, S., et al. (2015), "Hierarchical Folding and Reorganization of Chromosomes are Linked to Transcriptional Changes in Cellular Differentiation," *Molecular Systems Biology*, 11, 852. [2473]

Gong, Y., Lazaris, C., Sakellaropoulos, T., Lozano, A., Kambadur, P., Ntziachristos, P., Aifantis, I., and Tsirigos, A. (2018), "Stratification of TAD Boundaries Reveals Preferential Insulation of Super-Enhancers by Strong Boundaries," *Nature Communications*, 9, 1–12. [2473]

Håstad, J. (1990), "Tensor Rank is NP-Complete," *Journal of Algorithms*, 11, 644–654. [2475]

Hong, D., Kolda, T. G., and Duersch, J. A. (2020), "Generalized Canonical Polyadic Tensor Decomposition," *SIAM Review*, 62, 133–163. [2475]

Kiers, H. A., Ten Berge, J. M., and Bro, R. (1999), "PARAFAC2-Part I. A Direct Fitting Algorithm for the PARAFAC2 Model," *Journal of Chemometrics: A Journal of the Chemometrics Society*, 13, 275–294. [2466]

Kim, H.-J., Yardımcı, G. G., Bonora, G., Ramani, V., Liu, J., Qiu, R., Lee, C., Hesson, J., Ware, C. B., Shendure, J., et al. (2020), "Capturing Cell Type-Specific Chromatin Compartment Patterns by Applying Topic Modeling to Single-Cell Hi-C Data," *PloS Computational Biology*, 16, e1008173. [2464,2465,2466,2467,2468,2470,2471]

Kolda, T. G., and Bader, B. W. (2009), "Tensor Decompositions and Applications," *SIAM Review*, 51, 455–500. [2466]

Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P.-r., and Raychaudhuri, S. (2019), "Fast, Sensitive and Accurate Integration of Single-Cell Data with Harmony," *Nature Methods*, 16, 1289–1296. [2475]

Lee, D.-S., Luo, C., Zhou, J., Chandran, S., Rivkin, A., Bartlett, A., Nery, J., Fitzpatrick, C., O'Connor, C., Dixon, J., and Ecker, J. (2019), "Simultaneous Profiling of 3D Genome Structure and DNA Methylation in Single Human Cells," *Nature Methods*, 16, 1–8. [2464,2466,2467,2468,2469,2471,2472,2476]

Li, G., Liu, Y., Zhang, Y., Kubo, N., Yu, M., Fang, R., Kellis, M., and Ren, B. (2019), "Joint Profiling of DNA Methylation and Chromatin Architecture in Single Cells," *Nature Methods*, 16, 991–993. [2464,2466,2467,2468]

Li, X., Feng, F., Pu, H., Leung, W. Y., and Liu, J. (2021), "scHiCTools: A Computational Toolbox for Analyzing Single-Cell Hi-C Data," *PLOS Computational Biology*, 17, e1008978. [2464,2467]

Lieberman-Aiden, E., Van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, B. R., Dorschner, M. O., et al. (2009), "Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome," *Science*, 326, 289–293. [2464,2465,2471]

Lister, R., Mukamel, E. A., Nery, J. R., Urich, M., Puddifoot, C. A., Johnson, N. D., Lucero, J., Huang, Y., Dwork, A. J., Schultz, M. D., et al. (2013), "Global Epigenomic Reconfiguration during Mammalian Brain Development," *Science*, 341, 1237905. [2472]

Liu, H., Zhou, J., Tian, W., Luo, C., Bartlett, A., Aldridge, A., Lucero, J., Osteen, J. K., Nery, J. R., Chen, H., et al. (2021), "DNA Methylation Atlas of the Mouse Brain at Single-Cell Resolution," *Nature*, 598, 120–128. [2464,2466,2468,2469]

Luo, C., Hajkova, P., and Ecker, J. R. (2018), "Dynamic DNA Methylation: In the Right Place at the Right Time," *Science*, 361, 1336–1340. [2472]

Luo, C., Liu, H., Xie, F., Armand, E. J., Siletti, K., Bakken, T. E., Fang, R., Doyle, W. I., Stuart, T., Hodge, R. D., et al. (2022), "Single Nucleus Multi-Omics Identifies Human Cortical Cell Regulatory Genome Diversity," *Cell Genomics*, 2, 100107. [2468]

Pombo, A., and Dillon, N. (2015), "Three-Dimensional Genome Architecture: Players and Mechanisms," *Nature Reviews Molecular Cell Biology*, 16, 245–257. [2465,2473]

Ramani, V., Deng, X., Qiu, R., Gunderson, K., Steemers, F., Disteche, C., Noble, W., Duan, Z., and Shendure, J. (2017), "Massively Multiplex Single-Cell Hi-C," *Nature Methods*, 14, 263–266. [2464,2466,2467]

Rao, S. S., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., Sanborn, A. L., Machol, I., Omer, A. D., Lander, E. S., et al. (2014), "A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping," *Cell*, 159, 1665–1680. [2473]

Rodriguez, C., Borgel, J., Court, F., Cathala, G., Forné, T., and Piette, J. (2010), "CTCF is a DNA Methylation-Sensitive Positive Regulator of the INK/ARF Locus," *Biochemical and Biophysical Research Communications*, 392, 129–134. [2473]

Rontogiannis, A. A., Kofidis, E., and Giampouras, P. V. (2021), "Block-Term Tensor Decomposition: Model Selection and Computation," *IEEE Journal of Selected Topics in Signal Processing*, 15, 464–475. [2466]

Shen, S., Zheng, Y., and Keleş, S. (2022), "scGAD: Single-Cell Gene Associating Domain Scores for Exploratory Analysis of scHi-C Data," *Bioinformatics*, 38, 3642–3644. [2464]

Stevens, T., Lando, D., Atkinson, L. P., Cao, Y., Lee, S., Leeb, M., Wohlfahrt, K. J., Boucher, W., O'Shaughnessy-Kirwan, A., Cramard, J., Faure, A. J., Ralser, M., Blanco, E., Morey, L., Sanso, M., Palayret, M. G. S., Lehner, B., Di Croce, L., and Laue, E. (2017), "3D Structure of Individual Mammalian Genomes Studied by Single Cell Hi-C," *Nature*, 544, 59–64. [2464]

Tan, L., Ma, W., Wu, H., Zheng, Y., Xing, D., Chen, R., Li, X., Daley, N., Deisseroth, K., and Xie, X. S. (2021), "Changes in Genome Architecture and Transcriptional Dynamics Progress Independently of Sensory Experience During Post-Natal Brain Development," *Cell*, 184, 741–758. [2464,2466,2467]

Tucker, L. R. (1966), "Some Mathematical Notes on Three-Mode Factor Analysis," *Psychometrika*, 31, 279–311. [2466]

Ulianov, S. V., Zakharova, V. V., Galitsyna, A. A., Kos, P. I., Polovnikov, K. E., Flyamer, I. M., Mikhaleva, E. A., Khrameeva, E. E., Germini, D., Logacheva, M. D., et al. (2021), "Order and Stochasticity in the Folding of Individual Drosophila Genomes," *Nature Communications*, 12, 1–17. [2464]

Wang, H., Maurano, M. T., Qu, H., Varley, K. E., Gertz, J., Pauli, F., Lee, K., Canfield, T., Weaver, M., Sandstrom, R., et al. (2012), "Widespread Plasticity in CTCF Occupancy Linked to DNA Methylation," *Genome Research*, 22, 1680–1688. [2472]

Wang, M., and Li, L. (2020), "Learning from Binary Mmultiway Data: Probabilistic Tensor Decomposition and its Statistical Optimality," *Journal of Machine Learning Research*, 21, 6146–6183. [2466]

Yang, T., Zhang, F., Yardımcı, G. G., Song, F., Hardison, R. C., Noble, W. S., Yue, F., and Li, Q. (2017), "HiCRep: Assessing the Reproducibility of Hi-C Data Using a Stratum-Adjusted Correlation Coefficient," *Genome Research*, 27, 1939–1949. [2471]

Yu, M., and Ren, B. (2017), "The Three-Dimensional Organization of Mammalian Genomes," *Annual Review of Cell and Developmental Biology*, 33, 265–289. [2471]

Yuan, M., and Lin, Y. (2006), "Model Selection and Estimation in Regression with Grouped Variables," *Journal of the Royal Statistical Society*, Series B, 68, 49–67. [2475]

Zhan, Y., Mariani, L., Barozzi, I., Schulz, E. G., Blüthgen, N., Stadler, M., Tiana, G., and Giorgetti, L. (2017), "Reciprocal Insulation Analysis of Hi-C Data Shows that TADs Represent a Functionally but not Structurally Privileged Scale in the Hierarchical Folding of Chromosomes," *Genome Research*, 27, 479–490. [2473]

Zhang, A., and Xia, D. (2018), "Tensor SVD: Statistical and Computational Limits," *IEEE Transactions on Information Theory*, 64, 7311–7338. [2466]

Zhang, R., Zhou, T., and Ma, J. (2022a), "Multiscale and Integrative Single-Cell Hi-C Analysis with Higashi," *Nature Biotechnology*, 40, 254–261. [2464,2467]

Zhang, R., Zhou, T., and Ma, J. (2022b), "Ultrafast and Interpretable Single-Cell 3D Genome Analysis with Fast-Higashi," in *International Conference on Research in Computational Molecular Biology*, pp. 300–301, Springer. [2464,2466]

Zheng, Y., Shen, S., and Keleş, S. (2022), "Normalization and De-noising of Single-Cell Hi-C Data with BandNorm and scVI-3D," *Genome Biology*, 23, 1–34. [2464,2467,2469,2475]

Zhou, J., Ma, J., Chen, Y., Cheng, C., Bao, B., Peng, J., Sejnowski, T. J., Dixon, J. R., and Ecker, J. R. (2019), "Robust single-Cell Hi-C Clustering by Convolution-and Random-Walk–based Imputation," *Proceedings of the National Academy of Sciences*, 116, 14011–14018. [2464,2467]