# UNDERSTANDING THE GENETIC BASIS OF HUMAN DISEASES BY COMPUTATIONALLY MODELING THE LARGE-SCALE GENE REGULATORY NETWORKS

By

Hao Wang

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Computational Mathematics, Science and Engineering — Doctor of Philosophy

2022

## ABSTRACT

UNDERSTANDING THE GENETIC BASIS OF HUMAN DISEASES BY COMPUTATIONALLY MODELING THE LARGE-SCALE GENE REGULATORY NETWORKS

By

Hao Wang

Many severe diseases are known to be caused by the genetic disorder of the human genome, including breast cancer and Alzheimer's disease. Understanding the genetic basis of human diseases plays a vital role in personalized medicine and precision therapy. However, the pervasive spatial correlations between the disease-associated SNPs have hindered the ability of traditional GWAS studies to discover causal SNPs and obscured the underlying mechanisms of disease-associated SNPs. Recently, diverse biological datasets generated by large data consortia provide a unique opportunity to fill the gap between genotypes and phenotypes using biological networks, representing the complex interplay between genes, enhancers, and transcription factors (TF) in the 3D space. The comprehensive delineation of the regulatory landscape calls for highly scalable computational algorithms to reconstruct the 3D chromosome structures and mechanistically predict the enhancer-gene links. In this dissertation, I first developed two algorithms, FLAMINGO and tFLAMINGO, to reconstruct the high-resolution 3D chromosome structures. The algorithmic advancements of FLAMINGO and tFLAMINGO lead to the reconstruction of the 3D chromosome structures in an unprecedented resolution from the highly sparse chromatin contact maps. I further developed two integrative algorithms, ComMUTE and ProTECT, to mechanistically predict the long-range enhancer-gene links by modeling the TF profiles. Based on the extensive

evaluations, these two algorithms demonstrate superior performance in predicting enhancer-gene links and decoding TF regulatory grammars over existing algorithms. The successful application of ComMUTE and ProTECT in 127 cell types not only provide a rich resource of gene regulatory networks but also shed light on the mechanistic understanding of QTLs, disease-associated genetic variants, and high-order chromatin interactions.

## ACKNOWLEDGEMENTS

I want to give my deepest thanks to my advisor, Dr. Jianrong Wang, for his support, guidance, and encouragement during my Ph.D. program. He led me into the field of bioinformatics and guided me to be a successful Ph.D. student, ranging from detailed algorithm design to high-level project design. I am very grateful for his focus on my development, as I can clearly feel the consideration behind every project we worked on together. I will never forget the discussions we had during the late nights and on the way home. It is a privilege to have such an advisor that cares more about my development than himself, and I deeply appreciate it. Beyond the academic development, his enthusiasm, dedication, and meticulous attitude toward work and life profoundly influenced me. I will remember the 'principles of doing things' that he taught me and continuously practice them no matter where I am. No words can comprehend my appreciation, and I feel blessed to learn from him for five years.

I am also grateful to my committee members, Dr. Jianliang Qian, Dr. Yuehua Cui, and Dr. Carlo Piermarocchi, for their insightful guidance during my Ph.D. career. Discussion with them brought me to several brand-new research areas, which significantly expanded my skill sets.

I wanted to thank my colleagues and friends, Jiaxin Yang, Wenjie Qi, Dr. Binbin Huang, Hongjie Ke, Zhongjie Ji, and many others, for their support and accompany.

I wanted to give my deepest thanks to my parents, Haihong Liu and Jianhua Wang for their support and understanding. Without their love and support, I could never achieve so much on my own.

## TABLE OF CONTENTS

## LIST OF FIGURES

# CHAPTER 1

# INTRODUCTION

Genetic disorders have been proved to be closely related to the disease risks of all individuals, and the change of a single nucleotide of the human genome may cause several diseases. Therefore, understanding the relationships between genetic variants and diseases plays a central role in proposing individualized clinical therapy. Over the recent 20 years, Genome-wide association studies (GWAS) have been widely applied and predicted millions of disease-associated SNPs. Traditionally, people mainly focused on the SNPs within the coding region of genes and used the genes containing the SNPs as mediators to explain the SNP-disease association. However, genes only take 2% of the human genome, and the mechanisms of SNPs within the non-coding region of the human genome remain unclear.

Recently, the development of the Next Generation Sequencing (NGS) technique has been the driving force in studies of functional genomics and generates large-scale genome-wide coverage epigenomic datasets measuring gene expression, chromatin opening, and transcription factor (TF) binding sites across diverse cell types. Taking advantage of big biological data, over a million enhancers were discovered within the non-coding regions, which can be bound by transcription factors and regulate the expression of both local and distal genes through 3D chromatin loops. The complex interplay between genes, enhancers, and TFs are summarized in the gene regulatory network, where nodes represent the biological factors and edges represent the regulatory

links. The gene regulatory network provides a clear roadmap of how genetic variants contribute to diseases by disturbing genes, enhancers, and TFs.

Predicting the interactions between genes and enhancers are challenging. Disputing one can assign the nearest gene to enhancers as the target gene in 1D space, it has been proved that the enhancers can be brought to the proximal of the distal target genes in 3D space through long-range chromatin loops and regulate the gene expression. Experimentally, the chromosome conformation capture technique, including Hi-C and Capture-C, has been used to profile the chromatin contact between DNA fragments. However, the experimental data can only predict chromatin interactions in low resolution. Furthermore, the experimental data can only predict short-range chromatin interactions (<500kb) and has low power in predicting long-range chromatin interactions. Therefore, a robust computation model for predicting the long-range interaction between enhancers and genes is in great need. To address this problem, we developed a series of machine learning models to predict and utilize the 3D chromosome structures enhancer-gene interactions: FLAMINGO, tFLAMINGO, ComMUTE, ProTECT, and APRIL. In Chapter 2 and 3, we introduce FLAMINGO and tFLAMINGO, which reconstruct 3D chromosome structures based on Hi-C contact maps. FLAMINGO is a highly scalable and accurate algorithm for predicting high-resolution 3D chromosome structures from Hi-C contact maps. Using FLAMINGO, we successfully reconstructed the 3D structures of all 23 human chromosomes in the highest resolution (1kb) in six cell types. tFLAMINGO further expands the reconstruction of 3D chromosome structures into single cells using low-rank tensor completion. The application of tFLAMINGO in four single-cell chromatin interaction datasets provides a unique opportunity to study the dynamic 3D chromosome structures

across single cells and the relationship with gene regulations. In Chapter 4 and Chapter 5, we introduce ComMuTE and ProTECT, which predict functional regulatory interactions between enhancers and genes. ComMuTE models the joint regulatory effect of multiple enhancers and TFs using a graphical statistical model. We applied ComMuTE in 127 cell types/tissues to predict enhancer-gene links and provided a mechanistic explanation of high-order chromatin interactions and epistasis QTLs. ProTECT predicts the enhancer-gene links by modeling the Protein-Protein Interactions (PPI) between TFs. The elucidation of TF-mediated enhancer-gene links provides new insights into understanding the trans-QTL.

# CHAPTER 2

# RECONSTRUCT HIGH-RESOLUTION 3D GENOME STRUCTURES FOR DIVERSE CELL-TYPES USING FLAMINGO

A modified version of this chapter was previously published (Wang H. et al, 2022): Wang H., Yang J., Zhang Y., Qian, J. and Wang. J. (2022) Reconstruct high-resolution 3D genome structures for diverse cell-types using FLAMINGO. Nature Communications.

## 2.1 INTRODUCTION

The three-dimensional (3D) architecture of genomes plays pivotal roles in DNA replication, genome stability and tissue differentiation[1-3]. Quantitative characterization of spatial chromosome conformations is crucial for deciphering the complex systems of spatially coordinated transcriptional and epigenetic activities[4-6], leading to the understanding of gene regulation mechanisms. The genome-wide high-throughput chromosome conformation capture technique such as Hi-C[7, 8] has been one of the driving forces in studies of 3D genome structures. The Hi-C datasets profiled from different cell-types and species[7, 9-13] have revealed structural components of genome organization[7, 10, 14], such as chromatin loops, topologically associated domains (TADs), and chromatin compartments. Although these findings have provided powerful insights into the governing rules of chromosome folding at large scales (~100kb-1Mb), such as the loop extrusion model[15, 16], it is still computationally difficult to accurately reconstruct high-resolution spatial conformations, such as at ~5kb resolution, for all chromosomes in large genomes.

Since the collection of Hi-C experiments is growing, the resulting massive Hi-C data call for efficient computational algorithms for modeling 3D genomes. Previous algorithms of 3D reconstruction using Hi-C data have been able to predict spatial distances mainly at low-resolutions or within specific genomic segments[17]. Typically, based on experimentally estimated conversion functions[14], the observed Hi-C contact frequency is converted into spatial distances, which we term as observed Hi-C distances in this paper. In general, a consensus structure or an ensemble of structures are inferred by maximizing the similarity between predicted and observed Hi-C distances using optimization-based (such as MDS-type or manifold learning techniques)[18-27] or probabilistic approaches (such as MCMC strategy) [28-32]. Representative state-of-the-art algorithms that have been shown to outperform other methods, along with some recent developments, include ShRec3D[33], GEM-FISH[34], Hierarchical3DGenome[35], RPR[36], SuperRec[37], ShNeigh[38] and PASTIS[28] (Methods, Supplementary Note 1). The accuracy of a predicted structure is mainly evaluated by its capability of recapitulating the measured pairwise distances between genomic loci from Hi-C. Spearman correlation is one of the widely used metrics to quantify the accuracy. However, four fundamental challenges still need addressing in developing an efficient algorithm: (1) High scalability to reconstruct high-resolution spatial configurations for all chromosomes from massive Hi-C datasets; (2) Superior performance to handle large fractions of missing data, which is a common drawback of Hi-C experiments; (3) Capability to make accurate cross cell-type structure predictions, since the vast majority of cell-types lack Hi-C data; and (4) Capability to predict high-resolution structures from low-resolution Hi-C contact maps.

To address the above four challenges, we have developed a low-rank matrix completion based methodology for reconstructing 3D genome structures from Hi-C data. Low-rank matrix completion has been found to be a powerful modeling framework for 3D shape inferences in different scientific fields[39-41]. One of the unique advantages of such a modeling method is that it is able to explicitly leverage the low-rank property of a pairwise-distance matrix (rank≤5 for Euclidean distance matrix, see Methods)[42] in an objective function for optimization, and such a low-rank property has not been explicitly utilized in previous approaches, such as multidimensional scaling based methods. Efficient incorporation of the low-rank constraint into the modeling process allows fast structure reconstruction from just a small subset of Hi-C data, making the algorithm scalable for high-resolution structure predictions for large chromosomes with high fractions of missing data.

Our efforts have led us to create a Fast Low-rAnk Matrix completion algorithm for reconstructINg high-resolution 3D Genome Organizations from Hi-C data, FLAMINGO (https://github.com/wangjr03/FLAMINGO), which has been implemented to generate both 5kb- and 1kb-resolution 3D chromosomal structures for the human genome. Based on extensive performance evaluations using data from both simulated structures and experimental Hi-C datasets from the human genome, the high-resolution chromosome structures generated by FLAMINGO demonstrate substantially improved accuracy, compared with other state-of-the-art methods. The predicted high-resolution spatial distances in 3D space are further justified by orthogonal experiments (such as ChIA-PET[43], Capture-C[44, 45] and SPRITE[46]), providing biological insights into long-range chromatin interactions in gene regulation. Beyond 2D contact maps, the predicted 3D

structures by FLAMINGO can help to identify higher-order multi-way chromatin interactions, interpret potential mechanisms of genetic QTLs, characterize the geometrical patterns of chromatin folding, and facilitate the understandings of structural variations. Moreover, even using only 10% of down-sampled Hi-C contacts, FLAMINGO

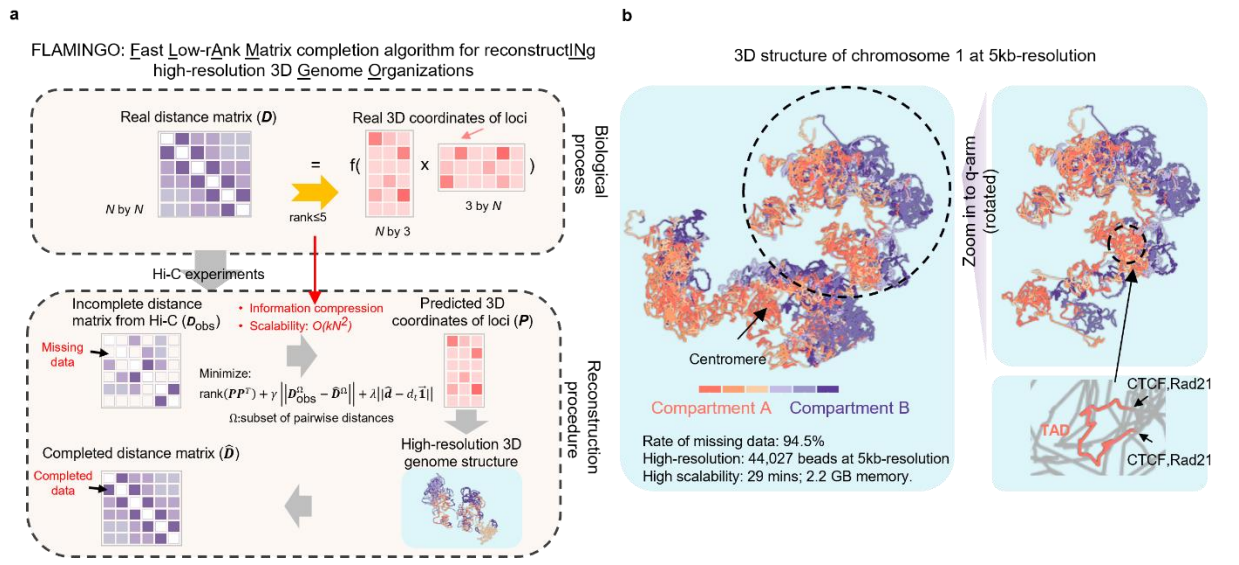



**Figure 2.1 Overview of FLAMINGO.** (**a**) Schematic figure of FLAMINGO. Biologically, the distance matrix (size $N$ by $N$) is induced by the 3D coordinate matrix of DNA fragments (size $N$ by 3), which guarantees that the rank of the distance matrix is no more than five (upper panel). The low-rank property suggests the potential of information compression ($N^2$ entries to $5N$ entries), and enables FLAMINGO to efficiently reconstruct structures from incomplete distance matrices and perform superiorly against large portions of missing data. Equipped with high scalability, FLAMINGO can quickly predict the optimal coordinate matrix that reproduces the observed distances from Hi-C data (middle panel), leading to the high-resolution 3D genome structure and the completed distance matrix (lower panel). (**b**) Reconstructed 5kb-resolution structure of chromosome 1 in the human genome by FLAMINGO. Chromatin compartments (A: orange; B: blue) demonstrate polarized positioning in the predicted structure. A representative example of predicted loop structures is shown in the zoom-in view, where both anchors interact with each other (supported by ChIA-PET interactions) and are bound by CTCF and Rad21. Color gradients represent consecutive TADs within each type of compartments.

still achieves higher accuracy than other methods, demonstrating its superior capability of handling missing data in Hi-C. In addition, an integrative version of our algorithm,

iFLAMINGO, is built to further combine 1D epigenomics data, such as DNase-seq signals, with Hi-C data, which allows us to make cross cell-type predictions of 3D genome architectures and boost the resolution of predictions. These algorithmic advantages will not only expand the coverage of cell-types for 3D genome modeling but also improve the information extraction from the fast-growing collection of experimental Hi-C data.

## 2.2 RESULTS

### 2.2.1 FLAMINGO algorithm to reconstruct high-resolution 3D genome architectures

Based on the 'beads on a string' polymer model[47], every chromosome is modeled as a chain of 'beads' consisting of DNA fragments or loci, and the pairwise distances between genomic loci are biologically induced from the Gram matrix of their 3D coordinates (Figure 2.1.a). To reconstruct the 3D spatial structure, the normalized chromatin contact maps from Hi-C experiments can be converted into an observed distance matrix as suggested by previous studies[10, 14], whose validity and robustness are justified by both computational model selections and empirical comparisons with image-based data (see Methods). The observed distance matrix typically contains large portions of unmeasured distances (namely, missing data), especially for high-resolution genomic loci (~5kb fragments)[10]. FLAMINGO predicts the optimal genome structure based on a low-rank matrix completion framework (Figure 2.1.a). The objective function contains three terms: (1) a term to impose the low-rank constraint on the Gram matrix of predicted 3D coordinates, since the 3D distance matrix has a rank at most five; (2) a term measuring the differences between predicted and observed distances, which is evaluated on the measured subset of pairs of loci; and (3) a penalty term penalizing unrealistic distances between adjacent DNA fragments. FLAMINGO uses the alternating-direction method of multipliers[48] to solve the

optimization problem. At convergence, the optimal 3D structure that minimizes the objective function is identified, along with the completed pairwise distance matrix (Figure 2.1. a).

The key feature of FLAMINGO is to incorporate the low-rank constraint (rank≤5) of the 3D distance matrix of size $N \times N$ into the optimization process, where $N$ is the number of genomic loci, such as the number of 5kb DNA fragments. Since the pairwise spatial distances are generated by the 3D coordinate matrix of genomic loci (rank≤3), the resulting symmetric Euclidean distance matrix has a rank at most 5[42]. It is because the squared Euclidean distance matrix is a sum of three matrices: one being the Gram matrix of rank at most 3 and each of the other two being of rank at most 1 (see Methods). And thus, it has intrinsic degrees of freedom at most $5N$, which, compared to the size $N \times N$ of the entire full matrix, is extremely small when $N$ is large. Therefore, in order to recover the entire distance matrix, we may just need the number of measurements of the distance matrix to be proportional to the intrinsic degrees of freedom. In fact, as long as the information of the underlying distance matrix is not concentrated on a few entries, each randomly selected measurement of pairwise distances will be equally informative, suggesting that the information can be substantially compressed[49] (Figure 2.1.a). Hence, by minimizing the rank of the inferred Gram matrix, low-rank matrix completion models[49] offer at least two benefits (Methods): (1) accurate 3D structures can be reconstructed from subsets of observed distances; and (2) fast matrix calculations can be carried out based on sparsity and low-rankness of the underlying matrices. Remarkably, both benefits are heavily needed for high-resolution structure predictions. By dividing the genome into high-resolution DNA fragments such as at 5kb-resolution, the size of the

distance matrix becomes huge, many entries of which have no data due to the limited sequencing depth of Hi-C experiments. Thus, FLAMINGO is able to build high-resolution 3D structures from the fast-growing collection of Hi-C datasets with decent scalability at computational complexity $O(N^2)$ without demanding increased sequencing depths (Figure A.1 and Figure A.2).

To enable parallel computations, FLAMINGO also employs a hierarchical strategy by dividing each chromosome into 1Mb domain-level fragments that are further divided into 5kb DNA fragments, where we define a 1Mb fragment as a domain (Methods). The same low-rank matrix completion algorithm is applied on both the inter-domain hierarchy consisting of 1Mb fragments, which leads to a basic structural skeleton, and the intra-domain hierarchy of 5kb fragments, which results in intra-domain structures. Different from other methods that only align the endpoints of domain fragments[34] or whose refinement processes are dominated by intra-domain distances[35], an iterative rotation algorithm along the three spatial directions is developed to assemble intra-domain structures into the inter-domain skeleton, by aligning all measured off-diagonal distances so as to maximize the consistency with inter-domain 5kb-resolution Hi-C contacts (Figure A.3, Methods). At convergence, the iterative rotation algorithm leads to the full high-resolution structures for each chromosome.

FLAMINGO has been applied on the normalized Hi-C datasets from six human cell-types (GSE63525[10]) to generate 3D structures for chromosomes 1-22 and X at 5kb-resolution (Figure A.3), which are the largest resources of reconstructed 3D structures for the human genome at high-resolution (https://github.com/wangjr03/FLAMINGO). For example, at 5kb-resolution, chromosome 1 contains 44,027 DNA fragments, excluding the

centromere and telomere regions, and 94.5% entries of the observed distance matrix in GM12878 are missing data. The structure of chromosome 1 can be predicted quickly by FLAMINGO (Figure 2.1.b). The two types of chromatin compartments (A/B) are organized into separable positions in the predicted structure, consistent with the polarized architecture observed from the multiplexed FISH[14]. By zooming into the high-resolution structure, predicted loop structures are found corresponding to previously annotated TADs (Figure 2.1.b), where the pairs of CTCF-associated Hi-C loop anchors (CTCF-CTCF pairs) are predicted with significantly shorter spatial distances, compared to genomic-distance controlled pairs in two cases: 1) pairs between a CTCF-anchor and a random anchor with the same genomic separation ( CTCF-random pairs, Figure A.1

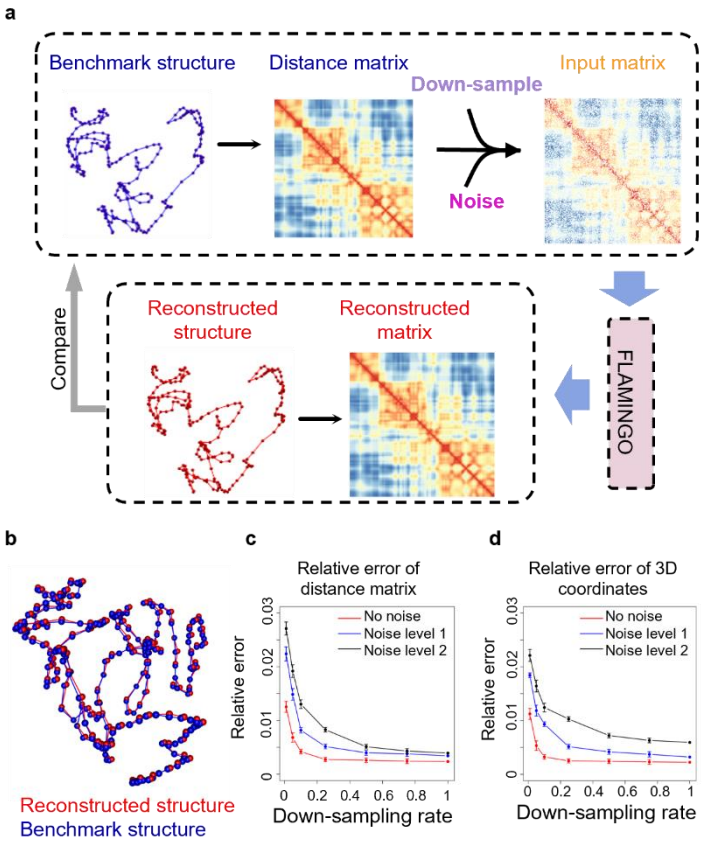



**Figure 2.2 Simulation analyses of FLAMINGO.** (**a**) Given a benchmark structure, the distance matrix is down-sampled using different down-sampling rates and mixed with

**Figure 2.2 (cont'd)**

different levels of noise (Noise level 1: low-level; Noise level 2: high-level; see Methods). The incomplete noisy distance matrices are used as inputs for FLAMINGO. The reconstructed 3D structures are compared with the benchmark structure by calculating relative errors and correlations. (**b**) One example of the reconstructed structure by FLAMINGO (down-sampling rate=0.5, noise level 1, see Methods), which aligns with the benchmark structure almost identically (correlation=0.9999999, relative error=0.0037). (**c-d**) The performance of FLAMINGO (relative errors: the y-axis) under various down-sampling rates and noise levels, with respect to the accuracy of 3D distance matrices (**c**) and 3D coordinates of DNA fragments (**d**). Error bars represent the standard deviations of relative errors examined based on n=10 independently down-sampled distance matrices under each down-sampling rate. Data are presented as mean values +/- SD. Source data are provided as a Source Data file.

boxplot, right, p-value=$5.21 \times 10^{-4}$, one-sided Wilcoxon test), and 2) pairs between random anchors with the same genomic separation (random-random pairs, Supplementary Fig 1 boxplot, left, p-value=$2.78 \times 10^{-5}$, one-sided Wilcoxon test). In addition, FLAMINGO has also generated 3D chromosomal structures at 1kb-resolution in GM12878 for all chromosomes (Figure A.2), which represent spatial reconstructions with the highest resolution to date. Moreover, FLAMINGO is robust to the choice of conversion factors for converting interaction frequency to distance, where the conversion factor is chosen within the range suggested by previous studies[14, 28] (Figure A.4).

### 2.2.2 Benchmark performance based on simulated structures

The performance of FLAMINGO was benchmarked on simulated structures. The distance matrix generated from the benchmark structure was randomly down-sampled and then mixed with noise (Figure 2.2.a, Methods). By applying FLAMINGO on the noisy incomplete distance matrices, the reconstructed 3D structures can be identified with fast convergence (Figure A.5), and they are in strong agreement with the original benchmark structures (relative error<0.03 and correlation>0.999) (Figure 2.2.b). In addition, the

accuracy is robust against a wide range of down-sampling rates and different levels of noise (Figure 2.2.c and 2.2.d, Figure A.5, correlation>0.999), demonstrating that FLAMINGO is capable of handling missing data. The high accuracy is also found to be robust when FLAMINGO is applied to a series of simulated structures with different sizes (Figure A.6), suggesting the performance is not affected by the number of genomic loci along chromosomes. Furthermore, to validate the iterative assembly algorithm for organizing intra-domain structures, we partitioned the benchmark structure into different domains and then reconstructed the whole structure using the assembly algorithm. The assembled structures recapitulate the benchmark structure with high accuracy (relative

error < 0.005, correlation > 0.999) and are independent of specific choices of domain partitions (Figure A.7a and 7b).

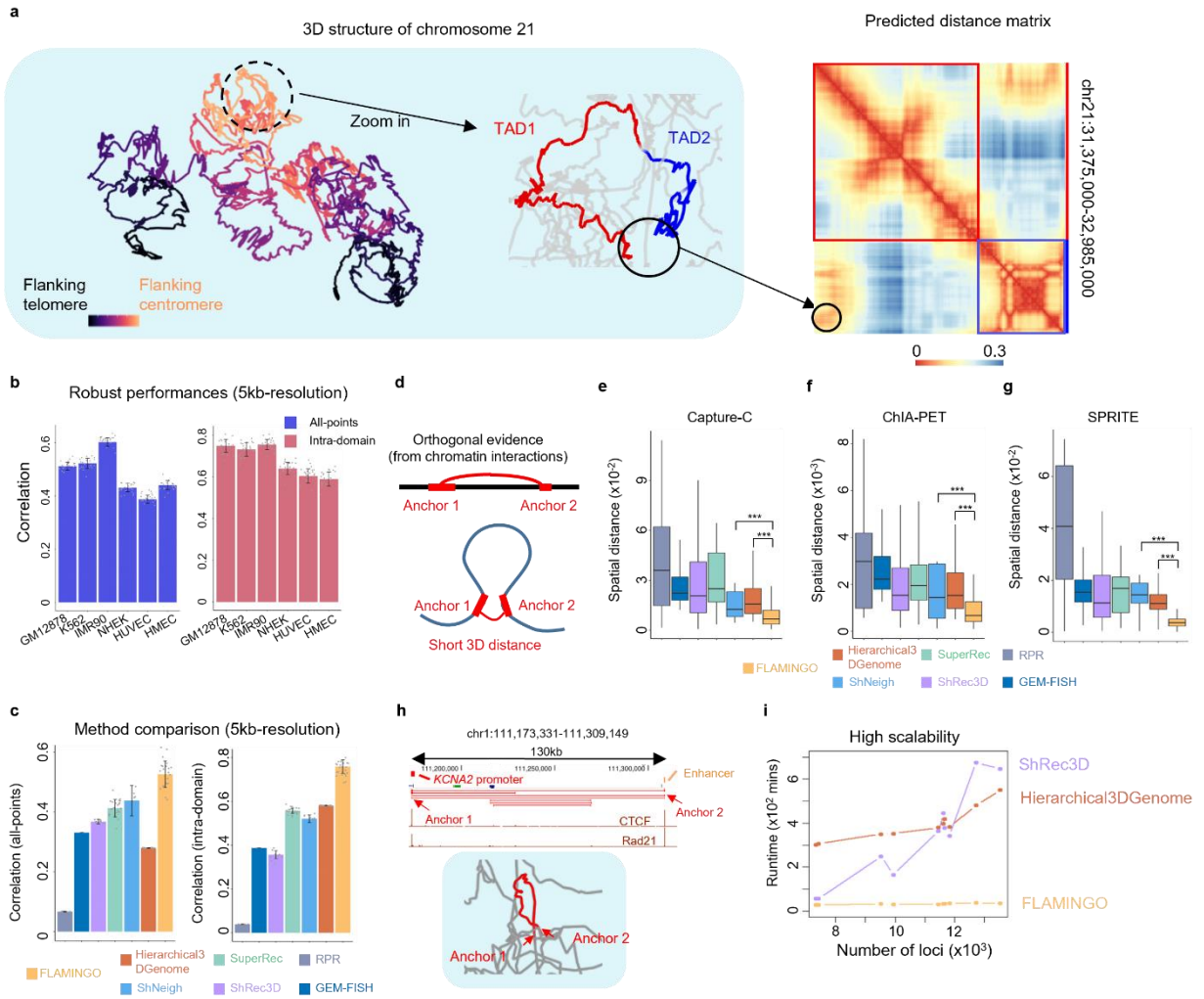



**Figure 2.3 Superior accuracy and scalability of FLAMINGO.** (**a**) The reconstructed structure of chromosome 21 at 5kb-resolution (left). The color gradient represents the genomic distance to the centromere (flanking centromere; yellow; flanking telomere: black). As an example, FLAMINGO recovers the chromatin loop formed by two TADs (chr21:31,375,000-32,985,000; middle), corresponding to inter-TAD hotspots in the reconstructed 3D distance matrix (right). (**b**) Robust performance of FLAMINGO across six cell-types at 5kb-resolution. Correlations between predicted and observed distance matrices are calculated for all 5kb fragments (all-points: blue) and fragments within domains (intra-domain: salmon). Error bars represent the standard deviations across n=23 chromosomes. (**c**) Performance comparison with the state-of-the-art algorithms based on Hi-C data in GM12878 at 5kb-resolution (all-points: left; intra-domain: right). Error bars represent the standard deviations across chromosomes with complete predictions (n=23 for FLAMINGO, Hierarchical3Dgenome and SuperRec; n=10 for ShRec3D; n=9 for ShNeigh; n=6 for RPR). GEM-FISH does not have error bars because it can only complete the prediction for chromosome 21. (**d**) Orthogonal chromatin

**Figure 2.3 (cont'd)**

interaction data provides additional evaluation metrics: anchors of chromatin interactions are expected to have short 3D distances. (**e-g**) FLAMINGO predicts significantly shorter distances between anchors of chromatin interactions profiled by Capture-C (n=3,692) (**e**), ChIA-PET (n=214) (**f**) and SPRITE (n=871) (**g**). The statistical significance (***) is calculated by one-sided Mann-Whitney test: (**e**) p-value=$9.4x10^{-25}$ (orange) and p-value=$7.6x10^{-24}$ (blue); (**f**) p-value=$2.8x10^{-22}$ (orange) and p-value=$5.1x10^{-20}$ (blue); (**g**) p-value=$7.4x10^{-31}$ (orange) and p-value=$6.5x10^{-42}$ (blue). The 3D structures of different methods are normalized for fair comparison. The center lines of boxplots show the median, the upper and lower box limits show the $25^{th}$ and $75^{th}$ percentiles respectively. The whiskers extend up to 1.5 times the interquartile range away from the limits of the boxes. Outliers outside this range were removed from the figure. (**h**) One example of chromatin loops predicted by FLAMINGO for a significant ChIA-PET interaction (red links) linking the KCNA2 promoter (red) with a distal enhancer (orange). (**i**) Comparison of the computational scalability by measuring the runtime (y-axis) as a function of different numbers of genomic loci (x-axis). Source data are provided as a Source Data file.

### 2.2.3 Superior reconstruction accuracy across diverse cell-types

The performance of FLAMINGO on experimental Hi-C data in the human genome was then systematically evaluated and compared with the state-of-the-art methods. As demonstrated in Figure 2.1.b and Figure A.1, FLAMINGO is able to quickly reconstruct 3D chromosome structures at 5kb-resolution, which are qualitatively consistent with both large-scale chromatin properties, such as compartments and TADs, and small-scale structural details, such as chromatin loops and CTCF/cohesin bindings. The predicted structural skeletons (1Mb-resolution) of chromosomes are strongly supported by results from both Hi-C (average correlation=0.95, Figure A.8.a) and FISH[14] (average correlation=0.80, Figure A.8.b), consistently higher than other methods. The reconstructed structures also vary across different cell-types, consistent with cell-type specific chromatin contact patterns from Hi-C (Figure A.9). Taking the predicted structure of chromosome 21 in GM12878 as an example, FLAMINGO reconstructs clear loop structures for TADs and predicts short 3D distances for inter-TAD chromatin contacts

(Figure 2.3.a). Compared to the fuzzy input distance matrix converted from Hi-C (Figure A.10), the distance matrix derived from the predicted 3D structure shows substantially improved resolution (Figure 2.3.a), and the reconstructed long-range inter-TAD contacts are supported by experimental Capture-C interactions (Figure A.10).

To quantitatively evaluate the genome-wide accuracy at 5kb-resolution, the predicted 3D chromosome structures were evaluated according to their consistency with the observed distance matrices derived from Hi-C (Methods). Similarities between structures are quantified by Spearman correlations, which have been widely used as accuracy metrics in structure analysis. To note, achieving high correlations at 5kb-resolution is a much harder problem than at low-resolutions (e.g. 100kb- or 1Mb-resolution), because Hi-C signals at 5kb-bins are much noisier and the number of high-resolution constraints in optimization is huge. Remarkably, the Spearman correlations between the predicted and observed 3D distances at 5kb-resolution, including both diagonal sub-matrices for intra-domain structures and off-diagonal sub-matrices for inter-domain structures, are robustly high across all six cell-types (Figure 2.3.b, left). The predicted structure in IMR90 shows the highest correlation (average correlation=0.603 across 23 chromosomes), followed by structures predicted in GM12878 and K562 (average correlations=0.512 and 0.525 respectively). The Spearman correlations based on off-diagonal points alone (i.e. inter-domain distances) also show similar levels (correlations>0.42), except for HUVEC (correlation=0.32). These results are significant achievements, considering the extensive noisy constraints imposed by the huge number of pairwise distances at 5kb-resolution. For example, in chromosome 1, there are $6.7 \times 10^7$ pairs of 5kb fragments with measured Hi-C contacts as constraints. Furthermore, the predicted intra-domain structures

demonstrate higher correlations across the six cell-types (Figure 2.3.b, right), especially in GM12878, K562 and IMR90 (average correlation>0.73). In addition, even at 1kb-resolution, the reconstructed 3D structures achieve high correlations with the observed spatial distances for both whole chromosomal structures and intra-domain structures (Figure A.2, all-points correlations ~0.4 and intra-domain correlations ~0.6). These consistently high correlations indicate that FLAMINGO is able to capture both long-range genome folding patterns and detailed structures within domains.

FLAMINGO was then compared with other methods, GEM-FISH[34], ShRec3D[33], Hierarchical3DGenome[35], ShNeigh[38], RPR[36] and SuperRec[37], which are state-of-the-art and recently developed algorithms representing different modeling strategies (Methods, Supplementary Note 1). Strikingly, FLAMINGO achieved substantially higher correlations than all the other methods, for both whole chromosome structures and intra-domain structures, at 5kb-resolution (Figure 2.3.c, Figure A.11 and A.12). For example, FLAMINGO achieved a correlation of 0.53 for whole chromosome structures in GM12878, while the other methods only achieved correlations below 0.45 (Figure 2.3.c, left). Similar advantage of FLAMINGO is also observed when the performance comparison is restricted to off-diagonal long-range inter-domain distances (Figure A.11). Moreover, focusing on detailed intra-domain structures, FLAMINGO achieved a correlation of 0.76 in GM12878, while the other methods only achieved correlations below 0.6 (Figure 2.3.c, right). Similarly, FLAMINGO outperformed across all the other five cell-types at 5kb-resolution (Figure A.12).

To further leverage orthogonal data for performance comparisons, high-resolution chromatin interactions profiled by Capture-C[45], ChIA-PET[50] and SPRITE[46] experiments

were used to evaluate whether the reconstructed structures assign short 3D distances between interacting anchors (Figure 2.3.d, Methods). Remarkably, FLAMINGO consistently demonstrated higher accuracy than other methods across all three sets of experimental metrics (Figure 2.3.e-g). The reconstructed structures from FLAMINGO assign statistically significant shorter 3D distances between anchors of chromatin interactions ($p - value < 2 \times 10^{-16}$, one-sided Mann-Whitney test), while other methods are less likely to capture the structural proximity for chromatin interactions. As an example (Figure 2.3.h), a long-range ChIA-PET interaction (~130kb) on chromosome 1 links a distal enhancer element (the anchor 2) to the promoter region of gene KCNA2 (the anchor 1), where both anchors are bound by CTCF and Rad21. Interestingly, in the reconstructed high-resolution structure by FLAMINGO, the enhancer and the promoter are in close proximity with each other and the genomic region in between forms a smooth chromatin loop. As comparison, the Hierarchical3DGenome algorithm does not assign a short spatial distance between the interacting enhancer and the KCNA2 promoter (Figure A.13.a). Additional examples can be found in Figure A.13 b-c. These results not only provide rigorous evidence to validate the superior accuracy, but they also underscore the impacts of FLAMINGO on decoding the mechanisms underlying orchestrated gene regulation in 3D space.

### 2.2.4 Advanced scalability for large-scale chromosome conformations

High-resolution 3D structure modeling places stringent demands for performance, reliability, and more importantly, scalability on algorithms, since a large number of genomic loci and pairwise distances are used in the optimization procedure. Based on efficient information compression and matrix computation, the computational complexity

of FLAMINGO is $O(kN^2)$, where $N$ is the number of genomic loci, such as the number of 5kb DNA fragments, and $k$ is a small constant. For example, it only took 42 minutes and 2.2GB memory for FLAMINGO to reconstruct the 5kb-resolution 3D structure for chromosome 1, the largest chromosome in the human. For chromosomes 2-22 and chromosome X, FLAMINGO was able to predict their structures even faster (Figure A.14a). As comparison, the state-of-the-art algorithms all have inferior scalability. The running times for Hierarchical3Dgenome and ShRec3D increase rapidly when the number of genomic loci becomes large (Figure 2.3.i), while the other methods (i.e. SuperRec, ShNeigh, RPR and GEM-FISH) are even slower (Figure A.14b). Most of these methods can only make predictions for short chromosomes (e.g. chr12-22) at 5kb-resolution. Furthermore, because FLAMINGO can accurately predict the 3D structures based on a small subset of pairwise distances, the scalability of FLAMINGO can be improved further by down-sampling the distance matrix from Hi-C (Figure A.14.c). In addition, based on our tests of 1kb-resolution reconstruction for all chromosomes in GM12878 (Figure A.2), FLAMINGO can generate complete predictions for large chromosomes fast. For the largest chromosome (chr1), it takes less than 25 hours using 200GB memory to reconstruct the 1kb-resolution 3D structure. Therefore, FLAMINGO provides drastic improvements on the computational scalability, which is much desired since a large number of Hi-C datasets are to be generated in the near future[51, 52].

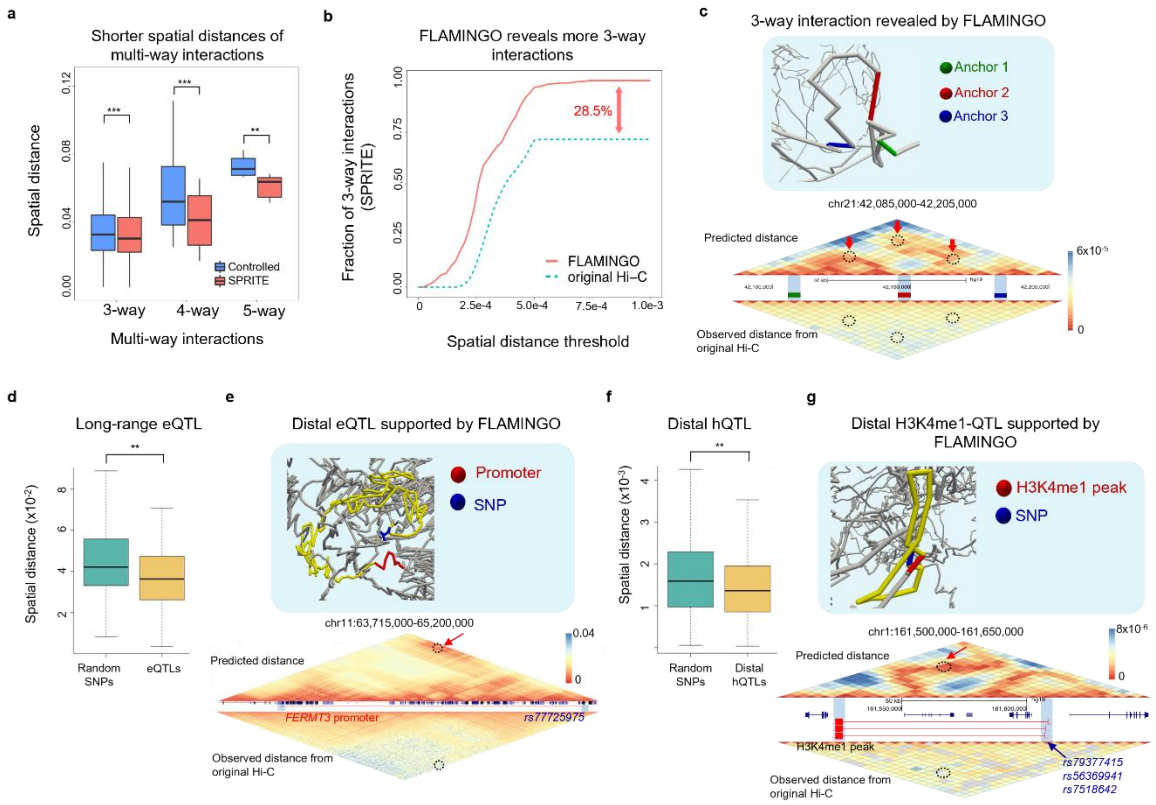



**Figure 2.4 Interpretation of multi-way chromatin interactions and QTLs.** (**a**) SPRITE multi-way interactions on chr21 are predicted with shorter spatial distances than the genomic-distance controlled background (**: p-value < $10^{-2}$, ***: p-value < $10^{-3}$; one-sided Wilcoxon test). The x-axis corresponds to 3-way (p-value=$2.3x10^{-9}$, n=302), 4-way (p-value=$1.9x10^{-4}$, n=17), and 5-way interactions (p-value=$7.8x10^{-3}$, n=7). The center lines of boxplots show the median, the upper and lower box limits show the 25$^{th}$ and 75$^{th}$ percentiles respectively. The whiskers extend up to 1.5 times the interquartile range away from the limits of the boxes. Outliers outside this range were removed from the figure. (**b**) FLAMINGO captures more 3-way interactions across different distance thresholds, compared to using normalized Hi-C contact map derived distance matrix. (**c**) One example of a SPRITE 3-way chromatin interaction captured by FLAMINGO. (**d**) The SNP-promoter pairs of long-range eQTLs (>900kb) are assigned with significantly shorter spatial distances by FLAMINGO, compared to genomic-distance controlled random pairs (**: p-value=$4.1x10^{-3}$; one-sided Wilcoxon test, n=1,227). The center lines of boxplots show the median, the upper and lower box limits show the 25$^{th}$ and 75$^{th}$ percentiles respectively. The whiskers extend up to 1.5 times the interquartile range away from the limits of the boxes. Outliers outside this range were removed from the figure. (**e**) One example of long-range eQTLs interpreted by FLAMINGO. The SNP rs77725975 (blue) and the promoter of FERMT3 (red) are placed in close 3D proximity. (**f**) The SNP-H3K4me1 pairs of distal hQTLs are assigned with significantly shorter spatial distances by FLAMINGO, compared to genomic-distance controlled random pairs (**: p-value=$6.3x10^{-3}$; one-sided Wilcoxon test, n=20,950). The center lines of boxplots show the median, the upper and lower box limits show the 25$^{th}$ and 75$^{th}$ percentiles respectively.

**Figure 2.4 (cont'd)**

The whiskers extend up to 1.5 times the interquartile range away from the limits of the boxes. Outliers outside this range were removed from the figure. (**g**) One example of distal H3K4me1-QTLs interpreted by FLAMINGO. The SNPs rs79377415, rs56369941, rs7518642 (blue) and the H3K4me1 ChIP-seq peak (red) are placed in close 3D proximity. Source data are provided as a Source Data file.

### 2.2.5 Analysis of multi-way interactions and QTLs by FLAMINGO beyond 2D Hi-C contact maps

To demonstrate the biological discoveries enabled by FLAMINGO that are not directly visible from 2D contact maps, the reconstructed 3D chromatin structures are used to resolve two important questions. First, we analyzed the predicted 3D structure's capability of capturing multi-way chromatin interactions. Spatially coordinated molecular processes frequently form multi-way interactions (e.g. 3-way, 4-way or 5-way interactions) in 3D space[46, 53, 54], which play pivotal roles in coupled transcriptional and epigenetic activities[55]. However, Hi-C contact maps can only reveal pairwise 2-way chromatin interactions. Moreover, the high rates of missing data in Hi-C result in large genomic regions with almost no measured interactions, further limiting the capability of finding multi-way interactions from 2D contact maps. Since FLAMINGO recovers the whole spatial structure, we hypothesize that the predicted 3D structures can improve the identification of multi-way interactions. The multi-way chromatin interactions profiled by SPRITE experiments in GM12878[46] are used to justify this hypothesis. In addition to pairwise interacting anchors (Figure 2.3.g), the GM12878 structure predicted by FLAMINGO consistently assigns significantly shorter spatial distances among anchors of multi-way interactions in SPRITE (Figure 2.4.a, p-value$<10^{-2}$, one-sided Wilcoxon test), compared to genomic-distance controlled random samples, suggesting the predicted 3D structures are in strong

agreement with the higher-order organizations of multi-way interactions. More importantly, compared to using the Hi-C contact map derived distance matrix, the predicted 3D structure by FLAMINGO can capture more multi-way interactions (Figure 2.4.b, Figure A.15.a-b). Here, a multi-way interaction is considered to be captured if all interacting anchors are located in the same 3D spatial neighborhood, where all pairwise spatial distances between anchors are smaller than a specified threshold. As shown in Figure 4b, across a wide range of thresholds on normalized spatial distances, FLAMINGO consistently demonstrates higher capabilities of discovering more 3-way interactions. Even if relaxed distance thresholds are used, 28.5% 3-way interactions from SPRITE experiments can not be identified based on Hi-C contact map derived distance matrix, while being captured by FLAMINGO (Figure 2.4.b). It is because these 3-way interactions involve distal interacting anchors across very long-range genomic regions (median genomic distance=2.32Mb), where Hi-C contact maps suffer from high rates of missing data. Similar results are also found for 4-way and 5-way interactions (Figure A.15.a and 15.b), where FLAMINGO achieves much higher advantages. Figure 4c shows a representative example of a 3-way interaction that has been identified by SPRITE experiments[46]. The three interacting anchors are brought into spatial proximity based on the predicted loop structures, which are also highlighted in the predicted distance matrix (Figure 2.4.c, right). As comparison, the distance matrix based on the Hi-C contact map shows no signals of spatial closeness for the three anchors. As another interesting example, a candidate 4-way interaction mediated by CTCF across a 12Mb genomic region in chr1 is discovered by FLAMINGO, while the Hi-C based distance matrix shows no spatial patterns (Figure A.15.c). These results suggest that, by reconstructing 3D

spatial structures, FLAMINGO can help to identify multi-way chromatin interactions and reveal higher-order genome organizations, beyond 2D Hi-C contact maps.

Second, we analyzed the predicted 3D structure's utility in interpreting genetic associations, such as long-range expression QTLs (eQTL) and distal histone QTLs (hQTL) in matched cell-types or tissues. QTLs statistically link genetic variants to molecular phenotypes and facilitate understandings of disease genetics. But it has been challenging to delineate the underlying molecular mechanisms of genetic associations. Spatial proximity between genetic variants and target genes or histone modification peaks have been suggested to mediate genetic associations[56, 57]. Similar to the approach of multi-way chromatin interaction analysis, the predicted 3D structure is evaluated with respect to its ability of interpreting QTLs[58-62] based on predicted short spatial distances, compared to using the Hi-C contact map derived distance matrix. Interestingly, across a wide range of thresholds on normalized spatial distances, substantially higher fractions of eQTLs and hQTLs are found to have their genetically associated loci (i.e. SNP-promoter or SNP-histone pairs) placed into small 3D neighborhoods by FLAMINGO (Figure A.16). Focusing on the long-range eQTLs[61] whose SNPs and target gene promoters are >900kb away, these SNP-promoter pairs are found to be assigned with significantly shorter spatial distances, compared to genomic-distance controlled random pairs (p-value=$1.3 \times 10^{-3}$, one-sided Wilcoxon test, Figure 2.4.d), suggesting the effectiveness of FLAMINGO in interpreting genetic associations. For each specific long-range eQTL (>900kb), a random set of SNP-promoter pairs with the same genomic-distance from the same chromosome is generated (Methods). Among these long-range eQTLs (n=1,227), 671 of them (54.7%) are predicted to have spatial distances that are at least 2-fold shorter than the median

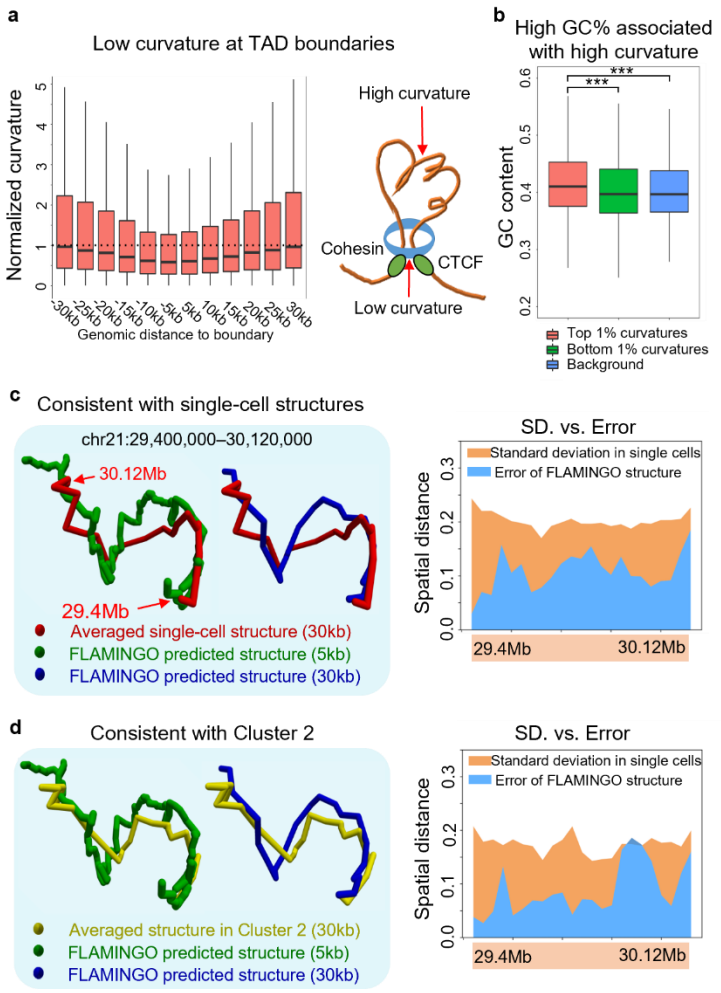



**Figure 2.5 Geometrical signature of predicted chromatin conformations.** (**a**) TAD boundaries demonstrate lower curvatures than flanking genomic regions. The center lines of boxplots (n=11,208) show the median of normalized curvatures, the upper and lower box limits show the 25$^{th}$ and 75$^{th}$ percentiles respectively. The whiskers extend up to 1.5 times the interquartile range away from the limits of the boxes. Outliers outside this range were removed from the figure. (**b**) The regions with high curvatures show higher GC-content compared with genomic background. One-sided Mann-Whitney test (***): p-value=2.7x10$^{-29}$ (green, n=5,261) and p-value=3.4x10$^{-34}$ (blue, n=5,261). The center lines of boxplots (n=5,261) show the median, the upper and lower box limits show the 25$^{th}$ and 75$^{th}$ percentiles respectively. The whiskers extend up to 1.5 times the interquartile range away from the limits of the boxes. Outliers outside this range were removed from the figure. (**c**) The consensus structure predicted by FLAMINGO consistently aligns with the average structure across single cells in K562. Right: the errors between the predicted consensus structure and the average structure (blue) are smaller than the intrinsic standard deviations among single cells (orange). (**d**) The consensus structure predicted by FLAMINGO is in strong agreement with the average structure across the subset of cells in cluster 2. Right: the errors between the predicted consensus structure and the cluster-2 specific average structure (blue) are smaller than the intrinsic standard

**Figure 2.5 (cont'd)**

deviations among single cells in cluster 2 (orange). Source data are provided as a Source Data file.

spatial distances of genomic-distance controlled random pairs. As a representative example (Figure 2.4.e), the SNP rs77725975 is a significant long-range eQTL to the gene FERMT3 (p-value=$2.6 \times 10^{-4}$) in whole blood cells[61] with a genomic distance of 983kb. This eQTL is placed into 3D proximity by FLAMINGO in GM12878, where the SNP rs77725975 and FERMT3's promoter are located spatially close to each other, while the Hi-C based distance matrix fails to provide structural basis to interpret this eQTL. Similarly, distal hQTLs[62] are also found to be assigned with significantly shorter spatial distances by FLAMINGO, compared to genomic-distance controlled random pairs (p-value=$2.84 \times 10^{-3}$, one-sided Wilcoxon test, Figure 2.4.f). Among the distal hQTLs (n=20,950), 11,797 of them (56.3%) are predicted to have spatial distances that are at least 2-fold shorter than the median spatial distances of genomic-distance controlled random pairs. As shown in Figure 4g for a set of distal hQTLs (p-value<$1.8 \times 10^{-4}$), FLAMINGO reconstructs a loop structure which brings the SNPs close to the specific target H3K4me1 peak that is ~75kb away. In contrast, the distance matrix derived from Hi-C contact maps shows no signal of long-range interactions in this region. These results strongly support the FLAMINGO's ability of interpreting the potential mechanisms of distal QTLs by leveraging the reconstructed spatial proximity information, a critical step further to decipher genetic associations with molecular phenotypes.

### 2.2.6 Geometrical property of chromatin structures

To gain additional insights into genome folding, 3D geometrical metrics are needed to describe the complex shapes of chromatin structures, which can not be directly obtained from Hi-C contact maps. The reconstructed 3D structures provide a systematic platform for dissecting geometrical signatures of chromatin organization. To do this, we calculated the curvatures for every 5kb genomic bin along the 3D curves of chromosomes. A larger curvature around a genomic region indicates the chromatin bends more sharply, while a smaller curvature suggests the region is relatively straight. Interestingly, the curvatures around TAD boundaries show significantly lower curvature than flanking genomic regions (Figure 2.5.a, p-value=$2.2 \times 10^{-16}$, one-sided Mann-Whitney test). Considering the loop extrusion model[16], it suggests that, when a loop is established and the extrusion complex stops sliding, the DNA located around the extrusion complex is maintained rigid. In addition, genomic regions with large curvatures show significantly higher GC-contents (Figure 2.5.b), consistent with the increased flexibility of GC-rich DNA sequences[63, 64] that may facilitate intra-TAD interactions.

### 2.2.7 Reference structure to interpret single-cell variabilities

Based on observations of recent single-cell Hi-C and imaging data[65-68], chromatin structure is dynamic and demonstrates variabilities across individual cells. The optimal consensus structure reconstructed from bulk tissue Hi-C by FLAMINGO thus provides a reference of chromatin folding aggregated from a pool of cells, which can be used as a basis to delineate and interpret the ensemble of chromatin configurations[69, 70]. We compared FLAMINGO's predicted consensus structure to the single-cell structures profiled by diffraction-limited 3D imaging[68] to analyze their relationship. The image-based

dataset[68] contains an ensemble of single-cell structures for a specific genomic region in chr21 at 30kb-resolution. The averaged structure is calculated from the ensemble and is then compared with FLAMINGO's prediction. Figure 5c shows the comparison for a loop structure in this region. Both 5kb- and 30kb-resolution predictions from FLAMINGO align well with the averaged structure of single cells (Figure 2.5.c, left). More importantly, the differences between these structures are consistently smaller than the intrinsic standard deviations among single-cells within the ensemble (Figure 2.5.c, right), suggesting that the consensus structure can sufficiently quantify the major patterns of structural configurations. In addition, it suggests that the distance information derived from Hi-C contact frequency is overall consistent with the spatial configurations obtained from imaging techniques. To further analyze the structural variations relative to the consensus structure, the single-cell structures are classified into five different clusters, where individual cells belonging to the same clusters have similar structures. Structural variabilities are observed across distinct single-cell clusters. Interestingly, for the subset of cells in cluster 2, the cluster-specific average structure is highly similar to the predicted consensus structure (Figure 2.5.d, left), with the differences largely smaller than the intrinsic standard deviations among single cells within this cluster (Figure 2.5.d, right), further supporting the biological relevance of the predicted structure. The other four clusters also similarly demonstrate the overall folding patterns, each of which contains specific variations relative to the predicted consensus structure (Figure A.17). Across all five clusters, the consensus structure consistently shows smaller differences to the cluster-specific average structures, than the intrinsic standard deviations of single cells within each cluster (Figure A.17). These results suggest that the predicted consensus

structures by FLAMINGO can facilitate improved interpretation of the structural heterogeneity in ensembles of single-cell structures.

### 2.2.8 Robust performance to handle missing data in Hi-C datasets

Due to limited sequencing depths of typical Hi-C experiments and low mappabilities of certain genomic regions, the observed distance matrices from Hi-C usually contain large

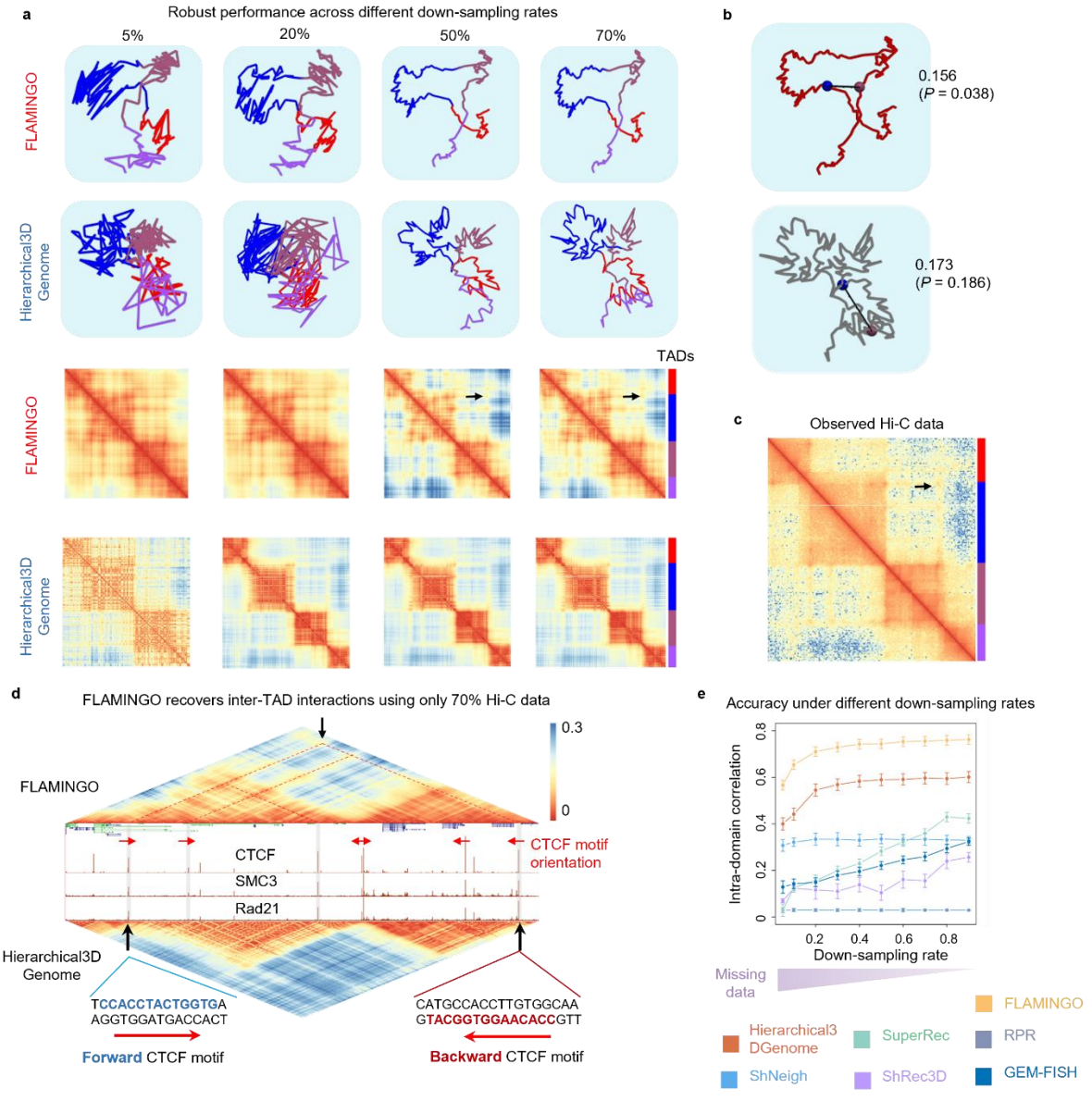



**Figure 2.6 Robust performance of FLAMINGO under different missing rates.** (**a**) Reconstructed 3D structures and completed distance matrices by FLAMINGO and

**Figure 2.6 (cont'd)**

Hierarchical3DGenome in chr21:34,000,000-35,000,000 using down-sampled data. As inputs, the observed distance matrix from Hi-C is down-sampled with different down-sampling rates (columns). Four TADs within this genomic region are annotated by colors. The inter-TAD interaction recovered only by FLAMINGO is highlighted by the black arrow. (**b**) FLAMINGO correctly recovers the short 3D distance between the two distal TAD boundaries (5' of the blue TAD and 3' of the brown TAD) as highlighted in (a), with 70% down-sampled data. After normalization, FLAMINGO predicts a 3D distance of 0.156 (p-value=$3.8 \times 10^{-2}$, n=1,000, permutation test, genomic distance controlled), while Hierarchical3DGenome predicts 0.173 (p-value=0.1862, n=1,000, permutation test, genomic distance controlled). (**c**) The observed distance matrix from Hi-C data, along with TAD annotations and the highlighted inter-TAD interaction. (**d**) The inter-TAD interactions recovered by FLAMINGO (zoom-in view of the blue and brown TADs within chr21:34,100,000-34,850,000) are supported by CTCF and cohesin bindings and the convergent CTCF motifs (red arrows). The inter-TAD interactions are missed by Hierarchical3DGenome. (**e**) FLAMINGO achieves higher reconstruction accuracy against missing data. Correlations between predicted and observed intra-domain structures (the y-axis) are calculated for FLAMINGO and the state-of-the-art methods under different down-sampling rates (the x-axis). The dots show the average correlations based on n=10 independently down-sampled input matrices and error bars correspond to the standard deviations across the ten random samples. Smaller down-sampling rates represent larger fractions of missing data. Source data are provided as a Source Data file.

portions of missing data[10, 71], which present a very challenging problem for high-resolution modeling. For instance, considering the same Hi-C dataset for chromosome 1, the rate of missing data is 21% at 100kb-resolution but quickly increases to 94.5% at 5kb-resolution. Overall, the rate of missing data is >80% across chromosomes 1-22 and X in the human genome at 5kb-resolution (Figure A.14.a). By incorporating the low-rank property of the distance matrix into the optimization procedure, FLAMINGO has the superior advantage of handling high rates of missing data.

To demonstrate FLAMINGO's capability of handling missing data, the observed distances derived from Hi-C were further down-sampled to check whether FLAMINGO still can reproduce the same high-resolution structures (Methods). As a representative example

on chromosome 21 (chr21:34,000,000-35,000,000), FLAMINGO was able to robustly reconstruct the structure even if 50% of the observed pairwise distances from Hi-C was further down-sampled (Figure 2.6.a). By further down-sampling the dataset to the levels with only 20% and 5% of observed data remaining, FLAMINGO was still able to infer the loop structures formed by the four TADs in this region, with slightly increased intra-TAD fluctuations. In contrast, Hierarchical3DGenome predicted fuzzy structures with substantial fluctuations across all down-sampling rates. In addition, specific intra-TAD chromatin contacts were also captured by FLAMINGO, as shown by the specific hotspots within the TAD blocks in the predicted distance matrices at 50% and 70% of down-sampling rates (Figure 2.6.a), while Hierarchical3DGenome only generated vague distance matrices without detailed structures within TAD blocks. More interestingly, FLAMINGO was also able to predict the short 3D distance for long-range inter-TAD contacts in the loop structure using only 70% of observed data (p-value=0.038, permutation test, genomic distance controlled) (Figure 2.6.b), while Hierarchical3DGenome predicted a much longer distance (p-value=0.186). The predicted inter-TAD distance is in agreement with the original Hi-C distance matrix (Figure 2.6.c) and demonstrates a higher level of specificity, although it was inferred from down-sampled data. As additional justifications of the predicted structure with missing data (down-sampling rate = 70% or 50%), the specific intra- and inter-TAD chromatin contacts recovered by FLAMINGO, but not predicted by Hierarchical3DGenome, are supported by CTCF and cohesin bindings, along with convergent pairs of CTCF motifs (Figure 2.6.d, Figure A.18.a).

As global quantitative evaluations, the recovered 3D structures and predicted distances by FLAMINGO using different down-sampled input matrices are compared with the originally observed distances. Strikingly, for the whole 5kb-resolution distance matrix including both inter- and intra-domain structures, the correlation coefficients remain stable and high (~0.49), until less than 30% of observed distances from Hi-C are kept for predictions (Figure A.18.b). Focusing on detailed intra-domain structures, the correlation coefficients still remain to be robustly high (>0.74), until less than 50% of observed distances are kept (Figure 2.6.e). Across the wide range of down-sampling rates, FLAMINGO robustly achieves higher accuracy than other algorithms, based on comparisons using observed Hi-C contact maps (Figure 2.6.e, Figure A.18.b) and also other chromatin interaction datasets, such as Capture-C, ChIA-PET and SPRITE (Figure A.18.c-e). For example (Figure 2.6.e), using only 10% of observed data, FLAMINGO achieved better accuracy than the state-of-the-art method, Hierarchical3DGenome, which used all of the observed data (Figure 2.3.c). These results clearly demonstrate FLAMINGO's ability to accurately reproduce high-resolution structures based on Hi-C with large fractions of missing data, which will significantly relax the demand of sequencing depths in Hi-C experiments and thus promote wide implementations of Hi-C in practice.

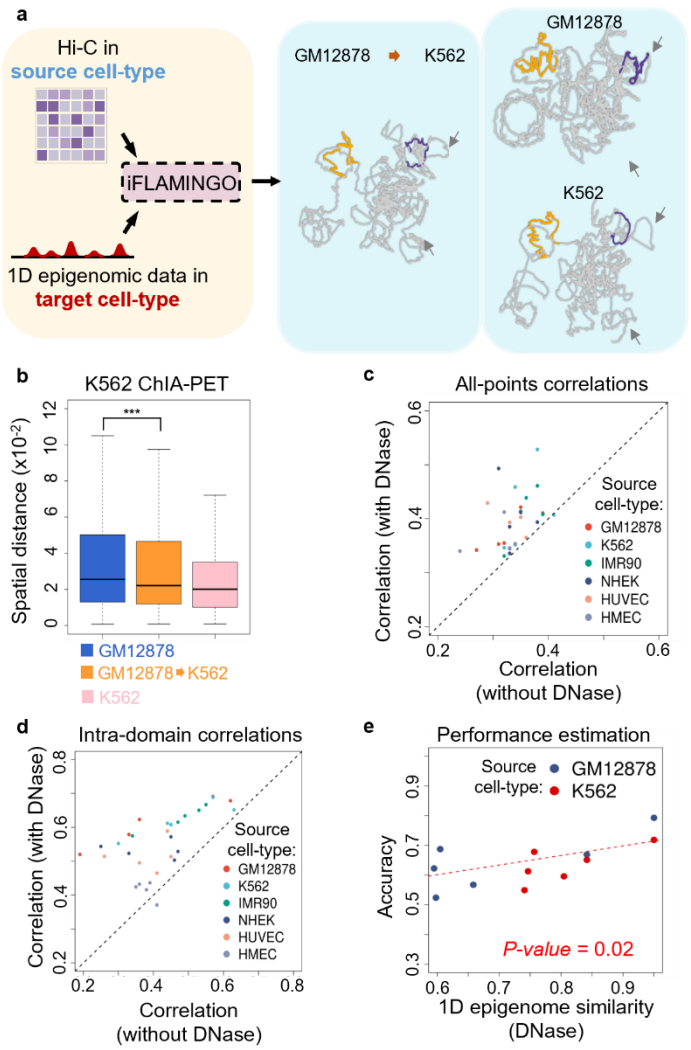



**Figure 2.7 Cross cell-type predictions by iFLAMINGO.** (**a**) Hi-C data from the source cell-type and 1D epigenomics data from the target cell-type are integrated by iFLAMINGO to predict the 3D genome structure in the target cell-type (left). An example of the 3D structure of chromosome 21 for K562 predicted from GM12878 is shown (GM12878->K562). K562-specific structural properties are highlighted by arrows where iFLAMINGO correctly captures, while the GM12878-specific structure shows substantial differences. Two intra-domain structures are further highlighted in the three structures (orange and purple). (**b**) Comparison of 3D distances between interacting ChIA-PET anchors based on the predicted 3D structures of GM12878 (blue), GM12878->K562 (orange) and K562 (pink). P-value=3.0x10$^{-4}$ (n=1,562, one-sided Mann-Whitney test). The center lines of boxplots show the median, the upper and lower box limits show the 25$^{th}$ and 75$^{th}$ percentiles respectively. The whiskers extend up to 1.5 times the interquartile range away from the limits of the boxes. Outliers outside this range were removed from the figure. (**c-d**) Performance comparisons between iFLAMINGO (the y-axis) and FLAMINGO (the x-axis) on cross cell-type predictions for n=30 source-target pairs. Source-target pairs are colored by source cell-types. The performance is quantified by correlations between predicted and observed distances for all DNA fragments, i.e. all-

**Figure 2.7 (cont'd)**

points, in (**c**) and fragments within the same domains, i.e. intra-domain, in (**d**). (**e**) Performance estimation (correlation of 3D distances, y-axis) for cross cell-type predictions of intra-domain structures as a function of 1D epigenomic similarities between cell-types (correlations of genome-wide DNase-seq data, x-axis). The regression line is fitted based on cross cell-type predictions from GM12878 and K562 (p-value=0.02, n=12, two-sided Student's t-test). Source data are provided as a Source Data file.

### 2.2.9 Cross cell-type prediction of 3D structures

Currently experimental Hi-C data have been collected only for a limited number of cell-types, due to the cost of the experiments or the difficulty of collecting sufficient numbers of cells for certain cell-types[71]. To enlarge the coverage of cell-types for 3D genome modeling, FLAMINGO is further extended to iFLAMINGO, an integrative version of the algorithm that can make cross cell-type predictions. To predict the 3D structure for a cell-type without Hi-C data, defined as target cell-type, iFLAMINGO combines two pieces of information (Figure 2.7.a, Methods): (1) Hi-C data from another cell-type, defined as source cell-type, which provides the overall structural backbone of the genome; and (2) chromatin accessibility data, such as DNase-seq, from the target cell-type, which provides the cell-type specific 1D epigenomic landscape. DNase-seq data are widely available across a large panel of cell-types and can characterize chromatin accessibilities at base pair resolution[6]. Since the levels of DNase-seq signals of a pair of genomic loci are associated with their 3D distances, for instance, co-accessible loci being significantly closer to each other in 3D space (Figure A.19.a), a regression model is built to impute approximate 3D distances based on DNase-seq signals in the target cell-type (Figure A.19.b). The imputed cell-type specific distances are then incorporated into iFLAMINGO to predict the 3D genome structure in the target cell-type (Methods).

iFLAMINGO was applied on the Hi-C data from GM12878 to predict the 3D genome structure in K562 by integrating K562-specific DNase-seq data into the modeling process. The resulting structure of chromosome 21 is shown in Figure 7a (GM12878->K562). The 3D structure predicted based on GM12878 Hi-C alone is shown as the negative control, and the structure predicted directly from K562 Hi-C is included as the positive control (Figure 2.7.a). The GM12878->K562 structure not only captures the global structural signatures of the K562 genome but also reconstructs detailed loop structures more similar to K562, both of which are highlighted in Figure 7a. By comparing with K562-specific chromatin interactions profiled by independent ChIA-PET experiments[72], the predicted 3D distances between interaction anchors from the GM12878->K562 structure are significantly shorter than the distances from the GM12878 structure (Figure 2.7.b, p-value=0.0003, one-sided Mann-Whitney test), suggesting the quantitatively improved similarity between the GM12878->K562 and K562 structures. Furthermore, the predicted spatial distances in the GM12878->K562 structure achieve a higher correlation with the experimentally-derived spatial distances of K562 Hi-C (correlation=0.62, Figure A.19.c), compared to the correlation achieved by the basic experimentally-derived spatial distances of GM12878 Hi-C with the experimentally-derived spatial distances of K562 Hi-C (correlation=0.55), suggesting that the predicted GM12878->K562 structure by iFLAMINGO captures the cell-type specificity of K562.

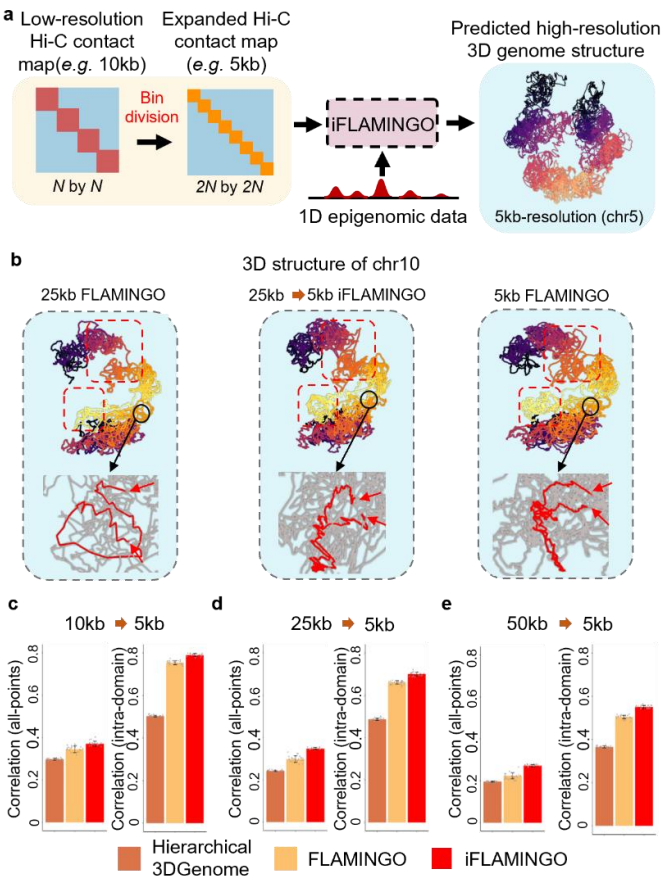



**Figure 2.8 iFLAMINGO improves the resolution of predicted 3D structures.** (**a**) Scheme of the high-resolution 3D structure prediction. Low-resolution distance matrix from Hi-C for $N$ large DNA fragments of size 10kb, are divided into smaller DNA fragments of size 5kb, resulting in a 2$N$ by 2$N$ distance matrix, where the small DNA fragments inherit the same distances to other fragments from the original large fragment. The high-resolution 1D epigenomics signals in each small DNA fragment are integrated into iFLAMINGO to predict the high-resolution 3D genome structures. As one example, the 3D structure of chromosome 5 at 5kb-resolution predicted from the 10kb-resolution distance matrix is shown. (**b**) Example of the predicted 5kb-resolution 3D structure of chromosome 10 from 25kb-resolution distance matrix (middle, 25kb->5kb), compared with the 25kb-resolution structure (left) and the 5kb-resolution structure (right). The large-scale structural differences are highlighted by red boxes. The comparisons of detailed intra-domain structures (red) are shown in inset. The red arrows represent the boundaries. (**c-e**) Performance comparison of predicting 5kb-resolution structures from 10kb-resolution (**c**), 25kb-resolution (**d**), and 50kb-resolution distance matrices (**e**). Correlations between predicted and observed 5kb-resolution distances are calculated for all DNA fragments, i.e. all-points, and for fragments within the same domains, i.e. intra-domain. The bar plot shows the average correlations across n=23 chromosomes and the error bars show the standard deviations across 23 chromosomes. Data are presented as mean values +/-SD. Source data are provided as a Source Data file.

iFLAMINGO was further applied on all source-target pairs from the six cell-types with Hi-C data, and the performance was evaluated based on the correlations between predicted and observed distance matrices in target cell-types (Methods). As comparison, the optimal structures predicted by FLAMINGO without using DNase-seq data are included as negative controls. Among all the 30 source-target cell-type pairs, iFLAMINGO achieved a higher accuracy for almost all the cross cell-type predictions (Figure A.20), not only for the whole distance matrices (Figure 2.7.c) but also for intra-domain structures (Figure 2.7.d). These consistent improvements underscore iFLAMINGO's ability of cross cell-type structure predictions and highlight the importance of 1D epigenomic information in 3D genome modeling.

To further demonstrate iFLAMINGO's potential on enlarging the cell-type coverage for 3D structure reconstructions, the accuracy of cross cell-type 3D predictions is plotted as a function of 1D epigenomic similarities between the source and target cell-types (Figure 2.7.e). Using GM12878 or K562 as source cell-types, the accuracy of predicted intra-domain 3D structures in target cell-types is significantly associated with the 1D epigenomic correlations to the source cell-types (p-value=0.02). Based on the fitted linear function, to obtain a cross cell-type prediction with accuracy>0.6, which is a level already higher than the state-of-the-art methods using Hi-C directly from the target cell-types (Figure 2.3.c), iFLAMINGO only requires Hi-C data available from a source cell-type with medium 1D epigenomic similarities (correlation>0.65). Combined with the ongoing experimental efforts of chromatin characterizations, such as the 4D Nucleome Consortium[51], iFLAMINGO will substantially expand the catalog of cell-types with high-resolution 3D structures.

### 2.2.10 Boost the resolution of 3D structures from low-resolution Hi-C

Since another limiting factor of experimental Hi-C data is the resolution of contact maps being low[73, 74], a tradeoff of genome-wide coverage of sequencing reads, it is much desired to predict high-resolution 3D structures from low-resolution contact maps of Hi-C. By incorporating high-resolution 1D epigenomic data, such as DNase-seq, iFLAMINGO is able to boost the resolution of the predicted 3D genome structures (Figure 2.8.a, Figure A.19). After splitting low-resolution DNA fragments into high-resolution bins, DNase-seq signals help delineate the distance ambiguity across consecutive bins and fine-tune the structures through optimization (Methods).

As a representative example, FLAMINGO was applied to the 25kb-resolution distance matrix for chromosome 10, resulting in a 25kb-resolution 3D structure (Figure 2.8.b, left). On the other hand, based on the 5kb-resolution distance matrix, the 5kb-resolution structure was generated by FLAMINGO as the benchmark structure (Figure 2.8.b, right). Finally, applying iFLAMINGO on the 25kb-resolution distance matrix, along with the DNase-seq data, led to a 5kb-resolution structure, the 25kb->5kb structure (Figure 2.8.b, middle), which shows increased similarity to the 5kb-resolution benchmark structure. The 25kb->5kb structure not only captures large-scale structural properties but also recovers the detailed high-resolution loops in the 5kb-resolution structure, which are missing in the 25kb-resolution structure (Figure 2.8.b).

To quantitatively evaluate the accuracy of boosted resolution genome-wide, a series of low-resolution distance matrices, at 10kb, 25kb and 50kb resolution, respectively, were generated from the same Hi-C datasets. The reconstructed structures were then compared with the original 5kb-resolution distance matrix. Across all the tests,

iFLAMINGO achieved the highest correlations to the benchmark structures (Figure 2.8.c-e). For instance, using 10kb-resolution distance matrices as inputs, iFLAMINGO achieved an average correlation of 0.37 for the whole reconstructed 5kb-resolution matrices and an average correlation of 0.79 for intra-domain matrices, both of which are higher than the state-of-the-art methods even when they were directly applied on 5kb-resolution input matrices (Figure 2.3.c). Therefore, iFLAMINGO not only substantially improves the information extraction from low-resolution Hi-C data but will also widely facilitate the implementation of Hi-C protocols without stringent constraints on resolution.

## 2.3 DISCUSSION

In this study, we have developed an algorithm, FLAMINGO, to reconstruct high-resolution spatial conformations for large genomes in 3D space. Using low-rank matrix completion techniques, FLAMINGO is able to substantially improve data mining efficiency for Hi-C experiments. Based on a series of rigorous performance evaluations, FLAMINGO consistently demonstrates superior accuracy and advanced scalability compared to other state-of-the-art methods. The strong agreements between the predicted genome architectures and orthogonal experimental evidence, such as Capture-C, ChIA-PET and SPRITE, further highlight FLAMINGO's ability of capturing high-resolution spatial signatures of chromatin. Biologically, the reconstructed 3D structures facilitate additional discoveries and understandings, beyond 2D contact maps, such as higher efficiency of identifying multi-way chromatin interactions, interpretation of long-range QTLs, geometrical properties associated with TAD boundaries, and providing structural references to analyze single-cell variabilities of chromatin folding. Furthermore, FLAMINGO, along with its integrative version iFLAMINGO, addresses four fundamental

challenges in 3D genome modeling: (1) high scalability to reconstruct high-resolution 3D structures for all chromosomes from massive Hi-C datasets; (2) robust performance to handle large portions of missing data in Hi-C; (3) accurate cross cell-type prediction of 3D structures for cell-types lacking Hi-C datasets; and (4) boosting the resolution of reconstructed 3D structures from low-resolution Hi-C contact maps. Given all these advantages, FLAMINGO will be an important tool for both computational and experimental studies on 3D genomes. The reconstructed high-resolution structures across different cell-types will significantly facilitate biological insights into the spatial organization of chromatin and its underlying mechanisms.

As one of the major benefits of FLAMINGO, the generated high-resolution 3D structures can serve as a platform to understand how transcriptional regulation is modulated in 3D space. Overlaid with functional genomics data, FLAMINGO predictions provide high-resolution structural supports for long-range regulatory links between enhancers and promoters (Figure 2.3.e-h), and recover the short 3D distances between CTCF-associated boundaries of chromatin loops (Figure 2.6.a-d, Figure A.1). Moreover, beyond 2D chromatin contact maps, FLAMINGO can help to analyze higher-order multi-way interactions (Figure 2.4.a-c) and long-range cis-regulatory QTLs (Figure 2.4.d-g), and characterize geometrical signatures of chromatin shapes (Figure 2.5.a-b). In recent years, deep learning models have been developed to predict regulatory interactions in gene regulation and TAD organization from DNA sequences[75, 76]. Since FLAMINGO and deep learning models have complementary algorithmic strengths, it is expected to gain system-level knowledge on the relationship between gene regulation and chromatin organization by combining FLAMINGO with these deep learning algorithms.

The optimized consensus structure provides an efficient representation of the 3D genome for biologists with the advantage of high interpretability. Another type of methods aim to infer variations of the underlying chromatin structures, namely ensemble structures, using either polymer simulation models[77-79] or machine learning algorithms[69, 70]. While modeling structural variations is important, it is sometimes difficult to biologically interpret an individual structure from a pool of predictions and to delineate experimental cell-to-cell variations from the increased noisy fluctuations. As shown in the comparisons between the reconstructed structure and the ensemble of single-cell structures, including both ensemble average structures and variable cluster-specific structures (Figure 2.5.c-d), FLAMINGO's predictions can serve as effective reference structures to standardize the relative variabilities across single cells. Equipped with the complementary advantages of accuracy and robustness against noise, FLAMINGO can help the ensemble-structure learning algorithms to improve both the predictive performance and the interpretation of structures.

There are currently two limitations of FLAMINGO, which require future methodology developments. First, although the transformation function from Hi-C contact frequency to spatial distance has been justified for intra-chromosomal contacts by previous studies[14, 34] and our analyses (Figure 2.5.c-d, Figure A.4, Figure A.17), there is currently no systematic estimation of the function for inter-chromosomal contacts. Thus, FLAMINGO can only reconstruct 3D structures for each chromosome separately, while it is difficult to assemble the structure for the whole genome including inter-chromosomal distances. Similarly, due to the lack of sequencing reads, centromere and telomere regions are excluded from the reconstruction of spatial chromosome conformations. These regions,

especially centromere regions that have been demonstrated to be important in regulating chromatin organization by previous studies[69, 80], are components that should not be excluded if organizations for the whole genome are to be assembled. In order to achieve complete reconstructions of 3D genome, future algorithmic developments will be needed to overcome this limitation. Second, the consensus structure predicted by FLAMINGO represents the population-average architecture from large numbers of cells, which can not capture the highly dynamic property of 3D chromatin[81, 82] (such as the dynamic chromatin loops and TADs). The multi-scale spatial conformation of chromosomes varies from cell to cell[83] and the variability plays important roles in epigenetics, gene regulation and DNA damage repair[84]. A series of ensemble-structure prediction algorithms have been developed to explore the dynamic conformations[69, 70, 77-79]. As a future development that can help to further overcome this limitation, single-cell Hi-C datasets will be needed to predict 3D structures for individual cells. Single-cell Hi-C datasets are highly sparse and raise significant challenges in handling missing data. Although FLAMINGO demonstrates superior performance against missing data for bulk tissue Hi-C datasets even with ~98% missing rate at 5kb-resolution (Figure 2.6.e, corresponding to 50% down-sampling rate), typical single-cell Hi-C experiments have >99.99% missing rates at 100kb-resolution. Therefore, the highly sparse single-cell Hi-C datasets require further algorithmic improvements, in order to characterize the detailed structural variations across individual cells.

Overall, the combined strengths of handling large rates of missing data, making cross cell-type predictions, and boosting resolutions, suggest high impacts of FLAMINGO on 3D genome analyses. High-resolution structures can be inferred for diverse panels of cell-

types spanning different differentiation lineages, without increasing sequencing depths or requiring closely similar cell-types. Thus, it will not only improve the data mining of existing Hi-C data but also address the urgent need from large-scale Hi-C data resources to be generated in the near future, such as the 4D Nucleome Consortium. Together with the recent image-based 3D genome information[4] and the high-dimensional epigenomics data[6, 85], FLAMINGO is expected to substantially expand our understandings of the spatially orchestrated genome architectures across cell-types.

## 2.4 METHODS

### 2.4.1 Chromatin contact maps and epigenomics datasets

We collected the Hi-C chromatin contact maps of six human cell-types, including GM12878, K562, IMR90, HMEC, HUVEC, and NHEK, from the GEO database[10] (GEO:GSE63525). To remove potential biases in the Hi-C data, we normalized chromatin interaction-frequency matrices using the Knight-Ruiz normalization method as suggested by previous studies[10]. The normalized Hi-C interaction frequencies are then transformed into 3D Euclidean distances based on the exponential function[14] : $D_{ij} = IF_{ij}{}^{(-\eta)}$, where $D_{ij}$ represents the squared pairwise 3D distance between DNA fragments $i$ and $j$, $IF$represents the interaction frequency, and $\eta$ is a free parameter. In fact, after testing our model by taking different values of $\eta$ in the range suggested by previous experimental estimates[14, 28], we have found that the accuracy of reconstruction is robust to the choice of $\eta$ (Figure A.4). Therefore, by default, $\eta$ is set to 0.5 ($\eta/2 = 0.25$) as suggested by previous literature[14]. The validity of 3D distances converted from Hi-C contact maps, which are termed as observed distances from Hi-C in this paper, are also supported by the high similarity between the reconstructed structure and averaged structures of single

cell clusters, whose 3D configurations are directly obtained from imaging data (Figure 2.5.c and 2.5.d, Figure A.17).

The genome-wide DNase-seq datasets of chromatin accessibility from the six cell-types were collected from the ENCODE and Roadmap consortia[50, 86]. In a specific cell-type, for each DNA fragment, the averaged DNase-seq signal (namely fold-change over genomic background) within the fragment is used to represent the cell-type specific chromatin accessibility in the genomic locus. Additional details on data collection and preprocessing are given in Supplementary Note 1.

### 2.4.2 Model framework of FLAMINGO

FLAMINGO reconstructs 3D genome structures based on Hi-C chromatin contact maps using the low-rank matrix completion technique (Figure 2.1.a), which can efficiently delineate underlying low-rank structures from the large and noisy pairwise distance matrices. The cell-type specific 3D coordinates of high-resolution DNA fragments for each chromosome are predicted by solving a constrained rank-minimization problem using the augmented Lagrangian method[48], which can converge fast and can robustly handle large amounts of missing data.

To enable parallel computation, a hierarchy of two scales (1Mb and 5kb) is used to model each chromosome and an integrative assembly strategy is designed to build optimal high-resolution chromosomal structures from these two scales (Figure A.3). Based on simulated benchmark analysis, the performance of FLAMINGO does not rely on specific choices of resolutions or domain partitions (Figure A.7). In addition, an integrative variant of FLAMINGO, iFLAMINGO (Figure 2.7.a, Figure A.19), is also developed to incorporate

cell-type specific DNase-seq datasets into the model so as to (1) enable cross cell-type predictions and (2) boost resolution of predicted 3D genome structures.

### 2.4.3 Reconstruct 3D genome structures based on low-rank matrix completion

Each chromosome is modeled as a 'beads-on-a-string' polymer chain, where each DNA fragment is modeled as a bead, and the centromere and telomere regions are removed from the analysis as suggested by previous studies[33-35]. Structure reconstruction requires inferring the optimal 3D coordinates of consecutive DNA fragments along a chromosome, which maximally align with the pairwise 3D distances between DNA fragments observed from Hi-C data. A unique property of FLAMINGO is its capability to leverage the low-rank nature of a pairwise distance matrix from Hi-C; namely, the high-dimensional pairwise distance matrix is biologically generated by the underlying low-rank coordinate matrix of DNA fragments (rank≤3). Defined by the coordinate matrix ($\boldsymbol{P}$), the Gram matrix ($\boldsymbol{X} = \boldsymbol{P}\boldsymbol{P}^T$) has a rank≤3. The squared Euclidean distance matrix ($\boldsymbol{D}$) is a sum of three matrices: $\boldsymbol{D} = \text{diag}(\boldsymbol{X})\mathbf{1}^T + \mathbf{1}^T\text{diag}(\boldsymbol{X}) - 2\boldsymbol{X}$ where $\text{rank}(\boldsymbol{X}) \leq 3, \text{rank}(\text{diag}(\boldsymbol{X})\mathbf{1}^T) \leq 1, and\ \text{rank}(\mathbf{1}^T\text{diag}(\boldsymbol{X})) \leq 1$. Due to the property of ranks for matrix addition, the Euclidean distance matrix has a rank≤5. Based on the theory of matrix completion[42], the low-rank property of both the pairwise Euclidean distance matrix (rank≤5) and the Gram matrix (rank ≤ 3) guarantees that, under certain randomness assumptions on measurements, the underlying 3D structure can be predicted using a small fraction of data from Hi-C (Figure 2.1.a).

We define $\boldsymbol{P}$ as the $N\ by\ 3$ coordinate matrix for $N$ consecutive DNA fragments along a chromosome. We also define $D_{i,j}$ as the squared 3D spatial distance between DNA fragments $i$ and $j$. Thus, the objective function for 3D genome reconstruction is:

$$\min ||\boldsymbol{X}||_*$$

subject to $X_{i,i} + X_{j,j} - 2X_{i,j} = D_{i,j}, (i,j) \in \Omega;\ \boldsymbol{X}\mathbf{1} = 0; \boldsymbol{X} = \boldsymbol{X}^T; \boldsymbol{X}\ is$ positive semidefinite,

( 1 )

where $\boldsymbol{X} = \boldsymbol{P}\boldsymbol{P}^T$ is the Gram matrix, $||\boldsymbol{X}||_*$ represents the nuclear norm Tr($\sqrt{\boldsymbol{X}^T\boldsymbol{X}}$), which is related to the rank of matrix $\boldsymbol{X}$, and the measurement set $\Omega$ represents a subset of indices of DNA fragment pairs. We further introduce a linear sampling operator $A$ as:

$$A(\boldsymbol{X}) = f \in R^{|\Omega|*1}, f_i =< \boldsymbol{X}, \boldsymbol{\omega}_{\alpha_i} > \ for\ \alpha_i \in \Omega,$$

( 2 )

where $\alpha_i$=($\alpha_{i,1}, \alpha_{i,2}$) is the index of a DNA fragment pair. The matrix basis $\boldsymbol{\omega}_{\alpha_i}$ is defined as:

$$\boldsymbol{\omega}_{\alpha_i} = \boldsymbol{e}_{\alpha_{i,1},\alpha_{i,1}} + \boldsymbol{e}_{\alpha_{i,2},\alpha_{i,2}} - \boldsymbol{e}_{\alpha_{i,1},\alpha_{i,2}} - \boldsymbol{e}_{\alpha_{i,2},\alpha_{i,1}},$$

( 3 )

where $\boldsymbol{e}_{i,j}$ represents a matrix which has 1 at entry $(i,j)$ and 0 otherwise. For later use, we define the adjoint of $A$ as $A^*$, where $A^*\boldsymbol{y} = \sum_i y_i \boldsymbol{\omega}_{\alpha_i}$. The subset of DNA fragment pairs ($\Omega$ and $\alpha_i$) is randomly down-sampled from all measured pairs of DNA fragments with specified down-sampling rates. Intuitively, by defining $\boldsymbol{\omega}$ and $\alpha_i$, the linear operator $A$ summarizes all the constraints in one notation so that the objective function can be re-written in a compact form:

$$min_{\boldsymbol{P}}\ \text{Trace}(\boldsymbol{P}\boldsymbol{P}^T), \text{subject to } A(\boldsymbol{P}\boldsymbol{P}^T) = \boldsymbol{b},$$

( 4 )

where $\boldsymbol{b} = A(\boldsymbol{M})$ and $\boldsymbol{M}$ represents the true underlying low-rank Gram matrix from Hi-C data satisfying $M_{i,i} + M_{j,j} - 2M_{i,j} = D_{i,j}$.

A penalization term is further added to the objective function to control unexpected large distances predicted between adjacent DNA fragments caused by low Hi-C data quality at certain genomic locations. Therefore, the final objective function is:

$$\min_{\mathbf{P}} \text{Trace}(\boldsymbol{P}\boldsymbol{P}^T) + \lambda/2\|B(\boldsymbol{P}\boldsymbol{P}^T) - d^t\mathbf{1}\|_2^2, \text{ subject to } A(\boldsymbol{P}\boldsymbol{P}^T) = \boldsymbol{b}, \quad (5)$$

where $\lambda$ represents the penalization parameter, and the scalar $d^t$ represents the maximal allowed distance between adjacent DNA fragments. The linear measurement operator $B$ projects the Gram matrix to the sub-diagonal elements:

$$B(\boldsymbol{X}) = g(\boldsymbol{X}) \in R^{(n-1)*1}, \text{ where } g_i(\boldsymbol{X}) =< \boldsymbol{X}, \boldsymbol{\omega}_{\beta_i} > \; for\; \beta_i = (i, i+1), \text{ and } \mathbf{1} \in R^{(n-1)*1}. \quad (6)$$

The adjoint of $B$ is denoted as $B^*$, where $B^*\boldsymbol{y} = \sum_i y_i \boldsymbol{\omega}_{\beta_i}$.

Intuitively, the low-rank matrix completion model only needs a subset of the whole set of pairwise distances, which is indexed by $\Omega$, to reconstruct the Gram matrix $\boldsymbol{P}\boldsymbol{P}^T$, and it requires the optimal matrix $\boldsymbol{P}\boldsymbol{P}^T$ to follow three properties (Figure 2.1.a): (1) The rank of matrix $\boldsymbol{P}\boldsymbol{P}^T$ should be as small as possible by minimizing the trace of $\boldsymbol{P}\boldsymbol{P}^T$. This property is consistent with the low-rank assumption for 3D chromatin structures; (2) The pairwise distances based on the reconstructed 3D coordinates of DNA fragments should align with the subset of 3D distances indexed by $\Omega$ by satisfying the optimization constraints. This ensures that the model can accurately reconstruct 3D genome structures consistent with observed pairwise distances; (3) The 3D distances between adjacent DNA fragments are

bounded. This constraint removes unrealistically stretched structures of chromatin and guarantees a smooth genome structure.

Since the trace function Trace($\boldsymbol{P}\boldsymbol{P}^T$) is convex with respect to $\boldsymbol{P}$, we solve the optimization problem by the alternating-direction method of multipliers[49]. The augmented Lagrangian is given by:

$$L(\boldsymbol{P};\Lambda) = \text{Trace}(\boldsymbol{P}\boldsymbol{P}^T) + \lambda/2\|B(\boldsymbol{P}\boldsymbol{P}^T) - d^t\mathbf{1}\|_2^2 + r/2\|A(\boldsymbol{P}\boldsymbol{P}^T) - \boldsymbol{b} + \Lambda\|_2^2, \quad (7)$$

where $\lambda$ is the penalty parameter, $r$ is the regularization parameter, and $\Lambda$ is the Lagrangian multiplier. The gradient of the augmented Lagrangian with respect to $\boldsymbol{P}$ is given by:

$$2\boldsymbol{P} + 2\lambda B^*(B(\boldsymbol{P}\boldsymbol{P}^T) - d^t\mathbf{1})\boldsymbol{P} + 2rA^*(A(\boldsymbol{P}\boldsymbol{P}^T) - \boldsymbol{b} + \Lambda)\boldsymbol{P}. \quad (8)$$

Starting from $\Lambda = 0$ and a random initial guess for $\boldsymbol{P}$, the following iteration will continue until the error between the reconstructed and observed distances indexed by $\Omega$ is smaller than a specified threshold (default=$10^{-3}$): $\boldsymbol{P}$ is updated with the Barzilai-Borwein steepest descent method using the current $\Lambda$ and then $\Lambda$ is updated using the current $\boldsymbol{P}$ [49]. The accuracy of the model does not rely on the value of $r$ and $\lambda$, and we have set the parameters $r = 1$ and $\lambda = 10$ based on the previous study of low-rank reconstruction of the Euclidean geometry[49]. To tune the only free parameter of the model, $d^t$, which is the maximal allowed distance between adjacent DNA fragments, we test FLAMINGO on experimental Hi-C data using different values of $d^t$ to select the distance yielding the smallest objective function as the default value (Figure A.21.b), which is found to be

robust across different chromosomes and cell types (Figure A.21.c). This model demonstrates fast convergence when applied on both simulated data and experimental Hi-C data (Figure A.5; Figure A.21.b).

FLAMINGO has an intrinsic computational complexity $O(kN^2)$, where $k$ is a down-sampling rate to define the subset ($\Omega$) of DNA fragment pairs (Supplementary Note 1). Thus, FLAMINGO has sufficiently high scalability to predict high-resolution structures for large genomes, where $N$ is large. Moreover, by using the low-rank property of a 3D distance matrix, FLAMINGO can reconstruct 3D genome structures using a small down-sampling rate $k$, such as 0.2, which can substantially accelerate the optimization. Furthermore, the parallelized computation enabled by the hierarchical prediction strategy further boosts the reconstruction speed.

### 2.4.4 Assemble predicted structures from different scales

The same low-rank matrix completion algorithm is applied separately at two scales: (1) the 1Mb domain-level scale; and (2) the 5kb intra-domain scale. To construct the final 3D structure, the predicted intra-domain structures are assembled into the skeleton specified by the domain-level structures. At each 1Mb domain-level DNA fragment, the center of the corresponding intra-domain structure is assigned at the 3D coordinates predicted for the domain-level fragment. The assigned intra-domain structures are then rotated to minimize the overall reconstruction error between the predicted and the observed pairwise distances over DNA fragments across adjacent domains (inter-domain fragment distances) (Figure A.3). To identify the optimal 3D rotation matrices and control the corresponding computational cost, we search for a series of optimal 3D Givens rotation

matrices on each dimension. The 3D rotation matrices are then approximated by the multiplication of the 3D Givens rotation matrices.

Denote the predicted intra-domain structure for domain $i$ as $\boldsymbol{S}_i$. The optimal 3D Givens rotation matrices for the $x$-axis across domains are identified by:

$$\min_{\theta^i{}_x} \sum_{j,k}(||\boldsymbol{r}_{\theta^i{}_x}(\boldsymbol{S}_{i,j} - \boldsymbol{C}_i) + \boldsymbol{C}_i - \boldsymbol{S}_{i+1,k}||^2 - D_{\,i,j;i+1,k})^2, \quad (9)$$

where $\boldsymbol{r}_{\theta^i{}_x}$ is the 3D Givens rotation matrix of $\boldsymbol{S}_i$ for the $x$-axis with parameter $\theta^i{}_x$, $\boldsymbol{S}_{i,j}$ represents the DNA fragment $j$ within domain $i$, $\boldsymbol{C}_i$ represents the center of domain $i$ (which is inferred from the domain-level prediction), and $D_{i,j;i+1,k}$ represents the observed squared 3D distance between two inter-domain DNA fragments (fragment $j$ of domain $i$ and fragment $k$ of domain $i+1$) from adjacent domains. The same algorithm is applied to all domains consecutively to search for the rotation matrices of the $x$-axis for all domains. Intuitively, the objective function searches for the best rotation $\boldsymbol{r}_{\theta^i{}_x}$ of domain $i$ around its center $\boldsymbol{C}_i$ to match the distances between fragments across adjacent domains observed from the Hi-C data. The rotation matrices for the $y$-axis and $z$-axis are obtained similarly. Therefore, a series of 3D Givens rotation matrices are identified iteratively for the three axes. Multiplying the converged 3D Givens rotation matrices together yields the optimal 3D rotation matrices which are used to rotate the intra-domain structures, leading to the final genome structure. Since it jointly models all inter-domain distances between adjacent domains (i.e. off-diagonal points) and robustly identifies the global optimal rotation matrices for all intra-domain structures, the rotation algorithm will better align reconstructed structures with the Hi-C data and boost the accuracy of reconstruction.

### 2.4.5 Benchmark performance using simulated genome structures

To quantitatively benchmark the accuracy of FLAMINGO, we simulated 3D genome structures and generated matrices of squared pairwise distances between DNA fragments. The FLAMINGO algorithm was then applied to the squared pairwise distance matrices to reconstruct the 3D structures. The model performance was evaluated by comparing the reconstructed structure with the original structure in two ways. (1) The relative error between the reconstructed 3D coordinates ($\boldsymbol{C}_{\mathbf{re}}$) and the benchmark coordinates $\boldsymbol{C}_{\mathbf{benchmark}}$ of DNA fragments was calculated: $RE_{\text{coord}} = \|\boldsymbol{C}_{\mathbf{re}} - \boldsymbol{C}_{\mathbf{benchmark}}\|_2^2 / \|\boldsymbol{C}_{\mathbf{benchmark}}\|_2^2$. (2) The relative error between the reconstructed pairwise distance matrix ($\boldsymbol{R}$) and the original squared distance matrix ($\boldsymbol{D}$) was calculated: $RE = \left\|\boldsymbol{R} - \boldsymbol{D}^{(1/2)}\right\|_2^2 / \left\|\boldsymbol{D}^{(1/2)}\right\|_2^2$. Moreover, Spearman correlations between predicted and benchmark structures were also calculated to quantify the accuracy.

To test the performance of FLAMINGO with respect to missing data, we randomly down-sampled subsets of the squared pairwise distances as inputs and considered other squared pairwise distances as missing. Multiple down-sampled datasets were generated with different fractions of missing data in terms of different down-sampling rates. FLAMINGO was applied to these down-sampled squared pairwise distance matrices, and the resulting 3D coordinates of DNA fragments were used to calculate the relative errors and correlations.

To further test the performance of FLAMINGO on noisy inputs, we added two levels of white noise separately into the down-sampled squared pairwise distance matrices. As suggested by previous research[49], the first level of noise (Noise level 1) was generated

by the normal distribution $N(\delta, \delta)$, where $\delta$ represents the minimum value from the down-sampled squared pairwise distances. Similarly, the second level of noise (Noise level 2) was generated by the normal distribution $N(2\delta, \delta)$. In this way, the noisy down-sampled squared pairwise distances remain positive with high probability, consistent with the basic property of Euclidean distances. The simulations and down-sampling procedures were repeated 10 times for each benchmark setting.

To test the assembly algorithm, we divided the benchmark structure into different domains or fragments. The intra-domain structures were reconstructed separately and then assembled for the final structures, which were compared with the benchmark structure. The relative errors of pairwise distances and 3D coordinates were calculated to demonstrate the high accuracy of the assembly algorithm and its robustness with respect to different choices of domain partitions (Figure A.7).

### 2.4.6 Performance comparison based on experimental Hi-C data

For each of the six cell-types, we reconstructed the 3D structures using FLAMINGO at 5kb-resolution for each of the 23 chromosomes, based on the normalized Hi-C input datasets. To quantitatively evaluate the global reconstruction accuracy of FLAMINGO, we calculated the Spearman correlation coefficients between reconstructed and observed 3D distances for all pairs of DNA fragments, which are defined as all-points correlations. To further evaluate the accuracy of reconstructed intra-domain structures, we also calculated intra-domain correlations based on pairs of DNA fragments within the same domains. An accurately reconstructed structure is expected to demonstrate high correlations, at both all-point and intra-domain levels, which further suggest that the reconstructed structure quantitatively aligns with the observed Hi-C datasets.

We compared the performance of FLAMINGO with seven representative state-of-the-art algorithms:ShRec3D[33], GEM-FISH[34], Hierarchical3DGenome[35], SuperRec[37], ShNeigh[38] and RPR[36]. These methods were selected because they have been shown in previous studies to perform better than other methods using similar modeling strategies, and other existing methods are not included in the comparison because either they have been shown to have less accurate performance by previous studies or they do not practically converge at 5kb-resolution in our tests. All these methods were applied, based on their suggested parameters, on all of the 23 chromosomes in the six cell-types at 5kb resolution (Supplementary Note 1). GEM-FISH only finished for chromosome 21. ShRec3D, ShNeigh and RPR finished predictions only for short chromosomes (ShRec3D: chr13-22, ShNeigh: chr15-22 and chrX, and RPR: chr17-22). Hierarchical3DGenome and SuperRec finished predictions for all 23 chromosomes. The correlation coefficients based on those chromosomes with complete predictions were calculated using the same method as explained above. At 5kb-resolution, the run-times on an AMD EPYC processor with 25 cores were recorded. The maximum memory was set to be 100GB, sufficient for all algorithms.

To further quantify the performance of FLAMINGO with respect to large fractions of missing data, we randomly down-sampled the squared pairwise distance matrix with different down-sampling rates. Using the down-sampled input data, we tested the performance of FLAMINGO and other methods based on the correlation metrics described above. For each down-sampling rate, ten random samples with missing data were generated. The correlation coefficients were calculated for each random sample to evaluate the model performance. Because of impractically long computational times

needed by other methods for large chromosomes, only the chromosomes with complete predictions from these methods are included in this comparison.

As orthogonal biological information for model comparisons, we also collected significant long-range chromatin interactions profiled from different experiments, including ChIA-PET[72], Capture-C[45], and SPRITE[46]. For each chromatin interaction, we calculated the predicted 3D distances between the interacting DNA fragments from different reconstruction algorithms. Since interacting DNA fragments (anchors) are close to each other in 3D space, the algorithm is considered to have higher accuracy if it yields shorter predicted distances between interacting DNA fragments.

### 2.4.7 Analysis of multi-way chromatin interactions and QTLs

The multi-way chromatin interactions in GM12878 are collected from a dataset of SPRITE experiments[46]. To identify significant multi-way interactions, Market-Basket algorithm is used to search for higher-order associations of multiple genomic regions that are supported by SPRITE sequencing reads. Significant 3-way, 4-way and 5-way interactions are called based on confidence threshold=0.1 and support thresholds=$3x10^{-4}$, $2x10^{-4}$ and $1.7x10^{-4}$, respectively. The support thresholds are selected based on the curves of called significant multi-way interactions as a function of different thresholds, and the values corresponding to the elbow points are chosen. Genomic-distance controlled random samples of multi-way interactions are used to generate the background null distribution for statistical testing on the spatial distances among multi-way interacting anchors from the SPRITE data. To compare the fractions of SPRITE multi-way interactions captured by short predicted distances from FLAMINGO versus the fractions captured by short distances converted from Hi-C contact maps, distances are normalized by F-norm to

guarantee fair comparisons. A variety of thresholds of distances are used to define 3D spatial neighborhoods. A multi-way interaction is considered to be captured if all interacting anchors are located in the same 3D spatial neighborhood. The eQTL datasets[58-61] and hQTL datasets[62] are collected from matched cell-types, including whole blood cells and lymphoblastoid cells. The same normalization procedure is applied to compare the capability of assigning short spatial distances for QTLs based on the predicted distances versus the distances converted from Hi-C contact maps. Similarly, a variety of thresholds of distances are used to define 3D spatial neighborhoods. And long-range eQTLs (>900 kb) and distal hQTLs are evaluated whether it can be interpreted using the predicted spatial proximity by checking whether the SNP and the target region (i.e. a gene's promoter or histone modification peak) are predicted with shorter spatial distances, compared to samples of genomic-distance controlled random pairs. For every QTL, 1,000 random genomic-distance controlled pairs from the same chromosome are generated for comparison.

### 2.4.8 Curvature analysis for predicted 3D genome structures

To calculate the curvature in each 5kb genomic bin, a quadratic parametric function was fitted locally based on the specific genomic bin and the two neighboring upstream/downstream bins. Assume the parametric representation of the curve is $\overrightarrow{\boldsymbol{r}(t)} = (x(t), y(t), z(t))$, where each dimension can be written as a quadratic function, e.g. $x(t) = a_0 + a_1 t + a_2 t^2$ . By fitting the curve locally, the curvature is calculated as $\kappa = |\vec{\boldsymbol{r}}'' \times \vec{\boldsymbol{r}}'| / |\vec{\boldsymbol{r}}'|^3$. To have a fair comparison across different chromosomes, curvatures are normalized by the median values of each chromosome. Curvature is then calculated around TAD boundaries[10].

### 2.4.9 Comparison with image-based single-cell structures

3D coordinates of genomic bins at 30kb-resolution across single cells for a 2Mb region in chromosome 21 are collected[68] and compared with FLAMINGO's predictions. In K562, 797 single cells are kept for comparison by filtering out cells with >10% bins having no data (missing data). Linear interpolation is used to fill the missing coordinates in each single cell. To normalize the scales of structures, the 3D coordinate matrix ($\boldsymbol{P}$) of every single cell (30kb-resolution) is centered, and then scaled by the F-norm: $\boldsymbol{P}_{\text{scaled}} = \boldsymbol{P} / ||\boldsymbol{P}||_F$. Singular value decomposition (SVD) is then used to rotate and align the normalized single-cell structures (Supplementary Note 1). The average structure across single cells is calculated by taking the mean coordinates for each genomic bin. The predicted consensus structure by FLAMINGO (5kb-resolution) is centered, scaled and rotated using the same procedure, and is then aligned with the average structure of single cells or cluster-specific average structures. A 30kb-resolution version of the consensus structure is calculated by taking the average coordinates of six consecutive 5kb-resolution bins. Hierarchical clustering is applied on single-cell structures based on Euclidean distance to classify the ensemble of single cells into clusters, which can systematically represent the structural variabilities across single cells. After aligning the predicted consensus structure with variable single-cell structures, the differences of coordinates along the genomic region are calculated and compared to the intrinsic standard deviations among single cells.

### 2.4.10 Cross cell-type prediction of 3D genome structures

To predict 3D genome structures in cell-types without Hi-C datasets which are defined as target cell-types, we further expand the FLAMINGO algorithm to combine the Hi-C

dataset from a source cell-type and the DNase-seq dataset from the target cell-type, resulting in an integrative variant of FLAMINGO, named as iFLAMINGO. Intuitively, the Hi-C data from the source cell-type facilitate the inference of an approximate structure, which is fine-tuned by the cell-type specific DNase-seq data from the target cell-type.

Based on the observation that 3D distances between interacting DNA fragments are associated with chromatin accessibilities (Figure A.19a), we impute the 3D distances between any two DNA fragments in the target cell-type ($D_{i,j}$) based on DNase-seq signals and 1D genomic distances (Figure A.19b). The imputation is achieved by fitting a linear regression model in the source cell-type: $D_{i,j} = \alpha_1 S_i + \alpha_2 S_j + \alpha_3 G_{i,j}$, where $\alpha_1$, $\alpha_2$, and $\alpha_3$ are fitting parameters to be determined, $D_{i,j}$ represents the observed distance, $S_i$ represents the DNase-seq signal of DNA fragment $i$, and $G_{i,j}$ represents the 1D genomic distance between DNA fragments $i$ and $j$. Based on the fitted regression model, 3D distances between DNA fragments can be imputed in the target cell-type, using the target cell-type specific DNase-seq data, which are then summarized into a matrix $\boldsymbol{E}$. Therefore, the imputed 3D distance matrix $\boldsymbol{E}$ represents the target cell-type specific information which can be used to improve the reconstruction of the corresponding 3D structure.

The imputed 3D distance matrix is integrated into the original objective function as a penalization term, so that we will solve the following problem to reconstruct the 3D structure:

$$\min_{\mathbf{P}} \text{Trace}(\boldsymbol{P}\boldsymbol{P}^T) + \lambda/2\|B(\boldsymbol{P}\boldsymbol{P}^T - d^t\mathbf{1})\|_2^2 + \gamma\|A(\boldsymbol{P}\boldsymbol{P}^T) - A(\boldsymbol{E}^M)\|_2^2, \text{ subject to } A(\boldsymbol{P}\boldsymbol{P}^T) = \boldsymbol{b}\,, \quad (10)$$

where $\gamma$ is the penalization parameter and $\boldsymbol{E}^{M}$ is the Gram matrix of the imputed 3D distance matrix ($\boldsymbol{E}$) for the target cell-type. The penalization term tunes the reconstructed 3D structure in the target cell-type to align with the imputed 3D distances from DNase-seq. Hence, by borrowing information from the source cell-type Hi-C data, iFLAMINGO predicts the cell-type specific 3D genome structures in the target cell-type.

To validate the performance of cross cell-type predictions, iFLAMINGO was applied to 30 source-target cell-type pairs, based on the six cell-types with Hi-C data available. For each source-target cell-type pair, we predicted the 3D genome structure for the target cell-type based on the Hi-C data from the source cell-type and the DNase-seq data from the target cell-type. The reconstructed 3D structures for target cell-types were evaluated by calculating the correlation coefficients between the reconstructed 3D distance matrix and the observed one based on the Hi-C dataset from the target cell-type. As comparisons, we also evaluated the performance using the reconstructed 3D distance matrices solely based on Hi-C data from the source cell-type, without incorporating the DNase-seq information from the target cell-type.

### 2.4.11 Improve the resolution of 3D genome structures

iFLAMINGO integrates the high-resolution chromatin accessibility data to improve the resolution of predicted 3D genome structures, such as 5kb-resolution, based on relatively low-resolution Hi-C contact maps, such as 10kb-resolution. Given a Hi-C contact map at 10kb-resolution, we divide each 10kb genomic fragment into two consecutive 5kb fragments. The 5kb fragments inherit the same pairwise 3D distances from the original 10kb fragment. In this way, the $m\ by\ m$ 3D distance matrix at 10kb-resolution is expanded into a $2m\ by\ 2m$ 3D distance matrix at 5kb-resolution, which serves as the initial structure

for high-resolution reconstruction. The high-resolution DNase-seq datasets of chromatin accessibility are then incorporated to impute the 3D distances between 5kb DNA fragments, following the same method described above (Figure A.19b). By applying the iFLAMINGO algorithm on the expanded 3D distance matrix from a low-resolution Hi-C contact map and the imputed one from a high-resolution DNase-seq dataset, the 3D genome structure at 5kb-resolution is then reconstructed. We applied the model on the Hi-C dataset in GM12878 for all of 23 chromosomes at resolution of 10kb, 25kb, and 50kb, respectively. The model performance is evaluated using the correlation coefficients (all-points and intra-domain) between the reconstructed and the observed 3D distance matrices at 5kb-resolution.

# CHAPTER 3

# PREDICT HIGH-RESOLUTION SINGLE-CELL 3D CHROMOSOME STRUCTURES USING TFLAMINGO

## 3.1 INTRODUCTION

The 3D chromosome structures provide the structural foundation of gene regulation, DNA replication, and cell differentiation. Comprehensive profiling of the 3D chromosome structures is important for understanding the structural basis of the interplay between genes, regulatory elements, and genetic variants. Chromosome conformation capture-based methods, including Hi-C and Capture-C, have been widely used to profile the contacts between DNA fragments in different cell types/tissues and generate important observations of the genome structures, such as chromatin loops, topologically associated domains (TADs), and chromatin compartments. However, the chromatin contact maps generated in bulk tissue only represent the average structure of millions of cells, thus cannot reflect the dynamic 3D chromatin structures across single cells.

In recent years, the toolbox for measuring the chromosome conformations in single cells has been largely expanded, including Dip-C, single nucleosome Hi-C (snHi-C), single-nucleus methyl-3C sequencing (snm3C) and single-cell Hi-C (scHi-C). These experimental methods measure contact frequencies between DNA fragments in individual cells and generate massive single-cell chromatin contact maps. These datasets push the understanding of the chromosome conformation from bulk tissue to singe cells and innovate the variable cell-to-cell chromosome structures. However, limited by the low sequencing depth, the single-cell chromatin contact maps are highly sparse in high

resolution (i.e. >99.9% missing rate at 10kb resolution), making it highly challenging to further study the high-resolution 3D chromosome structures. To address this emerging question, computational methods to predict the high-resolution single-cell 3D chromosome structures are highly desired.

In general, previous efforts in computationally predicting 3D chromosome structures can be classified into two categories: MDS-based methods and simulation-based methods. In the first category, the observed interacting frequencies from the single-cell chromatin contact maps are firstly converted to the spatial distance. The 3D chromosome structures are reconstructed from the derived distance matrices using the MDS-based methods or the recurrent plots. ShRec3D and RPR are representative methods in this category. These methods are solely data-driven and do not have any additional assumptions on the 3D structures. However, these methods cannot handle a significant fraction of missing data and demonstrate a relatively low accuracy in reconstructing high-resolution single-cell 3D chromosome structures. In the second category, the observed contacts from the single-cell chromatin contact maps are used as constraints in simulating 3D chromosome structures based on the polymer simulation models. The representative algorithms include isdHi-C, Si-C, and NucDynamics. In the simulation process, the algorithms simulate a 3D chromosome structure based on the biophysics properties of the DNA sequences and further refine the structures to maximize the contact probabilities of the observed interacting anchors. Benefiting from the polymer simulation, these methods have been successfully applied to reconstruct the single-cell chromosome structures. However, the simulation-based methods have strong prior assumptions about the chromosome structures. Based on the objective functions of these methods, the invariant

biophysical properties of single cells are considered as equally important as the dynamic single-cell chromatin contact maps, which may result in a decreased ability in capturing the structural variations across single cells. Additionally, these algorithms are configured with pre-defined parameters, requiring additional parameter selection procedures for different datasets.

To address these problems, we developed a low-rank tensor completion-based method, tFLAMINGO, to reconstruct high-resolution 3D chromatin structures from single-cell chromatin contact maps. As a powerful tool in video reconstruction and compression, the low-rank tensor completion methods leverage the similarity between frames to infer the missing pixels. Similarly, tFLAMINGO models every single-cell chromatin contact as a frame, and models the whole dataset as a video, which facilitates the information sharing across single cells to complete the missing data. Apart from the low-rank property of the tensor, tFLAMINGO further utilized the low-rank property of the single-cell chromatin contact maps, which guarantees the underlying single-cell 3D chromosome structures can be recovered using a subset of pairwise distances. These two algorithmic advantages distinguish tFLAMINGO from existing methods with superior accuracy in reconstructing the single-cell 3D chromatin structures and strong abilities in capturing the dynamic structural variabilities.

We applied tFLAMINGO on four single-cell chromatin conformation datasets (Dip-C, snHi-C, snm3C and scHi-C) in three cell types and reconstructed the 3D chromatin structures for all single cells in 10kb and 30kb resolution. Based on the extensive simulated datasets and experimental bulk tissue chromatin contact maps, tFLAMINGO demonstrates superior performance over existing methods in predicting 3D chromosome

structures. Beyond the robust compartment and TAD structures across single cells, the predicted 3D structures by tFLAMINGO capture the dynamic single-cell chromatin interactions, which allow us to evaluate dynamic gene regulations in 3D space. Furthermore, the predicted 3D structures of tFLAMINGO provide new biological insights into the mechanistic interpretation of GWAS SNPs and high-order chromatin interactions.

## 3.2 RESULTS

### 3.2.1 tFLAMINGO reconstructs high-resolution single-cell 3D chromosome structures

Single-cell chromatin conformation capture (3C) experiments measure the 3D chromosome structures and generate the chromatin contact maps for tens to hundreds of cells simultaneously (Figure 3.1.a). Limited by the low sequencing depth, the single-cell chromatin contact maps are highly sparse and contain large fraction of missing data in high resolution (>99.9% in 10kb resolution). Unlike existing methods, which models every single cell separately, tFLAMINGO jointly models the whole single-cell 3C dataset as a tensor, where frontal slices represent the single-cell chromatin contact maps (Figure 3.1.a). Such formalism of tFLAMINGO enables the imputation of missing contact frequencies in one cell to borrow information from other cells, thus mitigating the high missing rates of single-cell 3C datasets and accurately reconstructing single-cell 3D chromosome structures.

Given a sparse tensor of the single-cell 3C dataset, tFLAMINGO constructs a low-rank dense tensor that optimally aligns with the observed entries from the inputs, thus completing the missing values (Figure 3.1.a). Computationally, the construction process is achieved by minimizing the tensor tubal rank of the dense tensor and requiring the

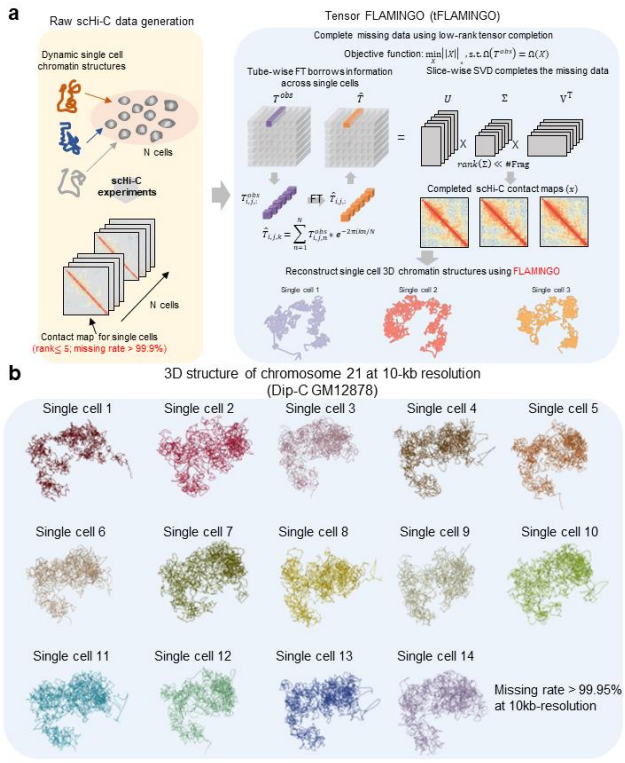



**Figure 3.1 Overview of tFLAMINGO.** (**a**) Schematic figure of tFLAMINGO. Biologically, the scHi-C experiment generates the highly-sparse contact maps for N cells. For every single cell, the distance matrix derived from the scHi-C experiment is a low-rank matrix (rank≤5). Thus, the tensor organizing the distance matrices of N cells is a low-rank tensor and the missing values can be completed using the low-rank tensor completion method. To accurately reconstruct the 3D chromatin structure of single cells at high resolution, tFLAMINGO utilizes the tube-wise Fourier Transformation to borrow information across single-cells, while keeping the geometric character of every single cell. Based on the completed distance matrix, FLAMINGO is used to reconstruct the 3D chromatin structure for every single cell. (**b**) Reconstruction of the 3D chromatin structure for 14 single cells at 10kb-resolution by tFLAMINGO using GM12878 Dip-C data.

reconstructed values equal to the observed values on the measurement set. Based on the completed single-cell chromatin contact maps, the 3D chromosome structures are predicted by our in-house 3D reconstruction algorithm, FLAMINGO, for every single cell.

The key design of tFLAMINGO is to model the whole single-cell 3C dataset as a tensor with dimension $M \times M \times N$, where $M$ represents the number of genomic loci and $N$ represents the number of single cells. The low-rank tensor completion method has been

widely used to represent large scale high-dimensional datasets with low-dimensional features, and its application includes video compression and reconstruction (Figure 3.1.a). In the single-cell 3C dataset, the low-rank properties are guaranteed in two aspects. Firstly, the tensor summarizing all single-cell chromatin contact maps is a low-rank tensor. This is because the cell-type-specific chromosome structures are observed to be robust at low-resolution (i.e. > 1MB resolution), suggesting single cells share a consensus backbone structure. In the single-cell 3D dataset, the information is redundant since the consensus structure is repeatedly measured in all single-cell chromatin contact maps. Therefore, the single-cell chromatin contact maps are complementary in terms of characterizing the consensus structure, concluding the low-rank property of the single-cell 3D dataset. Based on this property, the missing values in one cell can be inferred by borrowing information from the measurements of the same contacts in other cells. In tFLAMINGO, the information integration is facilitated by the tube-wise Fourier Transformation across all cells. Secondly, single-cell chromatin contact maps are low-rank matrices. According to the Euclidian geometry, the $M \times M$ chromatin contact map is induced by the $M \times 3$ coordinate matrix, thus having rank$\leq 5$. This property guarantees that the chromatin contact maps can be fully compressed and reconstructed using up to five singular values, which is far less than the number of genomic loci at high resolution. Thus, the chromatin contact maps can also be recovered based on a small fraction of the observed entries. By taking the advantage of the low-rank properties, tFLAMINGO facilitates large-scale information sharing within and across single cells, and completes the missing values of the sparse single-cell chromatin contact maps.

Our previously developed algorithm, FLAMINGO, is used to predict the 3D chromosome structures from the completed single-cell chromatin contact maps. FLAMINGO demonstrates superior performance and scalability in reconstructing high-resolution 3D chromosome structures. These features are especially important for reconstructing the high-resolution single-cell chromosome structures, which involves predicting the 3D

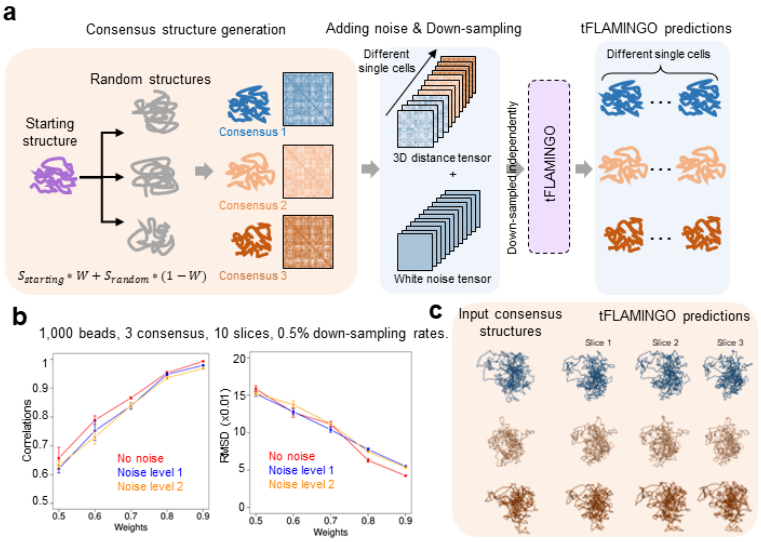



**Figure 3.2 Simulation analyses of tFLAMINGO.** (**a**) Schematic figure of simulations. Three consensus structures are generated from the same starting structure with parameter $W$ controlling the similarity between consensus structures. Each consensus structure is repeated ten times to generate a tensor with 30 frontal slices. The resulting tensor is further mixed with different levels of noise (no noise, noise level 1, and noise level 2) and down-sampled. The highly noisy and incomplete tensor is used as the input of tFLAMINGO to reconstruct the consensus structures. (**b**) Performance of tFLAMINGO under different weights with 1,000 beads and 0.5% down-sampling rates. (**c**) Examples of benchmark consensus structures and tFLAMINGO predictions (weights=0.6, correlations=0.783, RMSD=0.133).

location of several thousand genomic loci. Remarkably, FLAMINGO takes less than 25 hours to reconstruct the 3D structure of human chromosome 1 in 1kb resolution.

We applied tFLAMINGO on four single-cell 3C datasets to predict single-cell 3D chromosome structures in 10kb and 30kb resolution, providing the largest cohort of high-

resolution single-cell chromatin structures. As an example, tFLAMINGO predicted the 3D chromosome structures of chromosome 21 for 14 GM12878 cells based on the Dip-C data, whose missing rate is over 99.95% in 10kb resolution (Figure 3.1.b, Figure B.1-3).

### 3.2.2 Performance validation based on the simulation analyses

The performance of tFLAMINGO is firstly validated by reconstructing the simulated benchmark structures. We simulated a sparse tensor with three consensus structures, whose similarity is controlled by the weight $W$ (Figure 3.2.a). tFLAMINGO is applied on the simulated dataset to reconstruct the underlying 3D structures. Across a wide range of weight $W$, the predicted 3D structures of tFLAMINGO are highly coherent with the benchmark consensus structures, verifying that tFLAMINGO can capture the structural variations across single cells. As a representative example, at weight 0.6, the predicted 3D structures accurately capture the unique layouts of different consensus structures (Figure 3.2.c, Figure B.11). Moreover, the predicted 3D structures of the frontal slices are classified into three clusters based on the pairwise RMSD, which is consistent with their original identities during the data generation process. In addition to weights, tFLAMINGO demonstrates exceptional performance on simulated datasets with the different numbers of beads and frontal slices, as well as different down-sampling rates (Figure B.7-10). Remarkably, tFLAMINGO can accurately reconstruct the 3D structure with 3000 beads

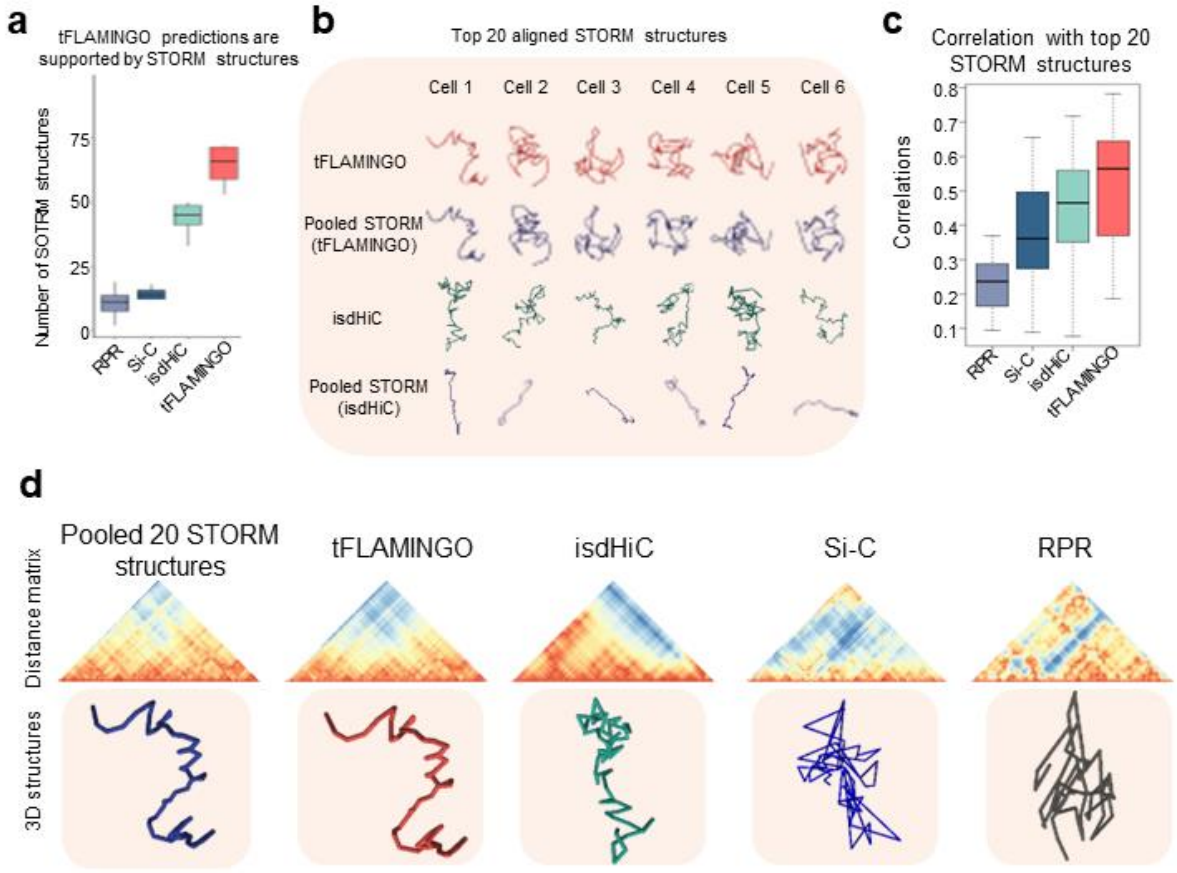



**Figure 3.3 Performance validation based on the STORM dataset.** (**a**) The number of STORM structures that are correctly captured by predicted single-cell structures for all methods. The center lines of boxplots show the median, the upper and lower box limits show the 25th and 75th percentiles respectively. The whiskers extend up to 1.5 times the interquartile range away from the limits of the boxes. (**b**) Examples of predicted 3D chromatin structures and top 20 aligned STORM structures for tFLAMINGO and isdHi-C. (c-d) tFLAMINGO accurately reconstructs the underlying 3D structures from snHi-C data. For each snHi-C single cell, the correlations between the raw snHi-C distance matrix and STORM distance matrices are calculated. The top 20 correlated STORM structures are considered to represent the true underlying 3D structures of the snHi-C distance matrix. (**c**) tFLAMINGO predictions show the highest correlations with the top 20 STORM structures. The center lines of boxplots show the median, the upper and lower box limits show the 25th and 75th percentiles respectively. The whiskers extend up to 1.5 times the interquartile range away from the limits of the boxes. (**d**) Example of the predicted 3D chromatin structure of snHi-C single cell 1. tFLAMINGO shows the highest correlation with the pooled STORM structures (correlation=0.676).

based on only 0.1% of the pairwise distances (Figure B.7-8, correlation=0.647, RMSD = 0.193), proving tFLAMINGO is able to reconstruct the high-resolution 3D chromosome structures from highly sparse chromatin contact maps.

Furthermore, we compared the performance of tFLAMINGO with the baseline method, which completes the missing values by averaging all frontal slices. Strikingly, tFLAMINGO shows significantly higher accuracy over the baseline method, demonstrating the algorithmic design of tFLAMINGO is necessary to reconstruct the single-cell 3D chromosome structures accurately (Figure B.10).

### 3.2.3 Performance comparison based on the STORM dataset

The performance of tFLAMINGO in reconstructing the 3D structures of human chromosome 21 for 14 K562 cells is benchmarked with the STORM dataset and compared with four state-of-art algorithms: RPR, ShRec3D, Si-C and isdHiC. The similarity between the predicted structures and STORM structures are quantified by the Spearman correlations, which has been widely used to quantify the accuracy of structure reconstructions. Since the STORM experiment measures the 3D structures of a 2MB region in thousand cells, we calculated the Spearman correlations between all pairs of predicted structures and STORM structures. Such a comparisons provides direct evidence of the model performance at the single-cell level. Firstly, we evaluated the consistency between the predicted structures and STORM structures. On average, tFLAMINGO predictions are supported by 73.4 STORM structures (Figure 3.3.a, correlation>0.8), while less than 50 STORM structures support other methods. For example, the predicted single-cell 3D structures of tFLAMINGO align well with the average of the top 20 STORM structures based on the correlations. In comparison, the

predictions of isdHiC demonstrate distinct structures with the STORM data (Figure 3.3.b). Secondly, the reconstruction accuracy of different algorithms is evaluated using the STORM structures. Since the underlying single-cell 3D chromosome structures of snHi-C data are unknown, the average of the top 20 STORM structures with the highest similarity with the raw snHi-C contact maps is used as the gold-standard to evaluate the model performance. Across all methods, tFLAMINGO demonstrates the highest correlations (Figure 3.3.c, median correlation 0.56). Figure 3.3.d shows one example where tFLAMINGO accurately reconstructs the underlying structures (correlation = 0.853) while other methods demonstrate lower accuracy (correlation < 0.4). These results not only validate the accuracy of tFLAMINGO in predicting single-cell 3D chromosome structures, but also suggest that the prediction of tFLAMINGO widely exists in the population of K562 cells, thus supporting the biological significance of tFLAMINGO.

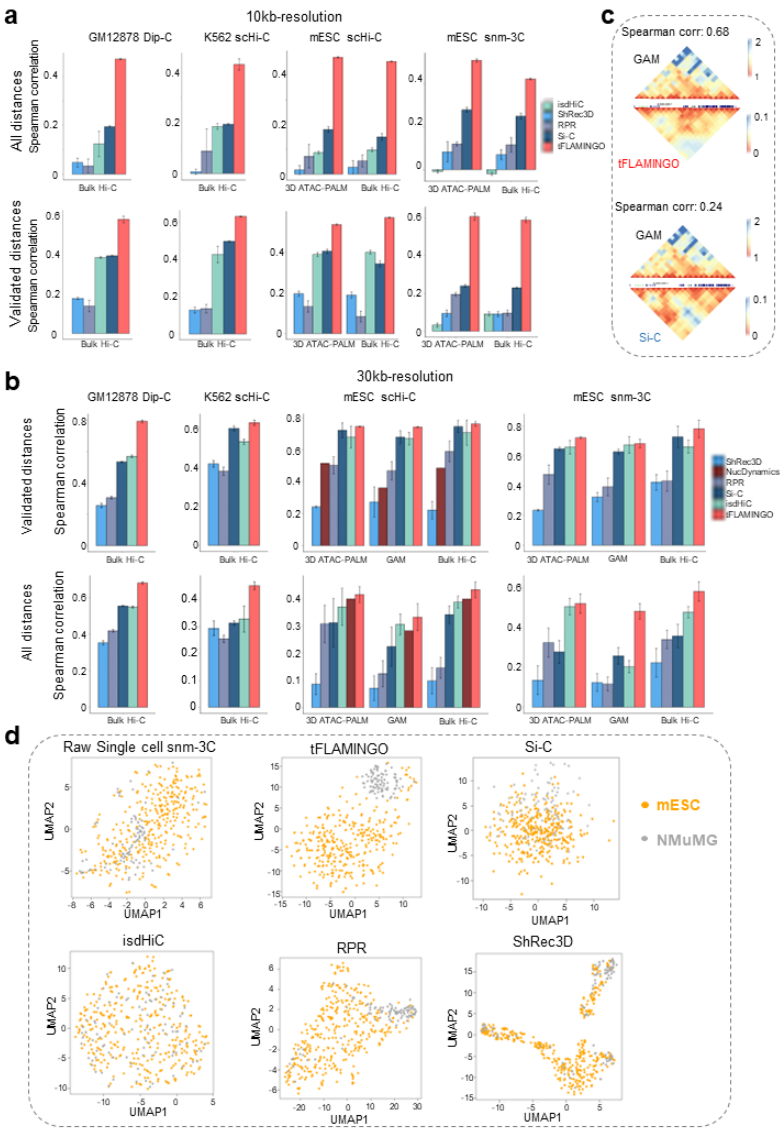



**Figure 3.4 Systematic performance comparison in reconstructing single-cell chromosome structures.** (**a**) Performance comparison in 10kb-resolution across four single-cell datasets (GM12878 Dip-C dataset, K562 snHi-C dataset, mESC scHi-C dataset, and mESC snm3C dataset). The chromatin contact maps generated in the matching tissues are used as gold standards. (**b**) Performance comparison in 30kb-resolution across four single-cell datasets (GM12878 Dip-C dataset, K562 snHi-C dataset, mESC scHi-C dataset, and mESC snm3C dataset). The predictions of NucDynamics provided by Si-C are directly used. (**c**) Example of the accurately reconstructed distance matrix of tFLAMINGO at chr19:8,200,000-9,100,000. (**d**) UMAP visualization of the reconstructed distance matrices. tFLAMINGO accurately predicts the cell-type-specific chromatin structures.

### 3.2.4 Performance comparison based on the bulk tissue chromatin contact maps

The performance of tFLAMINGO is systematically evaluated and compared with existing algorithms based on the bulk tissue chromatin contact maps. As the orthogonal evidence, the bulk tissue chromatin contact maps from Hi-C, 3D ATAC-PALM and GAM are collected to verify the predicted pairwise distance matrices of different algorithms. For every single cell, Spearman correlations based on two sets of distances are calculated to quantify the model performance: (1) Spearman correlations based on the measured distances in the bulk tissue datasets (termed as 'all distance correlations') and (2) Spearman correlations based on the measured distances in both bulk tissue datasets and each single-cell contact map (termed as 'validated distance correlations'). Comparing these two metrics, all distance correlations tend to quantify the accuracy of the completed missing values, while the validated distance correlations evaluate the accuracy of recapitulating the observed values. Strikingly, tFLAMINGO demonstrates superior performance in reconstructing the single-cell 3D chromosome structures, especially at 10kb resolution (Figure 3.4.a-b). More importantly, tFLAMINGO shows even more improvement over existing methods based on all distance correlations (Figure 3.4.a-b , tFLAMINGO: 0.52, other methods < 0.3), suggesting a highly enhanced ability in imputing the missing pairwise distances. The advanced performance stems from the algorithmic design of tFLAMINGO. Unlike the simulation-based methods, where the missing values are completed based on the polymer simulation, tFLAMINGO uses the low-rank structures learned from the observed data to impute the missing values, thus showing better consistency with the biological ground truth. At 30 kb resolution, tFLAMINGO still archives consistently high accuracy. Figure 3.4.c shows a representative example of

predicted chromatin contact maps predicted by tFLAMINGO and Si-C. At this 1MB genomic region, the distance matrix predicted by tFLAMINGO accurately recapitulates the domain structures and long-range chromatin interactions of the GAM chromatin contact map (correlation 0.68), which are not observed in the prediction of Si-C

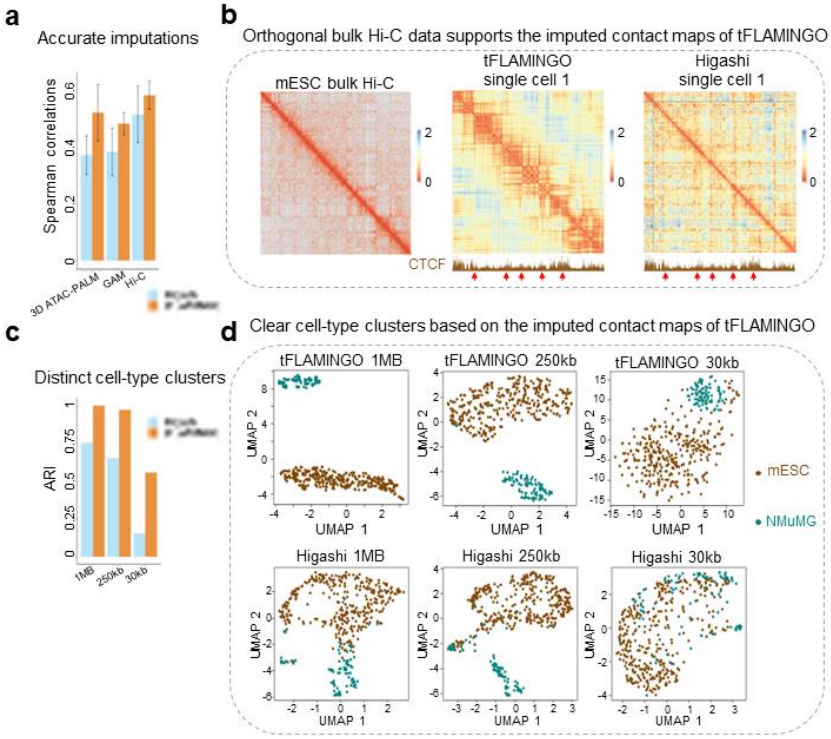



**Figure 3.5 Systematic performance comparison in imputing high-resolution single-cell chromatin contact maps.** (**a**) Evaluation of the accuracy of imputed single-cell contact maps on mESC snm3C datasets at 30kb-resolution. Bulk tissue chromatin contact maps generated by orthogonal experiments are used as gold standards. The error bar represents the standard deviations across 351 single cells. (**b**) Example of imputed single cell contact maps by tFLAMINGO and Higashi. The TAD pattern predicted by tFLAMINGO aligns with the bulk-tissue CTCF Chip-seq peaks. (**c-d**) Performance comparison in identifying cell types based on the imputed contact maps across different resolutions. (**c**) The quantitative accuracy is evaluated by the Adjusted Rand Index (ARI). (**d**) UMAP of distance matrices predicted by Higashi and tFLAMINGO. Dots represent single cells and are colored by the cell types.

(correlation 0.24). These results provide quantitative support for the superior performance of tFLAMINGO in reconstructing the single-cell 3D chromosome structures.

Furthermore, we evaluated the performance of different algorithms in capturing structural variations across cell types. All algorithms are applied to the snm3C dataset, including 351 mESC cells and 96 NMuMG cells, to reconstruct the single-cell 3D chromosome structures. The distance matrices induced by the predicted 3D chromosome structures are projected into the two-dimensional space using UMAP to discover clusters of cells. Since the great majority of pairwise distances are missing from the raw single cell snm3C chromatin contact maps at 30kb resolution (missing rate >99.9%), the observed values from the raw single cell snm3C dataset cannot reflect the cell-type specific structural variations and only one cloud of cells are observed (Figure 3.3.d). By jointly modeling all cells within the same cell type, tFLAMINGO identifies the cell-type-specific structures and correctly projects cells into the matching clusters(Figure 3.3.d). In comparison, no clear clusters of cells are observed in the UMAP plots based on the predictions of other methods(Figure 3.3.d). These results suggest that, by jointly modeling all cells, tFLAMINGO has better abilities to complete the missing data and capture the cell-type specific structural variations of single cells.

### 3.2.5 Performance comparison in imputing high-resolution chromatin contact maps

Currently, analyses of the single-cell 3D chromatin structures in high resolution are significantly hindered by the sparsity of the single cell datasets. To enhance the usability of single-cell datasets at high-resolution, the computational methods are developed to impute the single-cell chromatin contact maps and Higashi is the latest and most powerful one. In tFLAMINGO, the 3D distances between the DNA fragments are naturally induced by the predicted single-cell 3D chromosome structures, thus leading to a complete

distance matrix. By further converting the spatial distances to interaction frequencies using the observed negative exponent function, tFLAMINGO can help to impute the high-resolution chromatin contact maps.

To evaluate the performance of tFLAMINGO in imputing chromatin contact maps, tFLAMINGO is applied to the snm3C dataset of 351 mESC cells to impute the single-cell chromatin contact maps in 30kb resolution and compared with Higashi. The performance is evaluated based on the correlations between the imputed chromatin contact maps and bulk tissue chromatin contact maps measured by 3D ATAC-PALM, GAM, and Hi-C in bulk mESC cells. Compared with Higashi, tFLAMINGO demonstrates higher correlations (Figure 3.4.a, correlations > 0.44) across all comparisons. As Figure 3.5.b shows, the chromatin contact map imputed by tFLAMINGO for the single-cell 1 accurately captures the TAD structures of the bulk Hi-C contact maps. Furthermore, the TAD boundaries in the imputed single-cell chromatin contact maps are supported by the CTCF binding, which further illustrates the accuracy of tFLAMINGO. In comparison, no clear TAD structures are observed in the chromatin contact map imputed by Higashi, and the TAD boundaries are not consistent with the CTCF binding profiles.

The accuracy of imputation is also evaluated by the ability in identifying cell-type-specific structures. tFLAMINGO and Higashi are used to impute the chromatin contact maps for 351 mESC cells and 96 NMuMG cells in 1MB, 250 kb, and 30kb resolutions (Figure 3.4.c). The cells are further clustered based on the imputed distance matrices in the two-dimensional space to assign cell identities. Adjusted Rand Index (ARI) is used to evaluate the similarity between the predicted cell-type identities and the ground truth. At 1MB and 250 kb resolution, both Higash and tFLAMINGO demonstrate high ARI. However,

tFLAMINGO can still correctly predict the cell identify in 30kb resolution (Figure 3.4.c, ARI: 0.53, while Higashi fails to correctly identify clusters of different cell types (ARI: 0.13), suggesting tFLAMIGO enjoys a better ability in capturing the high-resolution cell-type-specific structural variations. As Figure 3.5.d shows, the chromatin contact maps imputed by tFLAMINGO can be robustly clustered into two cell clusters across all resolutions. In contrast, the cell clusters based on the imputation of Higashi are gradually merged as

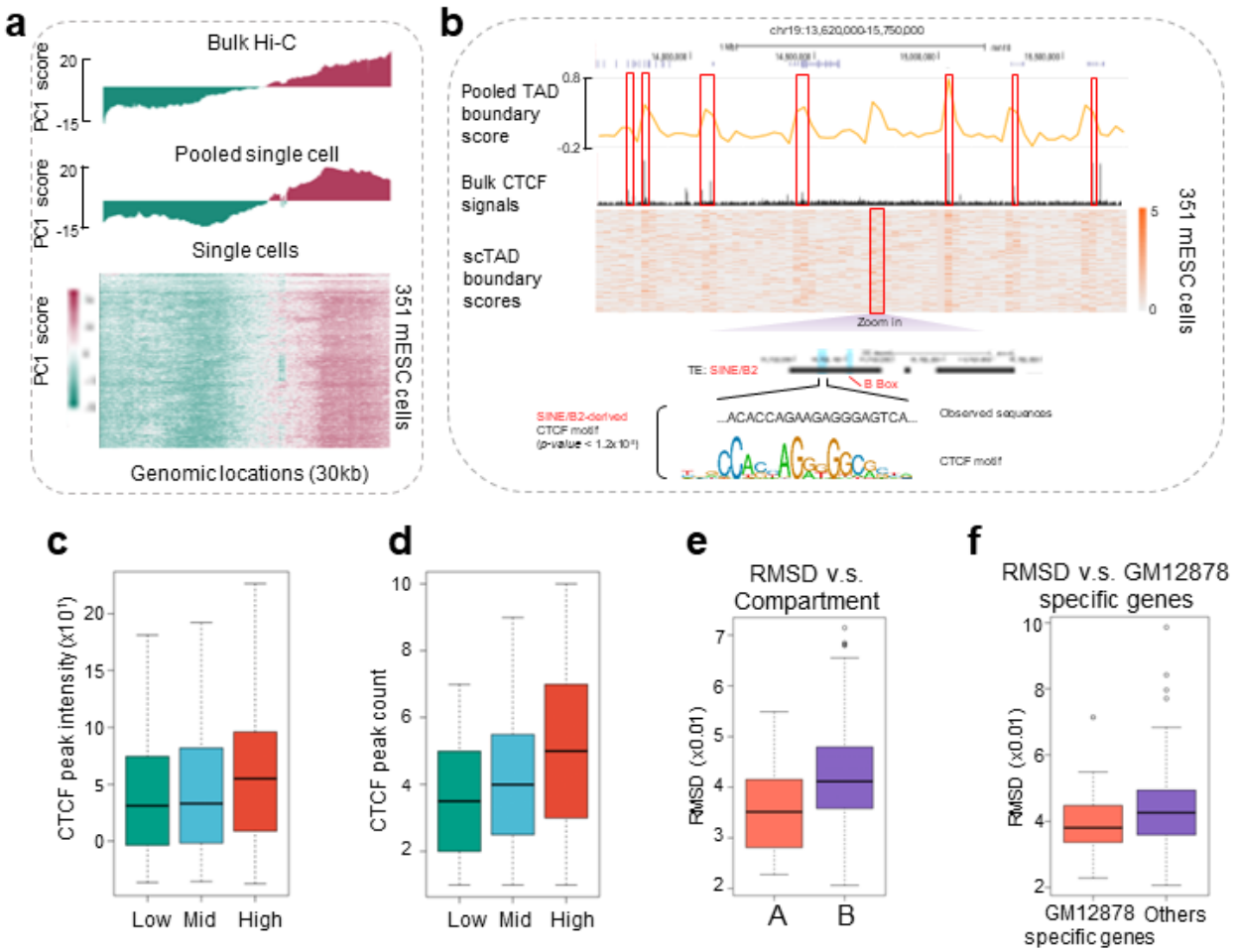



**Figure 3.6 Compartment analyses and TAD analyses in single cells.** (**a**) Compartment identification of chromosome 19 for 351 mESC single cells at 30kb-resolution. PC1 scores based on the bulk Hi-C contact maps pooled single-cell contact maps and all 351 single cells are shown. (**b**) Example of TAD boundary identification at chr19:13,620,000-15,750,000. Seven out of eight predicted TAD boundaries are

**Figure 3.6 (cont'd)**

supported by the bulk tissue CTCF Chip-seq dataset. The TAD boundary without CTCF Chip-seq peak contains a transposable B2 SINE element with B-box and a TE-derived CTCF motif. (**c-d**) Regions with higher TAD boundary scores tend to have (**c**) higher CTCF binding strength and (**d**) a higher number of CTCF peaks (High: <20% quantile; Mid: 20%-80%; Low: >80%). (**e-f**) Functional regions tend to have higher structural stabilities across single cells. The RMSD is calculated between the 3D chromatin structure of every single cell and the averaged structures of all cells. Lower RMSD represents the 3D location of DNA fragments that are stable across single cells. (**e**) Compartment A shows lower RMSD compared with compartment B. (**f**) Genes specifically expressed in GM12878 cells show lower RMSD compared with other genes.

resolution increases. These results clearly demonstrate the usability of tFLAMINGO in imputing single-cell chromatin contact maps at high resolution.

### 3.2.6 Single-cell compartment and TAD analyses of tFLAMINGO

Based on the bulk tissue Hi-C chromatin contact maps, the chromosomes are segregated into densely interconnected regions, i.e. compartments and topological associated domains (TADs), which delineate the outlines the chromosome structures in 3D space. Due to the high sparsity of the single-cell chromatin contact maps, discovering compartments and TADs for every cell is still challenging. Therefore, the completed high-resolution chromatin contact map imputed by tFLAMINGO provides a foundation to study the single-cell compartment and TAD structures. Based on the chromatin contact maps imputed by tFLAMINGO for 351 mESC single cells, single-cell compartments and TADs are called following the existing methodology. As Figure 3.6.a shows, chromosome 19 can be divided into two major compartments based on the bulk tissue Hi-C dataset in mESC. Interestingly, the same compartment structure is also observed in the pooled (average) single-cell chromatin contact maps, further verifying the accuracy of the predicted single-cell 3D chromosome structures. Moreover, all 351 mESC cells show

consistent distributions of the PC1 scores, suggesting the chromosome structures are highly stable across single cells at the compartment level. We also repeated the analyses in 14 GM12878 cells, and the similar distributions of the PC1 scores along the genome are observed across single cells and bulk tissue Hi-C data (Figure B.12).

To gain insights into chromatin structures across single cells at the TAD level, we further calculated the TAD boundary scores from the single-cell chromatin contact maps imputed by tFLAMINGO. As an example, Figure 3.6.b shows the distribution of the TAD boundary scores in a ~2MB genomic region. Eight regions with consistently high TAD boundary scores across all single cells are identified as TAD boundaries. Interestingly, seven out of eight TAD boundaries are intensively bound by CTCFs, which is consistent with the loop extrusion model. For the TAD boundary without CTCF binding, a CTCF motif (p-value < $1.2 \times 10^{-5}$) and B-Box regulatory element are observed in a SINE/B2 transposable element, which have been proved to shape the chromatin structures by serving as TAD boundaries. Quantitatively, DNA fragments with high TAD boundary scores show significantly higher CTCF binding intensity (p-value = $2.48 \times 10^{-5}$) and instances (p-value = $8.32 \times 10^{-4}$) compared with DNA fragments with medium and low TAD boundary scores (Figure 3.6.c and Figure 3.6.d). These results further highlight that the formation of the persistent TAD boundaries across single cells is mediated by CTCF binding, which is coherent with the loop-extrusion model.

We further explored the structural stabilities of 3D chromatin structures across single cells and their relationships with gene regulations in GM12878. RMSD between the single-cell 3D chromatin structures and the average structure across all cells is calculated along the

chromosome to quantify the structural stabilities at each genomic location. Interestingly, we found a differential distribution of the RMSD in compartment A/B. Compared with

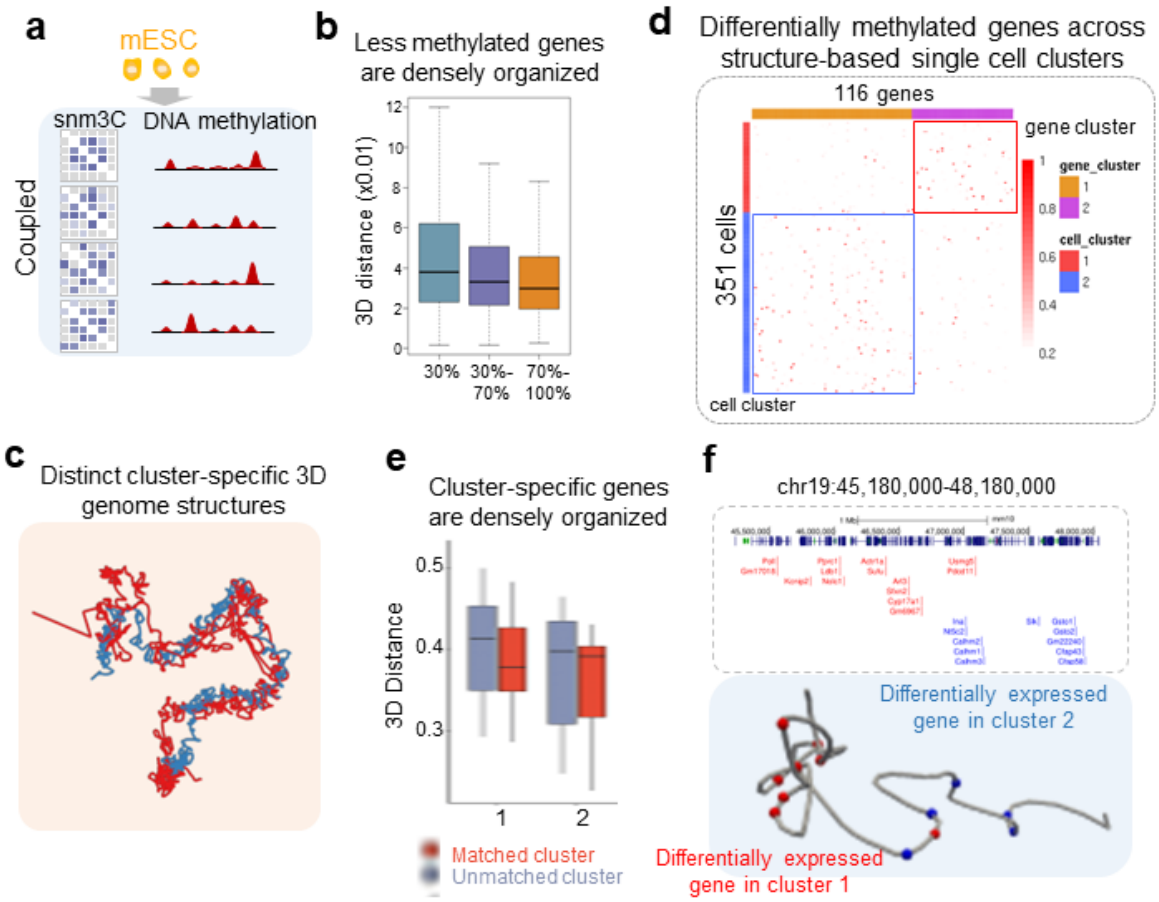



**Figure 3.7 Dynamic single-cell 3D chromosome structures reflects distinct methylation landscape of gene.** (**a**) In snm3C dataset, the contact map and coupled DNA Methylation are provided. (**b**) Less methylated genes show closer 3D distances across single cells. For every single cell, genes are divided into three groups based on the DNA methylation scores (<30% quantile, 30%-70% quantile, and >70% quantile). The pairwise distances between pairs of two genes within the same group are calculated across all single cells. The center lines of boxplots show the median, the upper and lower box limits show the 25th and 75th percentiles respectively. The whiskers extend up to 1.5 times the interquartile range away from the limits of the boxes. (**c**) 351 single cells are divided into two clusters based on the similarity of the predicted 3D chromatin structures. The consensus structures of the two cell clusters are visualized. (**d**) Identification of the differentially methylated genes across two cell clusters. (**e**) Distribution of the pairwise distances between the cluster-specific differentially methylated genes across cell clusters. Genes show shorter 3D distances in cells with lower DNA methylation scores. (**f**) Example of 3D chromatin structures facilitating the densely organized differentially expressed genes.

genomic regions in compartment B, compartment A shows significantly lower RMSD, implying more stable 3D chromosome structures across all single cells (Figure 3.6.e, p-value=$5.33 \times 10^{-5}$). This result suggests the open chromatin regions are more stable in the 3D space, probably because of their essential role in transcriptional regulations. In addition, the genomic regions harboring the genes specifically expressed in GM12878 show lower RMSD compared with other genes (p-value=$1.92 \times 10^{-3}$), further supporting the observation that the genomic regions with essential transcriptional and regulatory functions have more stable 3D structures across single cells. These results not only verify the accuracy of tFLAMINGO, but also provides new functional interpretations of the cell-to-cell structural variations.

### 3.2.7 Spatial analysis of gene activities in 3D space by tFLAMINGO

To further demonstrate the critical role of 3D chromosome structures in regulating gene expressions, we analyzed the spatial organizations of the differentially methylated genes using the snm3C dataset. Beyond profiling the single-cell chromatin contact maps, the snm3C dataset simultaneously measures the single-cell DNA methylation signals, which overlays the epigenomic information with the 3D structural information (Figure 3.7.a). The gene activity is quantified by the average DNA methylation signals within the promoter region for every cell. We divided protein-coding genes from chromosome 19 into three groups based on the strength of the DNA methylation signals and calculated the spatial distances between genes within each group based on the pooled 3D chromosome structures of 351 mESC cells. As Figure 3.7.b shows, genes with the lowest DNA methylation scores have shorter spatial distances, while genes with medium and high DNA methylation scores show relatively longer spatial distances. Considering the DNA

methylation signal of the promoter region is a counter metric of the gene activity, this result suggests the highly expressed genes are densely organized into 3D neighborhoods, which potentially enables the gene-gene regulations.

To further explain the variability of gene expressions across single cells in the 3D space, we clustered the 351 mESC cells based on the 3D chromosome structures and studied the spatial relationships between the differentially methylated genes. By projecting the distance matrices into the two-dimensional space using UMAP, two clusters of cells are identified with the maximized average silhouette width, where the first cluster contains 117 cells and the second cluster contains 234 cells. While these two clusters of cells show a similar backbone structure, local chromosome structures are extensively re-organized (Figure 3.7.c). By comparing the single-cell DNA methylation profiles across two clusters, 116 genes are identified as the cluster-specific differentially methylated genes, suggesting the distinct transcriptional landscapes in two cell clusters. As Figure 3.7.d shows, since 71 of 116 genes show strong DNA methylation signals in the second cell cluster but weak signals in the first cluster, they are considered to be specifically methylated in the second cluster. For the same reason, 45 genes are considered to be specifically methylated in the second cluster. To investigate the relative spatial localities of the differentially methylated genes, we calculated the pairwise distances between genes in two cell clusters based on two pooled cluster-specific structures. Strikingly, the differentially methylated genes exhibit shorter pairwise distances based on the matching 3D chromosome structures, while they are loosely scattered along with the unmatching 3D chromatin structure (Figure 3.7.e). As a representative example, Figure 3.7.f shows a 3MB genomic region, which contains 9 genes specifically methylated in the first cell

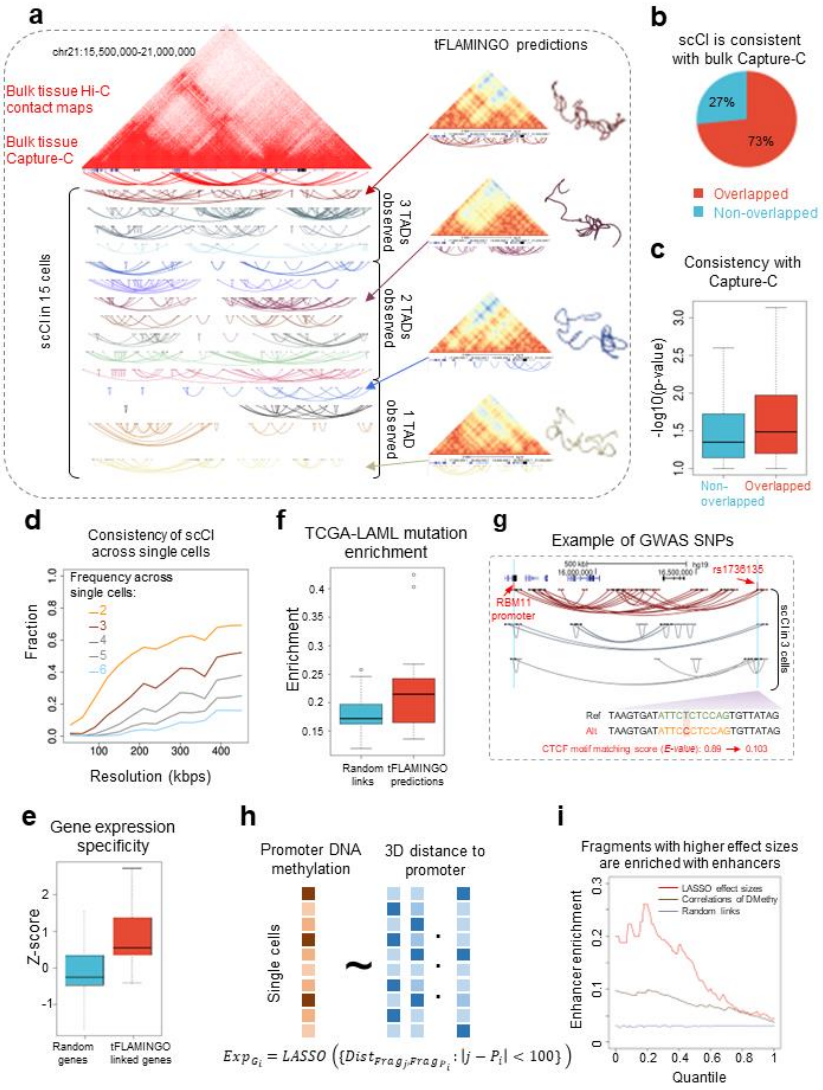



**Figure 3.8 Analyses of single-cell chromatin interactions.** (**a**) Predicted single-cell chromatin interactions across 15 GM12878 cells at 30kb-resolution. Altogether, the predicted single-cell chromatin interactions capture three TADs shown in the bulk Hi-C contact maps and Capture-C interactions. Across single cells, different 3D chromatin structures and associated single-cell chromatin interactions are observed, confirming the dynamicity of the 3D chromatin structures. Only statistically significant chromatin interactions are shown (p-value < 5x10-5). (**b**) Predicted single-cell chromatin interactions are strongly supported by bulk tissue Capture-C interactions. (**c**) Predicted single-cell chromatin interactions have a lower p-value in the bulk tissue Hi-C dataset, suggesting strong interactions are less dynamic across single cells. (**d**) Consistency of single-cell chromatin interactions across 15 cells under different resolutions. Under a certain resolution, fractions of the different number of cells containing a specific chromatin interaction are calculated. (**e**) Genes linked by the single-cell chromatin interactions have higher expression values in GM12878. (**f**) Single-cell chromatin interactions are enriched with TCGA-LAML somatic mutations compared with distance-controlled random interactions. (**g**) Example of GWAS SNP captured by single-cell chromatin interactions.

**Figure 3.8 (cont'd)**

rs1736135 is linked with gene RBM11 through chromatin interactions in three single cells. Potentially, rs1736135 controls the expression of RBM11 gene expression through chromatin interactions by creating a CTCF motif, thus associated with Crohn's disease. (**h**) Schematic of predicting the functional chromatin interactions based on the tFLAMINGO predicted contact maps and coupled DNA methylation scores using mESC snm3C data. LASSO regression is used to select the DNA fragments whose 3D distances to the gene promoter can best predict the DNA methylation scores of the target gene. The longest distances between DNA fragments and target gene promoters are limited to 3MB. (**i**) Enrichment of enhancers along effect sizes predicted by LASSO. As comparisons, the correlations between the DNA fragments and target gene promoters across single cells are used as the predictive score for enhancer enrichment analysis. The result of distance-controlled random chromatin interactions is also shown.

cluster and 14 genes for the second cell cluster. Interestingly, the two groups of genes are located in two distinct 3D neighbors and interact with each other through 3D chromatin loops. Apart from the 3D proximity of the differentially methylated genes, we further confirmed that the two sets of genes are enriched in different biological pathways, suggesting their unique roles in cell development at different stages. These results further demonstrate the biological utilities of tFLAMINGO in interpreting the dynamic activity of genes across single cells.

### 3.2.8 Dynamic single-cell chromatin interaction landscape identified by tFLAMINGO

As a direct contribution of tFLAMINGO, the predicted high-resolution 3D chromosome structures, as well as the chromatin contact maps, fully characterize the interaction landscape at 10kb resolution and facilitate the study of single-cell chromatin interactions. Therefore, we identified chromatin interactions based on the predicted chromatin contact maps of tFLAMINGO for 15 GM12878 cells in 30kb resolution (Figure B.15). As a representative example, Figure 3.8.a shows the single-cell chromatin interactions in a

5.5MB genomic region (chr21:15,500,000-21,000,000). In this region, the combined single-cell chromatin interactions form three TADs, consistent with the bulk Hi-C and Capture-C data. Surprisingly, chromatin interaction landscapes are drastically changed across single cells, and TADs are observed to be shifting, merging, and vanishing. For example, we observed three TADs from single-cell 1. The imputed single-cell chromatin contact maps accurately capture the bulk chromatin contact maps, and three compact domains are observed on the 3D structure. However, for single-cell 7, the second and third TAD from the bulk tissue Hi-C contact maps are merged into a larger compact domain, and only two TADs persist. In single-cell 12 and single-cell 15, the chromatin loops are untangled, resulting in the vanishment of two TADs. As a functional validation of the predicted single-cell chromatin interactions, we evaluated whether the genes linked by the single-cell chromatin interactions are specifically and highly expressed in GM12878. We found that the linked genes have high Z-scores of gene expression in GM12878, comparing with the randomly selected genes (Figure 3.8.e). This result highlights the regulatory effect of the predicted single-cell chromatin interactions. The systematic comparisons of TAD structures across single cells and bulk chromatin contact maps, along with the analysis of the gene expression specificity, unveil the dynamic chromosome structures in an unprecedented resolution, which cannot be observed in bulk tissue chromatin contact maps or the low-resolution single-cell chromatin contact maps.

### 3.2.9 Relationship between single-cell chromatin interactions and bulk Capture-C interactions

We further leverage the orthogonal Capture-C dataset to evaluate the consistency between chromatin interactions in bulk tissue and single cells. By overlapping the predicted single-cell chromatin interactions and Capture-C interactions, we found that the Capture-C dataset captures 73% of the single-cell chromatin interactions (Figure 3.8.b). On the other hand, the Capture-C interactions that overlap with the single-cell chromatin interactions tend to have lower p-values (p-value = 3.47x10-3), suggesting stronger chromatin interactions are conserved across single cells (Figure 3.8.c). In fact, a large fraction of chromatin interactions are shared across different cells. As shown in Figure 3.8.d, the fraction of single-cell chromatin interactions shared by different numbers of cells is calculated as the resolution increases. At 250 kb resolution, over 50% of the single-cell chromatin interactions are captured in two cells, and around 40% of single-cell chromatin interactions are shared by three cells (Figure 3.8.d). These analyses strongly suggest the bulk tissue chromatin contact maps only reflect the average of million cells and have no dynamicity. Therefore, the development of tFLAMINGO can largely boost the understanding of the dynamic chromatin interactions at the single-cell level.

### 3.2.10 Interpreting genetic variants based on single cell chromatin interactions

The 3D chromatin structures and chromatin contact maps predicted by tFLAMINGO provide the structural basis of the GWAS SNPs and disease-associated somatic mutations. Firstly, we used the single-cell chromatin interactions to interpret the LAML-associated somatic mutations. We overlapped the LAML-somatic mutations with the interacting anchors of the chromatin interactions and calculated the enrichment of somatic

mutations. Compared with the random chromatin interactions with genomic distances controlled, the single-cell chromatin interactions predicted by tFLAMINGO show a higher enrichment of somatic mutations (Figure 3.8.f), suggesting the disease-SNP associations are mediated by the chromatin interactions. Figure 3.8.g shows one representative example, where the SNP rs1736135 is associated with the Crohn's disease. Based on the predicted single-cell chromatin interactions, this SNP is linked to the promoter of an oncogene RBM11, whose overexpression can significantly decrease the survival rate of the patients. Interestingly, the SNP rs1736135 creates a CTCF motif by transiting a T to C in the alternate genome (E-value: 0.89 to 0.103), which potentially established the chromatin interactions with the RBM11 promoter and finally contributed to the Crohn's disease. This evidence further confirms the regulatory function of the single-cell chromatin interactions and provides a new approach to understanding the disease-associated genetic variants mechanistically.

### 3.2.11 Predicting functional gene regulatory links in single cells

In addition to the single-cell chromatin interactions, we further predict the functional regulatory links for gene expression. In bulk tissue, the regulatory elements are computationally linked to the promoter of genes based on the 1D genomic distances and co-activity patterns. However, these methods model the chromosomes as 1D strings and leave out the important 3D chromosome structures. According to the phase separation model, the gene expressions are controlled by the dynamic binding and unbinding events between genes and regulatory elements on the 3D chromosome structures. This transient binding process can not be explained by the static bulk tissue datasets but can be captured by spatial distances between the DNA fragments across single cells. Therefore,

single-cell 3D chromosome structures predicted by tFLAMINGO can help to predict the transcriptional regulations between genes and regulatory elements. Specifically, we linked the target genes and DNA fragments, whose close spatial distances to the target genes are associated with the high gene activities across single cells, based on the predicted single-cell 3D chromosome structures. The DNA methylation signals of every gene are considered to have a linear relationship with the spatial distances of all DNA fragments with 1D distance smaller than 3MB across all single cells. LASSO model is

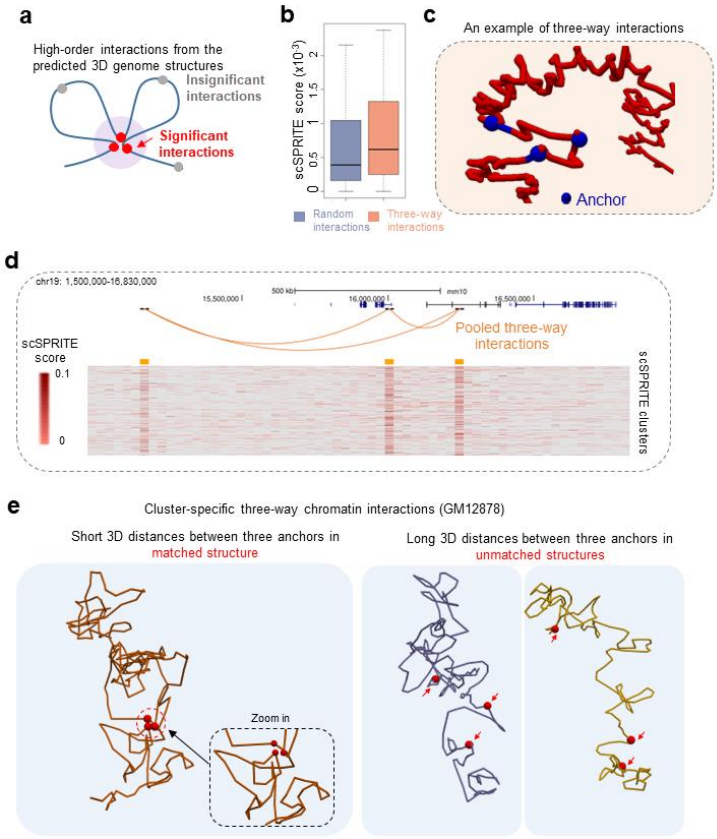



**Figure 3.9 Identification of the single-cell multi-way chromatin interactions based on the predicted chromosome structures.** (**a**) The three-way chromatin interactions are predicted from the 3D chromatin structures by evaluating the pairwise 3D distances. (**b**) Predicted three-way chromatin interactions are supported by scSPRITE data (n=3408, p-value<$1.4 \times 10^{-5}$). Higher scSPRITE scores represent the three-way interactions are observed in more single cells from scSPRITE dataset. (**c-d**) Example of (**c**) 3D chromatin

**Figure 3.9 (cont'd)**

structure of predicted three-way interactions and (**d**) scSPRITE scores. (**e**) Example of the 3D chromatin structure of cluster-specific three-way interactions.

used to prioritize the most influential DNA fragments, i.e. master regulators, of the gene activities (Figure 3.8.h). DNA fragments with larger effect sizes in LASSO are considered to have more substantial regulatory effects on the target gene expressions, as the smaller 3D distances are associated with lower DNA methylation signals of promoters and higher gene activities. The orthogonal enhancer annotation dataset in mESC is used to evaluate the regulatory effect of the linked DNA fragments. We calculated the enhancer enrichment among the predicted DNA fragment-gene links at different effect size cut-offs and used the enrichment to quantify the accuracy of predicted links. As shown in Figure 3.8.i, DNA fragments with high effect sizes are enriched with enhancers, suggesting the regulatory effects of the highly ranked DNA fragments. As comparison, the DNA fragments are also linked to the genes based on the co-activities and 1D genomic distances. In comparison, the predictions of tFLAMINGO demonstrate consistently higher enhancer enrichment across all cut-offs. These analyses clearly demonstrate that, overlaid with the gene activities, the ability of tFLAMINGO to accurately predict the functional chromatin interactions.

### 3.2.12 Analysis of single-cell multi-way interactions by tFLAMINGO

The high-order genome organization enables the multi-way interactions between DNA fragments, which play an important role in gene regulations. Beyond the 1D genomic distances and 2D chromatin contact maps, the single-cell 3D chromosome structures predicted by tFLAMINGO enable directly probing the spatial distances between multiple

DNA fragments and predicting the multi-way chromatin interactions from the 3D space. We depicted the three-way interactions by identifying sets of three closely allocated DNA fragments on the pooled 3D chromosome structure of 351 mESC cells. FOr each set of three DNA fragments, we calculated the average pairwise distances to quantify the compactness of the three-way interactions. To evaluate the statistical significance of the three-way interactions, 1000 sets of randomly selected DNA fragments were generated with the genomic distance controlled and their average pairwise distances were used to calculate the empirical p-values. Overall, we predicted 973 statistically significant three-way interactions (p-value < 0.05). To evaluate the predicted three-way interactions, we overlapped the scSPRITE clusters to the predicted three-way interactions and used the normalized counts of the overlapping scSPRITE clusters to quantify the accuracy (termed as the 'scSPRITE score' hereafter). We observed that the predicted three-way interactions show significantly higher scSPRITE scores than random three-way interactions with 1D genomic distances controlled. Figure 3.9.c shows an example of the predicted three-way interaction, where three anchors are brought to the same 3D neighborhood by a chromatin loop. Interestingly, the three-way interaction is also frequently observed in the scSPRITE cluster, validating the accuracy of the predicted three-way interactions.

To further understand the high-order organization of chromosomes across single cells, we divided 15 GM12878 cells into three clusters based on similar chromosome structures and predicted three-way interactions in each cluster. The three-way interactions are predicted based on the pooled structure of every cluster. We identified around 1000 three-way chromosome interactions in each cluster, suggesting that chromosomes display

dynamic high-order structures across single cells. As shown in Figure 3.9.e, the anchors of the predicted cluster-specific three-way interactions show short spatial distances on the matched 3D chromosome structure. However, they are far from each other on the two unmatched 3D chromosome structures, suggesting the complex high-order structures across single cells. Given these analyses, the single-cell 3D chromosome structures predicted by tFLAMINGO reveal the high-order chromosome conformation and innovate the study of multi-way chromosome interactions.

## 3.3 DISCUSSION

In this work, we developed tFLAMINGO to reconstruct the single-cell 3D chromosome structures at high resolution from the sparse single-cell chromatin contact maps. Equipped with the low-rank tensor completion method, tFLAMINGO mitigates the high missing rates of the single-cell chromatin contact maps by borrowing information from all contacts in all cells. The application of tFLAMINGO on four single-cell chromatin conformation capture datasets provides a rich resource of single-cell 3D chromosome structures at 10kb and 30kb resolution. Based on the extensive performance evaluations, tFLAMINGO achieves superior accuracy in reconstructing the single-cell 3D chromosome structures, imputing single-cell chromatin contact maps over the existing state-of-art methods, and capturing the cell-type-specific structural variations. The high consistency between the experimental super-resolution imaging data and tFLAMINGO predictions further confirms the accuracy of tFLAMINGO in predicting the highly dynamic single-cell 3Dchromosome structures. Biologically, tFLAMINGO confirms the robust compartment and TAD structures across single cells. Coupled with the DNA methylation dataset, tFLAMINGO unveils the interplay between dynamic gene regulations and 3D

chromosome structures across single cells. The detailed delineation of the high-resolution single-cell 3D chromosome structures by tFLAMINGO facilitates prediction of the single-cell chromatin interactions and provides mechanistic interpretations of GWAS SNPs and somatic mutations. Beyond the 2D chromatin contact maps, the characterization of the chromosome structure in the 3D space further enables the predictions of the high-order chromatin organizations and multi-way chromatin interactions.

Compared with existing methods, tFLAMINGO enjoys three unique advantages by modeling the low-rank structure of the single-cell chromatin contact maps: (1) substantially improved ability in handling high missing rate of the single-cell 3C datasets; (2) superior accuracy in predicting single-cell 3D chromosome structures and (3) robust performance in imputing the cell-type-specific high-resolution chromatin contact maps. Equipped with all these advantages, tFLAMINGO is designed for the single-cell 3C datasets and can aid the biological identification and interpretation of the cell-to-cell structural variations, differential single-cell gene expression, single-cell chromatin interactions, and the genetic variants.

As a data-driven model, tFLAMINGO solely relies on the input single-cell chromatin interactions and does not introduce any bias into the reconstruction. This feature is crucial for reconstructing the high-resolution single-cell 3D chromosome structures, as the information from the highly sparse single-cell chromatin contact maps poses fewer constraints on the predicted structures compared with the prior assumptions. As another important class of methods, the constrained polymer simulation-based model relies on both pre-determined biophysics properties of the DNA sequences and the observed single-cell chromatin contact maps to predict the 3D chromosome structures. This

strategy successfully reconstructs the low-resolution chromosome structures (i.e. 1MB) when most of the DNA fragments are constrained in the relatively dense low-resolution single-cell chromatin contact maps. However, the predictive accuracy in high resolution (i.e. 10kb) is drastically decreased for two reasons. Firstly, the high-resolution chromatin contact maps are incredibly sparse, and the simulation process is dominated by the pre-defined biophysics property, which is invariant across single cells and cell types and may contradict the observed chromatin contact maps. In this case, the model cannot find an optimal structure to satisfy both constraints, thus deviating from the observed values. Therefore, the simulated 3D chromosome structures cannot accurately capture the high-resolution structural variations across single cells. Secondly, the simulation-based methods tend to predict the chromosome structures as extended smooth strings, which violates the observed long-range chromatin interactions. For example, isdHi-C demonstrates a high accuracy (Figure B.16, correlation: 0.73) on a genomic region with 22% of long-range chromatin interactions (>200kb) but failed on the adjacent genomic region with 42% of long-range interactions (Figure B.16, correlation 0.18). Further simulation analyses confirm the limitation of isdHiC in predicting the condensed ball-type structures with massive long-range interactions (Figure B.17). Therefore, the simulation-based methods cannot accurately reconstruct the 3D structures from chromatin contact maps with lots of long-range interactions. In comparison, since no prior assumptions of the chromosome structures are made, tFLAMINGO demonstrates robust performance in reconstructing 3D structures with different geometrical patterns across different resolutions, which implies a high accuracy in reconstructing the dynamic 3D spatial structures.

We envision two future developments of tFLAMINGO. First, the information-sharing mechanism of tFLAMINGO requires that all cells are from the same cell type and share a similar backbone structure. In the current framework of tFLAMINGO, all cells are equally important in the Fourier transformation, and the low-rank features shared by all cells will be extracted for the reconstructions. This assumption can be satisfied when the cell-type identities of single-cell chromatin contact maps are provided, or the dataset only contains cells from one cell type. However, some highly complex tissue may contain cells from multiple cell types with unknown cell-type identities. For example, the human brain tissue consists of several highly differentiated cell types, and the cell-type deconvolution is challenging. In this case, the consensus structure of the dataset is essentially a mixture of multiple cell-type-specific chromosome structures, and the assumption of tFLAMINGO is not fulfilled. Although tFLAMINGO demonstrates superior performance on the simulated datasets with three similar consensus structures, it is still challenging to accommodate datasets with multiple fundamentally re-wired structures. Therefore, additional algorithmic improvement of tFLAMINGO is required to simultaneously de-convolve the single-cell 3D chromatin contact maps and reconstruct the single-cell 3D chromosome structures with the cell-type specificity retained. Secondly, tFLAMINGO can be improved to reconstruct the time-dependent single-cell 3D chromosome structures. Based on the recent single-cell RNA-seq data analysis, the differential gene expression patterns are observed at the different stages of cell differentiation along the lineage trajectory, suggesting the changing gene regulations and 3D chromosome structures. Reconstructing the time-dependent single-cell 3D chromosome structures will open new avenues to understand the dynamic gene expression during the cell cycle, cell

differentiation, and cellular activation from the 3D space. Currently, the experimental assessment of the lineage-specific single-cell 3D chromatin conformation is still challenging, and the computational methods are highly applaudable. tFLAMINGO has inherent algorithmic advantages in modeling the ordered single-cell chromatin contact maps in two aspects. Firstly, tFLAMINGO models all single-cell chromatin contact maps as the frontal slices of a low-rank tensor. The order of the frontal slices can be easily extended to incorporate the lineage information of single cells by assigning cells based on the time order. Secondly, tFLAMINGO demonstrates superior ability in preserving the cell-type-specific structures based on the simulation analyses. This critical advantage of tFLAMINGO guarantees that the subtle changes in 3D chromosome structures can be accurately captured. Therefore, the further development of tFLAMINGO can not only capture the structural variations across single cells but also recapitulate the time-dependent structural properties.

## 3.4 METHODS

### 3.4.1 Model framework of tFLAMINGO

tFLAMINGO reconstructs single-cell 3D chromatin based on the low-rank tensor completion method. In the framework of tFLAMINGO, the missing values of the sparse tensor summarizing all single-cell chromatin contact maps is firstly completed and the underlying 3D structures are predicted for every cell. This framework brings two algorithmic advancements: (1) tFLAMINGO jointly models the chromatin contact maps of all single cells. This ensures the information could be borrowed across single cells; (2) tFLAMINGO makes full use of the low-rank property of the single cell chromatin contact maps. This property guarantees the underlying 3D chromatin structures can be accurately

recovered from the sparse chromatin contact maps under high missing rates. Computationally, we solved a tensor rank-minimization problem using the ADMM method to complete the missing values and used our in-house 3D reconstruction algorithm FLAMINGO to reconstruct the 3D structures.

### 3.4.2 Chromatin contact maps and data preprocessing

Chromatin contact maps from four single-cell 3C studies are collected, including the Dip-C experiment in GM12878, the snHi-C experiment in K562, the snm3C experiment in mESC and scHi-C experiment in mESC (see Data availability). Although the interaction frequencies from single-cell 3C datasets show strong agreement with the bulk tissue dataset, the single cell chromatin contact maps tend to have much smaller interaction frequencies at high resolution, thus cannot be directly converted to the 3D distances between DNA fragments using the conversion transformation function observed from the bulk-tissue Hi-C data. More importantly, different linear relationships are observed for interaction frequencies with different 1D genomic distances, suggesting the potential confounding effect of the 1D distances. Based on this observation, tFLAMINGO maps the single-cell chromatin contact maps to the same scale as the bulk Hi-C contact maps using the band-wise log-linear regression. Interaction frequencies between DNA fragments with similar 1D genomic distances are jointly modeled, which can be represented as a diagonal band on the interaction frequency matrix. Apart from interaction frequencies, 1D genomic distances between interacting DNA fragments, missing rate of single cells and expected interaction frequencies between interacting DNA fragments are considered as covariates:

$$log\left(IF_{i,j}^{bulk}\right) = \alpha_l * log\left(IF_{k;i,j}^{sc}\right) + \beta_l * log\left(Dist_{i,j}\right) + \theta_l * MR_k + \gamma_l * log\left(IF_{k;i,i}^{sc} * IF_{k;j,j}^{sc}\right),$$

( 11 )

where $IF_{i,j}^{bulk}$ represents the interaction frequency between $ith$ DNA fragment and $jth$ DNA fragment in the bulk-tissue Hi-C contact map, $IF_{k;i,j}^{sc}$ represents the interaction frequency between $ith$ DNA fragment and $jth$ DNA fragment in the contact map of $kth$ single cell, $Dist_{i,j}$ represents the 1D genomic distance between $ith$ and $jth$ DNA fragment, and $MR_k$ represents the missing rate of $kth$ single cell contact map. To account for the different log-linear relationships under different distance ranges, the regression parameters are estimated in every distance band as suggested by previous studies. We applied the estimated log-linear transformation functions on different single cell 3C datasets across different resolution and observed high correlations between the transformed single-cell interaction frequencies and observed bulk Hi-C interaction frequencies, validating the robustness and generalizability of the estimated transformation functions.

The bulk-tissue chromatin contact maps generated by four studies are collected from GEO and 4DN databases, including bulk-tissue Hi-C experiments in GM12878 and K562 (GSE63525), GAM experiment in mESC (GSE64881), 3D ATAC-PALM experiment in mESC (GSE126112) and bulk-tissue Hi-C experiment in mESC (4DNFI5IAH9H1). The GAM data only provide chromatin contact maps at 30kb resolution. Except the GAM data, all bulk-tissue chromatin contact maps are used to validate the performance in 10kb and 30kb resolution. Whenever possible, the chromatin contact maps normalized by the Knight-Ruiz normalization are used. The scSPRITE data is collected from the GEO database (GSE154353) and preprocessed according to the instructions.

### 3.4.3 Complete single-cell chromatin contact maps based on the low-rank tensor completion

The missing rate of the single-cell contact map is high (>99.9% at 30kb resolution), making the reconstruction of the 3D chromatin structures in high resolution extremely challenging. tFLAMINGO mitigates the high missing rate by borrowing information in two directions: (1) the same contact of two DNA fragments across all contact maps and (2) all contacts between DNA fragments within the same contact maps. Biologically, every sparse single-cell contact map represents a randomly down-sampled ‘snapshot’ of the consensus 3D chromatin structure with structural variations. Therefore, the missing entry in one single cell contact map can be imputed by borrowing information from the same entry measured in other single cell contact maps. tFLAMINGO facilitates the information-sharing across all single cells using a Fourier transformation-based method. To borrow information from contacts within the same contact map, tFLAMINGO takes advantage of the low-rank property of the single-cell contact map. According to Euclidean geometry, the distance matrix derived from the single cell contact map is induced by a 3D coordinate matrix, thus having the low-rank property (rank$\leq$ 5). The low-rank property guarantees that the missing values can be reconstructed from a small fraction of observed values. Equipped with an SVD-based method, tFLAMINGO borrows information across all contacts within the same single cell contact map.

Computationally, the single cell contact maps are summarized into a tensor, where each frontal slice represents a single cell contact map and a tube that perpendicular to the plane of paper represents a contact between a pair of DNA fragments across all single cells. tFLAMINGO aims to recover a dense tensor with minimum error compared with the

sparse input tensor on the observed entries using a t-SVD-based method. The t-SVD method has been widely used to identify the low-rank structures of high-dimensional tensor. Similar to the matrix SVD, t-SVD decomposes the tensor into the multiplication of three tensors: $T^{obs} = U * S * V^T$, where $*$ represents the circular convolution product (t-product) of tensors. According to the tensor-completion theory, the tensor completion problem can be solved by calculating the matrix SVD across all frontal slices of the tensor in the Fourier domain. The observed tensor $T^{obs}$ is transformed into the Fourier domain using a tube-wise Fourier Transformation:

$$\hat{T}^{obs}_{i,j,k} = \sum_{n=1}^{N} T^{obs}_{i,j,n} * e^{-2\pi ikn/N}. \quad (12)$$

Intuitively, the contact between DNA fragment $i$ and DNA fragment $j$ for single cell $k$ in the Fourier domain ($\hat{T}^{obs}_{i,j,k}$) are calculated from the same contact across all single cells ($T^{obs}_{i,j,n}\ for\ all\ n$). Therefore, if any single cell chromatin contact map contains observed values for the contact $(i, j)$, all values in the tube $\hat{T}_{i,j,:}$ will be completed in the Fourier domain by aggregating the observations of all cells. Given the tensor $\hat{T}^{obs}$ in the Fourier domain, the SVD is applied on every fontal slice of $\hat{T}^{obs}$ ( $\hat{T}^{obs}_{:,:,k}$): $\hat{T}^{dense}_{:,:,k} = U^{obs}_k * S^{obs}_k * (V_k{}^{obs})^T$. The SVD procedure captures the low-rank structures of the frontal slices and borrows information across all contacts within each cell. The recovered tensor is then transformed into the original domain using the inverse Fourier Transformation, and the resulting tensor can maximally approximate the input one.

For single-cell chromatin contact maps, the high missing rate of the observed tensor $T^{obs}$ requires the completion process only relies on a few observed entries. In tFLAMINGO, the objective function of the low-rank tensor reconstruction is:

$$\left|\left|X\right|\right|_{TNN},\ s.t.\ \Omega(T^{obs}) = \Omega(X),$$

( 13 )

where $T^{obs}$ represents the sparse tensor summarizing all single cell chromatin contact maps, $X$ represents the recovered dense tensor, $\Omega$ represents the set of observed entries in $T^{obs}$ and $TNN$ represents the Tensor Nuclear Norm. To achieve fast and accurate convergence, tFLAMINGO simplifies the optimization problem by solving the equivalent optimization problem in the Fourier domain:

$$\left|\left|blkdiag\left(\hat{X}\right)\right|\right|_{*}\ s.t.\Omega\left(\hat{X}\right) = \Omega(\hat{T}),$$

( 14 )

where $\hat{X}$ represents the transformed tensor in the Fourier domain, $blkdiag$ represents the block diagonal matrix constructed by placing the frontal slices of the tensor $X$ into diagonal submatrices of a large matrix, $*$ represents the matrix nuclear norm. tFLAMINGO uses the Alternating Direction Method of Multipliers (ADMM) algorithm to solve the optimization problem and the original objective function can be re-written as:

$$\left|blkdiag\left(\hat{Z}\right)\right|_{*} + 1_{\Omega(\hat{T})=\Omega(\hat{X})}\ s.t.\hat{X} - \hat{Z} = 0,$$

( 15 )

where $Z$ is introduced as an intermediate variable. The iterative updating scheme can be derived as:

$$X^{t+1} = argmin_{X:\Omega(X)=\Omega(T^{obs})}\{\left||X - (Z^t - Q^t)|\right|_F^2\},$$

$$Z^{t+1} = argmin_Z\{\frac{1}{\rho}\left|\left|blkdiag(\hat{Z})\right|\right|_* + \frac{1}{2}\left|\left|\hat{Z} - (X^t + Q^t)\right|\right|_F^2\},$$

$$Q^t = Q^{t-1} + X^t - Z^t,$$

( 16 )

where $\rho$ is a free parameter. In tFLAMINGO, $\rho$ is set to 1 by default according to previous analysis. $X^{t+1}$ can be analytically solved as :

$$X_{i,j,k}^{t+1} = \begin{cases} Z_{i,j,k}^t - Q_{i,j,k}^t & (i,j,k) \notin \Omega_{T_{i,j,k}^{obs}} \\ T_{i,j,k}^{obs} & (i,j,k) \in \Omega \end{cases}$$

$Z^{t+1}$ can be solved by applying the soft-thresholded t-SVD method on $X^t + Q^t$. Through iterations, $Z$ borrows information across all contacts across all single cells using the t-SVD method and $X$ guarantees the imputed values are close to the observed values on the measurement set. Upon convergence, tFLAMINGO imputes a much denser contact map for every single cell which maximally aligns with the observed single cell chromatin contact map.

### 3.4.5 Reconstruct the single cell 3D chromatin structure based on low-rank matrix completion

Given the single-cell chromatin contact maps imputed by the low-rank tensor completion algorithm, tFLAMINGO reconstruct the 3D chromosome structures. The chromatin contact maps are converted to pairwise distance matrices using the observed conversion function: $IF_{i,j} = PD_{ij}^{-\alpha}$, where $\alpha$ is set to 0.25 based on the previous studies. Our in-house 3D chromatin reconstruction algorithm, FLAMINGO, is used to reconstruct the high-

resolution 3D chromatin structures for every single cell. Algorithmically, FLAMINGO reconstructs the 3D chromosome structures based on the low-rank matrix completion technique, which guarantees an accurate reconstruction of the 3D coordinate matrices from highly noisy and sparse chromatin contact maps. Compared with existing methods, FLAMINGO demonstrates superior accuracy and scalability in reconstructing high-resolution chromatin structures (up to 1kb) from extremely sparse chromatin contact maps (missing rate >99%).

### 3.4.6 Performance evaluation based on simulated chromatin structures

The performance of tFLAMINGO is extensively evaluated by reconstructing the simulated benchmark structures. In the simulation, a benchmark structure with $l$ beads is generated. The $l$ by $l$ benchmark distance matrix induced by the benchmark structure is down-sampled $n$ times with the down-sampling rate $\gamma$ and mixed with three levels of noise: (1) no noise, (2) noise level one, which is generated by the normal distribution $N(\delta, \delta)$, where $\delta$ is the minimum value of the down-sampled matrix, and (3) noise level two, which is generated by the normal distribution $N(2\delta, \delta)$. Thus, a sparse tensor is constructed to simulate the single-cell 3C dataset. tFLAMINGO is applied on the sparse tensor to reconstruct the benchmark 3D structures. The model performance is quantified by two metrics: (1) Spearman correlations between the pairwise distance matrices predicted by tFLAMINGO and benchmark pairwise distance matrices and (2) the RMSD between the predicted 3D coordinates ( $C^{pred}$ ) and benchmark 3D coordinates ($C^{benchmark}$): $RMSD = \sqrt{\frac{1}{n}\sum_{i=1}^{n} ||C_i^{pred} - C_i^{benchmark}||}$ . To demonstrate the robustness of tFLAMINGO, the simulated datasets are generated under different combinations of $l, n$ and $\gamma$.

Apart from the white noise, we further evaluated the performance of tFLAMINGO on inputs with noise generated from random structures. Given a benchmark structure, $n$ random structures are generated and corresponding pairwise distance matrices are mixed with the benchmark pairwise distance with weight $W$: $D = D^{benchmark} * (1 - W) + D^{random} * W$. The resulting noisy pairwise distance matrices are used as the input of tFLAMINGO to recover the benchmark structure. Compared with the white noise, the structured noise follows a similar pattern to the benchmark distances, i.e. consecutive points show lower distances, thus making the reconstruction of the benchmark structure more challenging.

Further, tFLAMINGO demonstrates excellent performance on datasets containing multiple consensus structures. Biologically, single-cell 3C data may contain cells in different developmental stages, which share similar backbone structures with structural variations. In the simulation, three consensus structures are generated with the weight $W$ to illustrate the heterogeneity of the single-cell 3D chromosome structures: $S_i = S_{starting} * W + S_{random} * (1 - W)$, where weight $W$ controls the similarity of the consensus structures. A sparse tensor with $3N$ frontal slices is further constructed by down-sampling the distance matrix induced by each consensus structure $N$ times and mixed with noise and used as the input of tFLAMINGO to reconstruct the underlying 3D structures. tFLAMINGO is applied on the sparse tensor to recover the 3D structures. A wide range of weight is tested in the simulation and the performance of tFLAMINGO is evaluated by the correlations of pairwise distances and RMSD of 3D coordinates.

### 3.4.7 Performance comparison based on the STORM 3D genome imaging data

We applied tFLAMINGO on K562 snHi-C data to reconstruct the 3D chromatin structures of chromosome 21 for 16 single cells. To evaluate the model performance, we benchmarked the predicted single cell 3D chromatin structures with the STORM 3D genome imaging data. As the data quality control, single-cell structures measured by STORM data with missing rates greater than 0.5 are excluded from the analysis. For each pair of the predicted structures and STORM structures, the Spearman correlations of spatial distances are calculated to evaluate the consistency. To approximate the underlying 3D structures of the snHi-C dataset, we calculated the Spearman correlations between the raw single-cell chromatin contact maps and STORM structures. The top 20 STORM structures with the highest Spearman correlations with each single-cell chromatin contact maps are considered to be the true underlying structures. Therefore, the Spearman correlations between the predicted structures and approximated ground-truth measure the reconstruction accuracy. Note that, these two metrics are complimentary, since the first Spearman correlation evaluates the consistency between the predictions and STORM data, and the second Spearman correlation validates the accuracy of reconstructing 3D structures of the snHi-C data. Therefore, a better algorithm is expected to achieve high values in both correlations.

### 3.4.8 Performance comparison in reconstructing 3D chromatin structures based on experimental single cell Hi-C data

We applied tFLAMINGO on four single cell Hi-C datasets (human GM12878 Dip-C and human K562 snHi-C, mESC scHi-C and mESC snm3C) to reconstruct the single-cell 3D chromatin structures (chr21 in human GM12878 and K562, chr19 in mESC) in 10kb and

30kb resolution. To evaluate the accuracy of completed missing data, we calculated the Spearman correlations based on all available entries from the bulk chromatin contact maps (termed as ‘all distance correlations’). The ability of recovering the observed values of the single-cell 3C datasets is further quantified by the Spearman correlations based on the observed distances from the single-cell 3C dataset (termed as ‘validated distance correlations’). Therefore, an accurately reconstructed single-cell 3D chromosome structures should demonstrate high all distance correlation as well as validated distance correlation.

The performance of tFLAMINGO is compared with six existing algorithms in reconstructing 3D chromatin structures: ShRec3D, NucDynamics, RPR, isdHi-C and Si-C. Another existing algorithm, MBO, is not included into the comparison due to no code availability. Predicted chromatin structures of NucDynamics based on mESC scHi-C data are directly used. ShRec3D, RPR, isdHi-C and Si-C are applied on the same dataset as tFLAMINGO in 10kb and 30kb resolution to predict the single cell 3D chromatin structures. To systematically compare the performance of tFLAMINGO with existing methods, we used two sets of experimental datasets as gold-standards: (1) the single cell 3D chromatin structures provided by the multiplexed STORM 3D genome imaging data in human K562 and (2) bulk tissue chromatin contact maps in human GM12878, human K562 and mESC.

We further evaluated cell-type specificity of the predicted 3D chromatin structures by different algorithms. All algorithms are applied on the snm3C data with 351 mESC single cells and 96 NMuMG cells to reconstruct the single-cell 3D chromatin structures. UMAP

plots based on the distance matrices predicted by different algorithms are used to visualize the clusters of single cells and evaluate the performance.

### 3.4.9 Performance comparison with Higashi in imputing high-resolution single cell chromatin contact maps

We benchmarked tFLAMINGO and Higashi on the snm3C dataset in 1mb, 250kb and 30kb resolution. For both algorithms, the original cell types of the single cells are provided as the inputs. The model performance is evaluated by: (1) the correlations with the observed distances in bulk-tissue chromatin contact maps and (2) the ability to recover cell-type identity based on the imputed chromatin contact maps. The Adjusted Random Index (ARI) is used to quantify the consistency between the predicted single cell clusters and the original cell-type identify.

### 3.4.10 Identification of the single-cell compartment A/B and TAD boundaries

The single-cell interaction frequency matrices are derived from the single cell distance matrices completed by tFLAMINGO using the conversion function $IF_{ij} = PD_{ij}^{-4}$ (corresponding to the conversion from interaction frequencies to distances with the conversion factor -0.25). The expected interaction frequency matrices then normalize the observed interaction frequency matrices following the standard procedure in previous studies. PC1 scores calculated from the normalized interaction frequency matrices are used to represent the compartment A/B. The single-cell TAD boundaries are called based on the single-cell interaction frequency matrices using the TADCompare software. The structural stability of the 3D chromatin structures is quantified by calculating the RMSD between the average structure from tFLAMINGO predictions and the single-cell 3D

chromatin structures along the genome: $RMSD_i^k$=|$Coord_i^k - Coord_i^{pooled}$|, where $RMSD_i^k$ represents the RMSD in single cell $k$ at the genomic location $i$.

### 3.4.11 Differential methylated gene analysis across clusters of single cells

To demonstrate the relationships between 3D chromatin structures and gene regulations, 351 mESC cells from the snm3C data are grouped into two clusters based on the pairwise distance matrices and two clusters are selected based on the highest average silhouette score. The coupled single-cell DNA methylation signals of snm3C data are overlapped with gene promoters to quantify the single-cell gene activities. The differential methylation analysis of genes across two clusters of single cells is performed using the DEGseq2 package with default settings.

### 3.4.12 Analyses of single-cell chromatin interactions and genetic variants

To predict the single-cell chromatin interactions, the distance matrices induced by the predicted single-cell 3D chromosome structures are converted to the interaction frequency matrices using the same conversion function as above. FitHi-C is used to predict the statistically significant chromatin interactions from the single-cell interaction frequency matrices. To control the false positive rates, the p-value threshold is set to $1x10^{-20}$, which is a stringent criterion compared with previous analyses. As validation, we overlapped the single-cell chromatin interactions with the bulk Capture-C dataset and calculated the fraction of overlapping. To provide a mechanistic interpretation of GWAS SNPs and somatic mutations, we overlapped the SNPs with the single-cell chromatin interactions and calculated the enrichment of the SNPs. As a comparison, links between randomly selected DNA fragments with genomic distances controlled are generated. The motif matching score of CTCF is calculated using the TOMTOM software.

Furthermore, we predicted the three-way chromatin interactions from the predicted 3D structures of tFLAMINGO. For every set of three DNA fragments, the average pairwise spatial distances were calculated to quantify the compactness. Specifically, for a set of DNA fragments $i, j, k$ ($i < j < k$), the averaged 3D pairwise distance is calculated as: $D_{i,j,k} = \frac{1}{3}\left(D_{ij} + D_{ik} + D_{jk}\right)$, where $D_{ij}$ represents the 3D genomic distances between DNA fragment $i$ and $j$. As comparison, the average spatial distances of 1,000 sets of DNA fragments with the same 1D genomic distances are calculated: $D^{bg}_{m,n,p} = \frac{1}{3}\left(D_{mn} + D_{mp} + D_{np}\right)$, where $n - m = j - i$ and $p - n = k - j$. The empirical p-values are calculated as $P_{ijk} = \frac{1}{1000}(1 + \#\{D_{ijk} > D^{bg}\})$. Similar to the identification of the two-way chromatin interactions, the adjacent three-way interactions are pruned and the most significant ones are selected: $SI_{i,j,k} = argmin_{m,n,p} P_{mnp}\ for\ all\ |m - i| = |n - j| = |p - k| < 5$.

# CHAPTER 4

# DECIPHER THE COMBINATORIAL GRAMMAR OF TRANSCRIPTION FACTORS IN LONG-RANGE MULTI-ENHANCER REGULATION

## 4.1 INTRODUCTION

The comprehensive profiling of the epigenomic landscapes has discovered millions of putative enhancer regions across hundreds of cell lines. In the 3D space, enhancers are brought to the proximity of the gene through DNA loops and regulate the gene expressions. Such enhancer-gene link exhibits strong cell-type specificity and is ubiquitous across the human genome, highlighting the crucial role of enhancer regulation in cell differentiation and development. Beyond the one-on-one interaction between enhancers and genes, recent case studies demonstrate that multiple enhancers can synergistically control the expression of a single gene[87-89]. For example, the kni gene in the Drosophila embryo is regulated by the orchestration of an intronic enhancer and a distal enhancer which is ~35kb away from the promoter[87]. Another experimental analysis further shows that multi-enhancer regulation is crucial for phenotypic robustness[88]. These important biological discoveries demonstrate the complex landscape of transcriptional regulations and highlight the vital role of multi-enhancer regulation.

To characterize the interaction landscape of the genome, several experimental techniques have been developed, including Hi-C, ChIA-PET, and Capture-C, to measure the interaction frequencies between pairs of two genomic loci and demonstrate the cell-type-specific genome organization. However, these methods rely on the ligation between the cross-linked DNA anchors to detect the chromatin interactions, thus demonstrating

less ability to capture long-range and multi-way interactions. To further characterize the high-order chromosome conformations, three techniques that do not require proximity ligations are invented: SPRITE, GAM, and ChIA-Drop. The successful applications of these methods in the human genome and mouse genome not only generate the high-resolution chromatin contact maps but also unravel the multi-way enhancer-promoter interactions and the functional relationships between the co-binding transcription factors (TFs). For example, IRF, STAT, AP1, and SMAT family motifs are frequently observed within the interacting anchors, suggesting the cooperative action of the TFs in shaping the 3D chromosome structures and regulating gene expressions.

Although these experimental techniques have revealed the multi-way chromatin interactions and largely expanded the understanding of the interplay between genes, enhancers, and TFs, they are only available in human GM12878 cells and mouse ESC cells, thus limiting the studies of other important tissues and cell lines. In addition, these techniques can only capture the interactions between relatively large DNA fragments (SPRITE: 5kb; GAM: 30kb; ChIA-Drop: 10kb), which is not sufficient to pinpoint the actual functional enhancers.

Given these limitations, computational methods are developed to predict the cell-type-specific enhancer-promoter interactions by integrating multi-omics datasets generated by the large consortia, e.g. ENCODE and Roadmap Epigenomics projects. By evaluating the gene expression, enhancer activates, genomic distances, and other DNA sequence features, these methods link the enhancers to the promoters and predict the cell-type-specific enhancer-promoter interactions. In light of the machine learning techniques, these methods can be classified into supervised learning methods and unsupervised

learning methods. For the unsupervised methods, the 1D genomic distances and the correlations between the enhancer activities and gene expression across diverse cell types are used as the predictive score to prioritize the enhancer-gene links. Compared with the supervised learning methods, the unsupervised learning methods do not require the experimental chromatin interactions to train the model but demonstrate lower accuracy based on comprehensive benchmarking analyses. For the supervised learning methods, the experimentally verified chromatin interactions are used to annotate the interacting enhancer-gene links. By learning the differential distributions of the multi-omics feature between the interacting and non-interacting enhancer-gene links, these methods can make genome-wide predictions. Currently, most computational methods in predicting long-range enhancer-gene interactions are supervised learning methods, including TargetFinder, JEME, IM-PET, ProTECT, RIPPLE, FOCS, and EAGLE. Besides the epigenomic signals, transcription factor (TF) binding sites are also used to improve the predictive accuracy. For example, TargetFinder uses the peaks of the TF ChIP-seq datasets within the enhancers, promoters, and intervening genomic windows as features to predict the enhancer-gene links. ProTECT further incorporates the Protein-Protein Interactions (PPIs) between the enhancer-binding and promoter-binding TFs to predict the TF-mediated enhancer-gene links. The expanded feature set largely improves the model performance. However, these methods can only be applied to cell types with experimental chromatin interaction datasets, which are unavailable in a significant fraction of cell lines. Although the trained models can be directly applied to another cell type to make cross-cell-type predictions, a recent analysis suggests that the cross-cell-type predictions are less accurate. In addition, a severe overfitting problem has been observed

for some methods, suggesting the high false-positive rates of the genome-wide predictions. Furthermore, the existing methods consider potential enhancer-gene links as independent samples and thus cannot capture the synergistic effect of multiple enhancers in regulating the same target gene. Finally, while TargetFinder and ProTECT use TF bindings as features, how TFs cooperate with each other to regulate distal target genes is still unclear, thus providing little mechanistic insights into the complex gene regulations.

According to the recent experimental studies, apart from the co-binding TFs in the same locus, TFs that bind to different loci can also synergistically regulate the expression of genes through multi-enhancer regulations. For example, GFI1b, RUNX1, and MYB are observed to regulate the expression of Myc gene by binding to a cluster of Myc enhancers. The deletion of a distal enhancer, which is ~ 1.7Mb away from the Myc gene, can significantly downregulate the expression of Myc gene, suggesting the critical role of the synergistic effect of TFs and enhancers in maintaining the gene expressions. In another example, the expression of -globin genes is regulated by five enhancers spanning a continuous 24kb genomic region. Interestingly, instead of binding to the same enhancer, four master erythroid transcription factors NF-E2, GATA1, SCL, and KLF1, show differential binding profiles across five enhancers, highlighting the cross-enhancer cooperation between these TFs. Globally, the integrative analysis of the long-range chromatin interactions and TF ChIP-seq datasets leads to the discoveries of TF clusters enriched in the interacting DNA anchors. These global analyses and case studies strongly support the need to model the multi-TF and multi-enhancer regulations in understanding the complex gene regulatory networks.

As stated above, existing methods can only model the one-on-one enhancer-gene links, except JEME. To quantify the additive effect of nearby enhancers, JEME jointly models the linear relationships between the activities of all nearby enhancers (+/- 1MB of the TSS) and gene expressions using a LASSO model. Together with the epigenomic signals and 1D genomic distances, the LASSO coefficients are used as a feature in the random forest model to predict the enhancer-gene links. By modeling the additive effect of multiple enhancers, the co-regulating enhancers tend to reside in the same TAD and super-enhancer, contain similar TF motifs and have correlated epigenomic signals across cell lines. However, JEME does not incorporate the information of TFs into the algorithm and thus cannot provide a mechanistic interpretation of the multi-enhancer regulations.

In this study, we developed a new unsupervised learning model, ComMUTE, to predict the long-range multi-way enhancer-gene links by mechanistically modeling the TF regulatory grammar of gene expressions. As a scalable Bayesian graphical model, ComMUTE integrates gene regulatory grammars by modeling the combinatorial TF modules of the co-regulating enhancers and thus mechanistically links multiple enhancers to the target genes simultaneously. In the framework of ComMUTE, genes are clustered into gene groups, and the enriched TF combinations across all genes are considered to be the group-specific TF grammars. Based on the gene group-specific TF combinations, a subset of enhancers that can synergistically provide the required TFs are prioritized as the co-regulating enhancers. Since ComMUTE is an unsupervised model, no experimental chromatin interaction dataset is required, which greatly expands the usability of ComMUTE in vast cell types without Hi-C datasets. We applied ComMUTE in 127 cell types/tissues to predict the cell-type-specific multi-way enhancer-gene links.

Compared with existing algorithms, ComMUTE demonstrates consistently improved performance by benchmarking with 19 cell-type-specific Hi-C and Capture-C datasets.

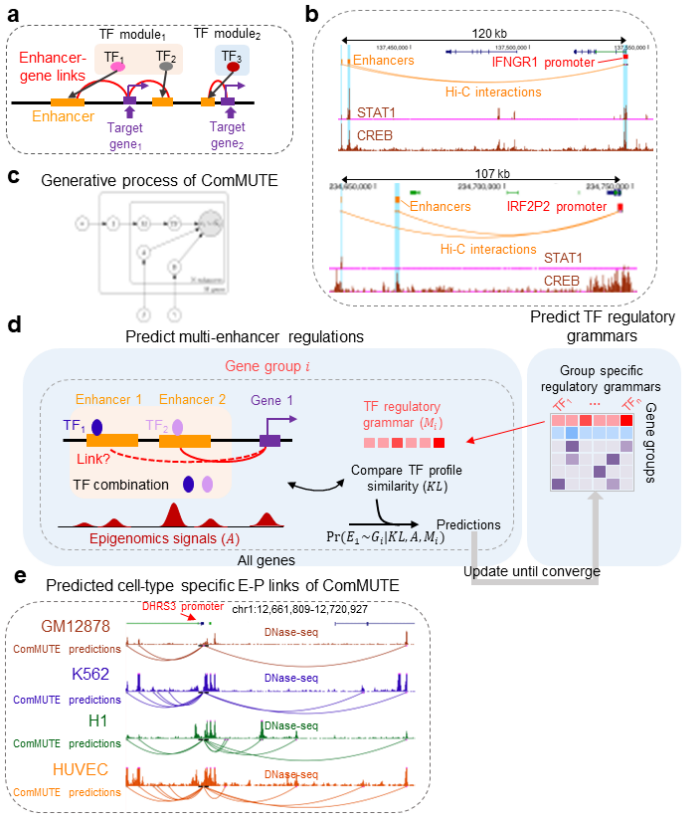



**Figure 4.1 Bayesian framework of ComMUTE in predicting multi-enhancer regulations.** (**a**) Genes are regulated by different combination of TFs (TF modules), which binds to multiple interacting enhancers with the same target gene. (**b**) Examples of Hi-C interactions linking enhancers (orange) and genes (red) showing the linked enhancers cooperatively provide the TF combination STAT1-CREB to regulate the target genes. (**c**) Plate diagram of ComMUTE. (**d**) Iterative scheme of ComMUTE. First, given the TF regulatory grammar, ComMUTE searches for a set of enhancers that cooperatively provides the required TFs (combinatorial TF profile) by comparing the similarity (KL). Combined with epigenomics data, ComMUTE predicts enhancer-gene interaction. Second, based on the predicted enhancer-gene interactions, the genes are assigned to different TF regulatory groups based on enhancer-binding TF profiles. ComMUTE repeats the two steps until convergence. (**e**) Example of predicted cell-type specific enhancer-gene interactions in GM12878, K562, HUVEC and H1. The ComMUTE predictions are consistent with cell-type specific DNase-seq data.

By randomly shuffling the TF bindings across enhancers in the inputs, we demonstrated that incorporating the TF module can significantly improve the accuracy of predicted enhancer-gene links. Interestingly, the co-regulating enhancers demonstrate high partial correlations of activities conditioned on the target gene expression, multi-correlations, and enrichment of Hi-C interactions, confirming the enhancers are directly associated and not mediated by the joint interacting promoters. Furthermore, the SPRITE multi-way interactions strongly support the multi-way enhancer-gene links. In addition to enhancer-gene links, we also evaluated the predicted TF regulatory grammars of gene groups. We observed a high PPI enrichment and clear co-expression patterns of the predicted combinatorial TFs. Moreover, the predicted enhancer-gene links are enriched with QTLs and GWAS SNPs. Strikingly, the epistasis eQTLs are precisely captured by the predicted multi-way enhancer-gene links, innovating new biological insights in mechanistically interpreting the high-order functional associations between SNPs.

## 4.2 RESULTS

### 4.2.1 ComMUTE predicts long-range multi-enhancer regulations based on TF regulatory grammars

In the framework of ComMUTE, the relationships between TFs, enhancers and genes are summarized as a three-layer network, where TFs bind to enhancers and enhancers interact with genes. The gene regulatory grammars are represented by the combinatorial TF profiles across all linked enhancers. Based on this model, the synergistic regulatory effect of multiple TFs is aggregated by the co-regulating enhancers (Figure 4.1.a). To verify the three-layer gene regulatory networks, we analyzed the Capture-C dataset and observed common TF combinations at different genomic loci, suggesting different genes

share similar regulatory grammars. As a representative example, we identified that the combination of STAT1 and CREB are collectively shared by the IFNGR1 gene and the IRF1P2 gene (Figure 4.1.b). At the IFNGR1 gene locus, two distal enhancers (~120 kb) are simultaneously linked to the gene promoter by the Capture-C interactions. While both enhancers show strong CREB binding signals, the STAT1 exclusively binds to the second enhancer. Similarly, two enhancers are linked to the promoter of the IRF1P2 gene (Figure 4.1.b). Interestingly, STAT1 and CREB show differential binding signals within two enhancer regions, suggesting the cross-enhancer cooperation is necessary for the activation of the target gene.

Based on the three-layer gene regulatory network, ComMUTE is specifically designed to model the multi-enhancer regulations and complex regulatory grammars. To predict the interacting probabilities of candidate enhancer-gene links, ComMUTE evaluates whether linking the enhancers to the target genes can improve the TF profiles of all co-regulating enhancers towards the required gene regulatory grammar. This unique design enables ComMUTE to model the joint regulatory effect of multiple enhancers and TF grammars, which distinguishes ComMUTE from existing algorithms. The interacting probability between enhancer $i$ ($e_i$) and gene $j$ ($G_j$) is calculated as $P(e_i \sim G_j | A, TF_{e_i}, TF_{G_j}^{-e_i}, M_I)$, where $A$ represents the activity based features(enhancer activity, gene expression and their correlation across 127 cell-types), $D$ represent the 1D genomic distance between enhancers and TSS of genes, $TF_{e_i}$ represents the binarized occurrence vector of TF motifs for the candidate enhancer and $TF_{G_j}^{-e_i}$ represents the TF profile of all enhancers linked to the $G_j$ except $e_i$. These features are selected because they demonstrate strong predictive power in dissecting Capture-C interactions and random enhancer-gene links

(Figure C.1). To model the complex regulatory grammar, we introduced two latent variables to represent the gene group membership ($I$) and the group-specific TF regulatory grammar ($M$). Since both $I$ and $M$ are unknown, ComMUTE further predicts the probabilistic membership of genes by comparing the gene-specific TF profiles to the group-specific regulatory grammars: $P(I = k|TF_{G_j}, M)$, where $TF_{G_j}$ represents the TF profile of all linked enhancers and $M$ represents the group-specific regulatory grammars. After the gene groups memberships are updated, $M$ is updated accordingly.

To efficiently infer the underlying distributions of features and latent variables, ComMUTE utilized an iterative Gibbs sampling framework (Figure 4.1.c). In each iteration, genes are assigned to different groups based on the predicted enhancer-gene links from the last iteration and the group-specific TF profiles are calculated. The prediction of enhancer-gene links are then calculated by evaluating the KL divergence between the gene regulatory grammar$M_I$ and TF profile of $G_j$ if the enhancer is linked to the gene: $exp\ (-R * KL(TF_{G_j}|M_I))$, where $R$ is a free scaling parameter. The predicted enhancer-gene links are then used to update the gene group membership. To avoid the searching for optimal enhancer combinations stuck at certain states through iterations, ComMUTE adopted a Simulated Annealing-based searching strategy and tested different combinations of enhancers and TFs before moving to the next iteration. To tune the unknown model parameters, we tested a wide range of values for the scaling parameter $R$ and number of gene groups, and selected the optimal values based on the highest AUROC when benchmarking predicted enhancer-gene links with Capture-C interactions in GM12878 (Figure C.2). Upon convergence, three sets of predictions are made by ComMUTE: (1) probabilistic score of enhancer-gene links, (2) mixture memberships of genes and (3) TF

regulatory grammar across gene groups. These outputs systematically delineate how multiple TF synergistically regulate the gene expression through multi-way enhancer regulations.

To expand the generalizability of ComMUTE, we predicted enhancer-gene links based on the imputed and non-imputed multi-omics dataset in diverse cell types. In total, four

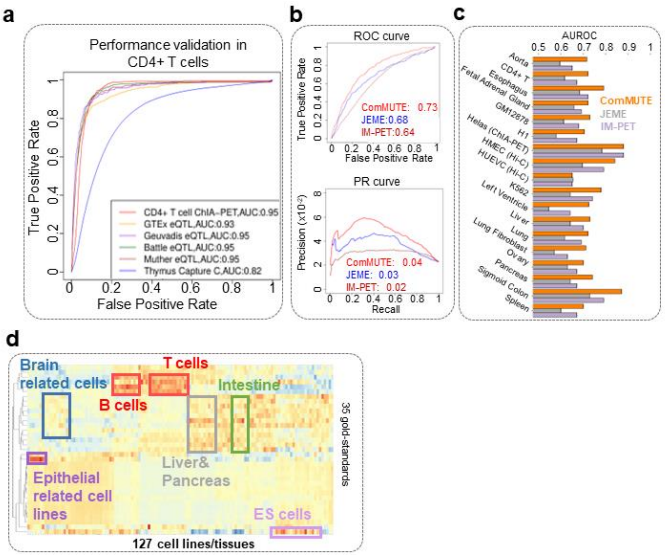



**Figure 4.2 Performance comparisons with JEME across 35 gold-standards support the superior performance of ComMUTE.** (**a**) Performance of ComMUTE with shuffled TFs (blue) and only one gene group (purple). For the shuffled TF version, the TF binding profiles are shuffled across all enhancers to disable the TF features of ComMUTE. For the one-gene-group version, all genes are assigned to the same gene groups, aiming to only capture master TF regulators for all genes. Compared with the shuffled TF and one-gene-group version, ComMUTE achieves higher AUROC (y-axis) in both K562 (upper) and GM12878 (upper). (**b-c**) Examples of ComMUTE predicted enhancer-gene links (blue) based on the combination of NF-κB-CREB (b) and SMAD4-ZBTB33-NF-κB (c). The predicted enhancer-gene links are supported by Hi-C (brown), ChIA-PET (purple) and Capture-C (red).

versions of predictions are generated: (1) imputed DNase-seq and RNA-seq across 127 cell-types/tissues; (2) imputed H3K27ac and RNA-seq across 127/tissue; (3) non-imputed DNase-seq and RNA-seq across 29 cell-types/tissues and (4) non-imputed H3K27ac and RNA-seq across 29 cell-types/tissues. These large-scale predictions of enhancer-gene

links provide a rich resource for understanding the complex multi-enhancer regulations. In this paper, we focused on the first version due to the broadest cell-type coverage. Across 127 cell-types/tissues, ComMUTE predicted ~80,000 cell-type-specific enhancer-gene links for each context. The predicted enhancer-gene links follow a similar 1D genomic distance distribution to Capture-C interactions. On average, over 20% of enhancers and over 85% of genes have a degree of more than one, suggesting the universal existence of multi-enhancer regulations (Figure C.3). As expected, the predicted enhancer-gene links show significantly higher correlations over the random chromatin interactions (Figure C.3), supporting the functional interactions between enhancers and genes. As an example, distinct cell-type-specific regulatory landscapes are predicted at the DHRS3 gene locus across GM12878, K562, H1 and HUVEC (Figure 4.1.e). The predicted enhancer-gene links precisely capture the cell-type-specific DNase peaks. Interestingly, a distal enhancer is linked to the DHRS3 gene (~40kb away) by skipping the nearest gene, suggesting the ability of ComMUTE to capture the long-range enhancer-gene links.

### 4.2.2 Robust performance in predicting enhancer-gene links

To systematically evaluate the performance of ComMUTE, we compared the genome-wide predictions of ComMUTE with 19 experimental chromatin interactions, i.e. Capture-C, Hi-C and ChIA-PET, and 16 tissue-specific eQTL annotations. As a commonly used metric, AUROCs that are calculated based on the Cross-Validations are used to evaluate the performance of the supervised learning models. However, significant concerns are raised about the inflated performance due to inappropriate segmentation of the training/testing sets and the selection of negative samples. Unlike the currently existing

methods, ComMUTE is an unsupervised algorithm that does not need experimental datasets for training and does not need to Cross-Validations for performance validation. Given this significant algorithmic advancement, ComMUTE is free from the risk of overfitting, and the genome-wide predictions can be directly evaluated based on the orthogonal experimental chromatin interactions. As a representative example, the predicted enhancer-gene links achieve high AUROC (>0.93) in GM12878 by benchmarking with ChIA-PET interactions and four eQTL datasets, supporting the superior performance of ComMUTE (Figure 4.2.a). The performance of ComMUTE is compared with six state-of-art algorithms: JEME, TargetFinder, IM-PET, RIPPLE, Ernst et al and FOCS. Since the predictive probabilities of JEME and IM-PET are available across 127 cell types, we compared the performance of ComMUTE with these two methods based on 19 chromatin interaction datasets. Based on the observation that the 1D genomic distances can significantly inflate the AUROCs, we limited the comparison to the commonly evaluated enhancer-gene links of all three methods. This strategy guarantees that these methods are benchmarked on the exact same set of enhancer-gene links, thus excluding the confounding effects induced by the differential distributions of the input features. Based on the rigorous evaluation strategy, ComMUTE demonstrated consistently improved AUROC over JEME and IM-PET (Figure 4.2.b). As shown in Figure 4.2.c, ComMUTE achieved an AUROC of 0.73 in the CD4+ T cell, which is higher than JEME (AUROC: 0.68) and IM-PET (AUROC: 0.64). In addition, the AUPR based on the predictions of ComMUTE (AUPR: 0.04) is also higher than the predictions of JEME (AUPR: 0.03) and IM-PET (AUPR: 0.02). Globally, ComMUTE outperforms JEME and IM-PET across all 19 comparisons, suggesting a strong agreement between the

ComMUTE predictions and experimental chromatin interactions. In addition to the AUROC, we further calculated the enrichment of experimental chromatin interactions and the ranked enhancer-gene links. In almost all scenarios, ComMUTE achieves higher enrichments of true positives over JEME, especially for the top-ranked links (Figure C.6). In addition to JEME and IM-PET, we also compared with RIPPLE, Ernst et al, and FOCS (Figure C.7). The enrichment of ten experimental chromatin interactions among the predicted enhancer-gene links is calculated to quantify the model performance. In fact, since all of the existing methods are trained on the experimental data, they naturally tend to prioritize the experimentally validated enhancer-gene links, leading to an inflated enrichment. Strikingly, ComMUTE still demonstrated the highest enrichment in eight of ten comparisons, excluding comparisons based on Hi-C datasets in HeLa cells and K562 cells (Figure C.8). The predictions based on different epigenomic features also show improved performance compared with the background pairs with 1D genomic distances controlled (Figure C.9).

To demonstrate the usability of ComMUTE in expanding the understanding of gene regulations to cell types without experimental chromatin interaction datasets, we compared the performance of ComMUTE with TargetFinder to predict cell-type-specific enhancer-gene links. TargetFinder is selected based on the improved performance in making cross-cell-type predictions compared with other methods. We trained TargetFinder in one of the six-cell types using Hi-C interactions and used the trained model to predict enhancer-gene links in the other five cell types. We compared 30 sets of cross-cell-type predictions of TargetFinder with the cell-type-specific predictions of ComMUTE. As shown in Figure C.8, ComMUTE shows consistently high AUPR in all cell

types, which further supports the superior performance of ComMUTE in predicting high-quality enhancer-gene links in understudied cell types.

We further studied the cell-type-specificity of the predicted enhancer-gene links in 127 cell types/tissues based on the enrichment of 35 functional interaction datasets. As shown in Figure 4.2.d, different clusters of cell type are identified, suggesting the predicted enhancer-gene links in this cell-type are highly similar and supported by the same sets of gold standards. For example, distinct clusters for B-cells, T-cells, Epithelial related cells,

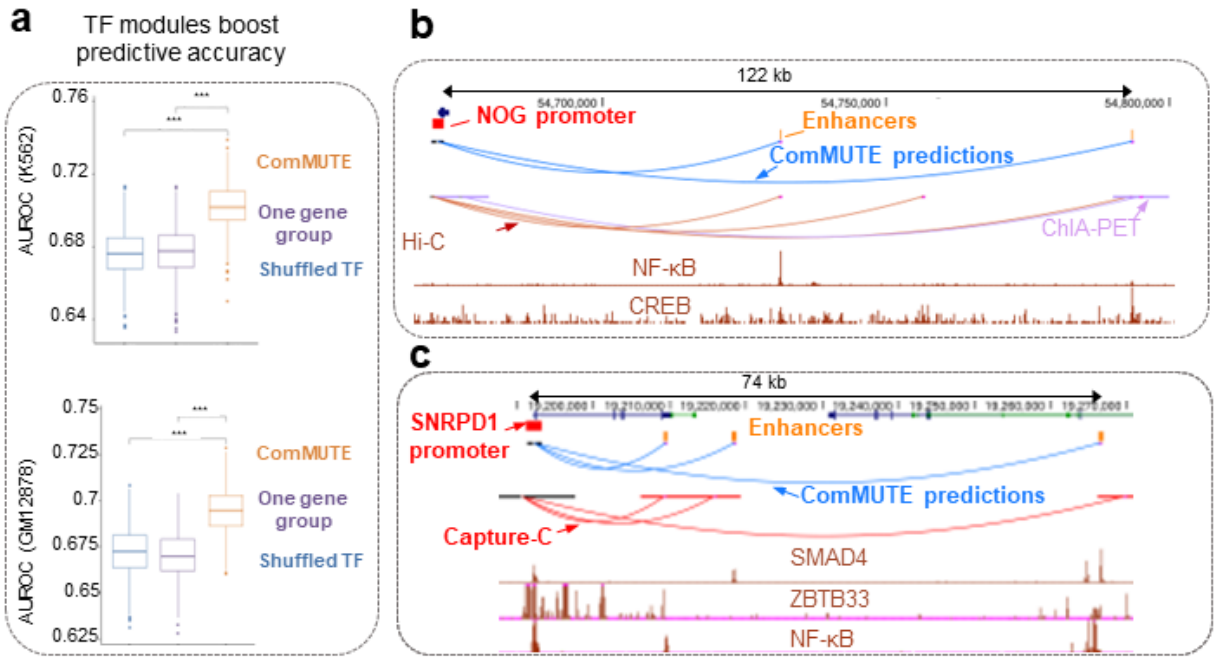



**Figure 4.3 Integration of TF modules improve the predictive accuracy.** (**a**) Performance of ComMUTE with shuffled TFs (blue) and only one gene group (purple). For the shuffled TF version, the TF binding profiles are shuffled across all enhancers to disable the TF features of ComMUTE. For the one-gene-group version, all genes are assigned to the same gene groups, aiming to only capture master TF regulators for all genes. Compared with the shuffled TF and one-gene-group version, ComMUTE achieves higher AUROC (y-axis) in both K562 (upper) and GM12878 (upper). (**b-c**) Examples of ComMUTE predicted enhancer-gene links (blue) based on the combination of NF-κB-CREB (b) and SMAD4-ZBTB33-NF-κB (c). The predicted enhancer-gene links are supported by Hi-C (brown), ChIA-PET (purple) and Capture-C (red).

and ES cells are observed, suggesting the fundamentally rewired regulatory links during the cell differentiation and development.

### 4.2.3 Integration of the TF regulatory grammar boost the predictive accuracy

As a significant algorithmic advancement of ComMUTE, the combinatorial TF modules of gene regulations are integrated to boost the prediction of enhancer-gene links. To demonstrate the contribution of TF-related features in improving the performance of ComMUTE, we specifically tested two cases. Firstly, we set the number of gene groups to one (termed as the 'one gene group' hereafter). In this case, only the TF regulatory grammar shared by all genes is identified and used to predict enhancer-gene links. In the second case, we shuffled the TF binding profile within enhancers to disable the TF feature (termed as the 'shuffled TF' hereafter). In this case, the prediction of enhancer-gene links is solely based on the epigenomic features and 1D genomic distances. Strikingly, the original configuration of ComMUTE achieves a median AUROC of ~0.7 in GM12878, while the one gene group setting and shuffled TF setting achieve AUROCs around 0.68 (Figure 4.3.a). Similar decreased performances are also observed in K562 (Figure 4.3.a). These permutation analyses demonstrate that the improved accuracy of ComMUTE is due to the integration of the TF module and prove that the TF modules predicted by ComMUTE can accurately capture the underlying gene regulatory grammar. For example, NF-κB and CREB are well-known co-factors in regulating gene expression. These two TFs are also predicted to synergistically regulate gene expressions in our predictions. Given the discovered regulatory grammar of NF-κB and CREB, the predicted multi-enhancer regulations accurately captured the cross-enhancer cooperation of these two TFs. At the NOG gene locus, ComMUTE linked two distal enhancers to the gene

promoter across a 122kb genomic window. The predicted enhancer-gene links are extensively supported by Hi-C and ChIA-PET interactions. Interestingly, a strong NF-κB binding peak is observed in the first enhancer but not in the second enhancer, and conversely, CREB only shows high signals in the second enhancer (Figure 4.3.b). The multi-enhancer regulations at the SMAD12 gene locus also capture the exclusive NF-κB and CREB binging signals across co-regulating enhancers (Figure C.10.a). A more complex TF module of SMAD4, ZBTB33, and NF-κB are captured at the SNRPD1 locus, where three enhancers are linked to the gene promoter and supported by Capture-C interactions (Figure 4.3.c). Although none of the linked enhancers contain binding sites of all three TFs, the multi-enhancer regulation facilitates the formation of the TF module

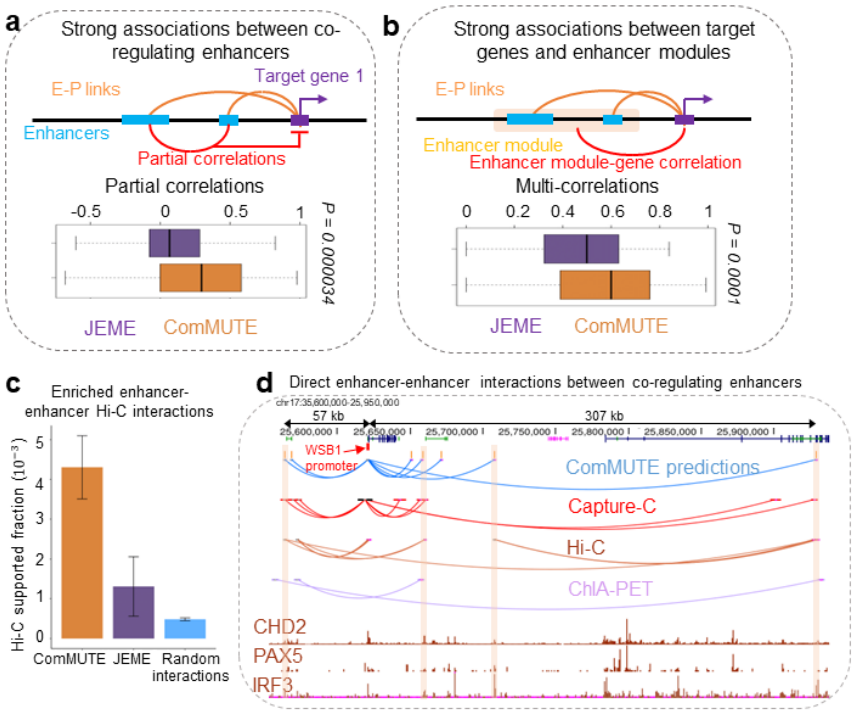



**Figure 4.4 Direct functional and physical interactions between predicted co-regulating enhancers.** (**a**) The enhancers regulating the same genes are directly interacted, which yields higher partial activity correlations conditioned on the common target genes (top). Compared with JEME, ComMUTE predictions show significantly

**Figure 4.4 (cont'd)**

higher partial correlations between co-regulating enhancers. (**b**) The co-regulating enhancers synergistically regulate the target gene, which yields higher multi-correlation between all enhancers and genes. Compared with JEME, ComMUTE predictions show significantly higher multi-correlations. (**c**) Co-regulating enhancers predicted by ComMUTE are enriched with direct Hi-C interactions (y-axis) compared with JEME and random enhancer-enhancer interactions (x-axis). (**d**) Example of multi-enhancer regulations predicted by ComMUTE. ComMUTE predicts the enhancer-gene interactions (blue curve) between seven enhancers (orange) to the WSB1 gene (red) based on the TF combination CHD2-PAX5-IRF3, and five of them are supported by Capture-C interactions (red curve). The direct interactions between enhancers are supported by Hi-C (brown curves) and ChIA-PET (purple curves).

and precisely controls the gene expression. A similar example of the PARS2 gene also shows that three linked enhancers collectively provide the TF module of SMAD4-ZBTB33-NF-κB (Figure C.10.b). Together with the permutation tests, these examples strongly support the superior ability of ComMUTE in decoding the regulatory grammar of the long-range gene regulations.

### 4.2.4 ComMUTE captures direct enhancer-enhancer interactions

In ComMUTE, multiple enhancers are linked to the same gene based on their functional cooperation in regulating gene expressions. Here, we demonstrate that the functional and physical interactions between co-regulating enhancers are direct and not mediated by the common target genes. Firstly, we calculated the partial correlations of the enhancer activities between the co-regulating enhancers conditioned on the target gene expression (Figure 4.4.a). The partial correlations evaluate the direct associations of enhancer activities between enhancers by removing the indirect associations mediated by the target gene. Compared with the enhancer modules predicted by JEME, the multi-enhancer regulations predicted by ComMUTE shows significantly higher partial correlations, suggesting the co-regulating enhancers have strong direct functional associations.

Secondly, the multi-correlations are calculated to measure the association between the target genes and all co-regulating enhancers, which evaluates the combinatorial regulatory effect of all co-regulating enhancers on gene expressions (Figure 4.4.b). In comparison, ComMUTE achieved the multi-correlation of 0.6 (median), while JEME only achieved ~0.48 multi-correlations, suggesting that co-activation patterns between the co-regulating enhancers and target genes. The results of the comparison based on partial correlations and multi-correlations strongly support the accuracy of the multi-enhancer regulations predicted by ComMUTE.

To further verify the direct physical interactions between multiple enhancers, we calculated the enrichment of the Hi-C interactions among all possible pairwise interactions between co-regulating enhancers. Compared with JEME and randomly linked enhancers, ComMUTE demonstrated the highest enrichment (Figure 4.4.c), supporting the enhancers are directly interacting with each other in the 3D space. As a representative example, seven enhancers are linked to the WSB1 gene promoter based on the TF combinations of CHD2-PAX5-IRF3 (Figure 4.4.d). Compared with Capture-C interactions, four out of seven predictions are supported, including the interaction of an enhancer that is ~307kb away from the gene promoter. Interestingly, three enhancer-enhancer links are supported by Hi-C and Capture-C interactions. The left-most enhancer is linked to the right-most enhancer across a 464kb genomic window, suggesting the existence of a DNA loop at this locus. In another two examples of the LMTK2 gene and the RALY gene, the predictions of ComMUTE are not only extensively supported by Capture-C but also capture the long-range enhancer-enhancer interactions of Hi-C and ChIA-PET interactions (Figure C.13). These results demonstrate that the co-regulating enhancers

directly interact in the 3D space and establish functional cooperation in regulating gene expressions.

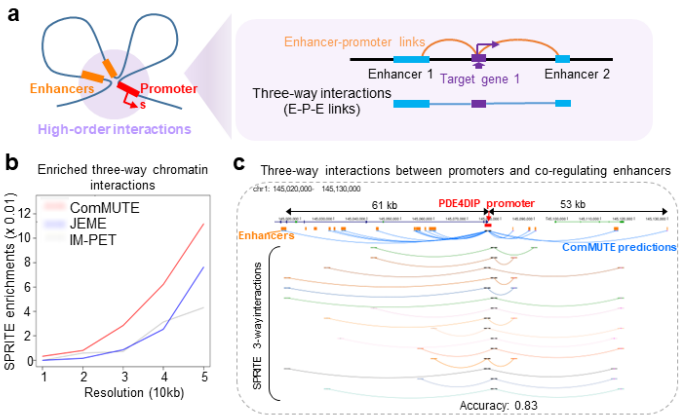



**Figure 4.5 Validation of the predicted multi-enhancer regulations based on SPRITE.** (**a**) The multiple interacting enhancers and genes are densely organized in 3D space and forming high-order chromatin interactions, e.g. three-way enhancer-promoter-enhancer interactions. (**b**) ComMUTE predictions (red) are enriched with SPRITE three-way chromatin interactions (y-axis) under all resolutions (x-axis), compared with JEME (blue) and IM-PET (grey). (**c**) Example of ComMUTE predicted multi-enhancer regulations (blue curves). The three-way enhancer-promoter-enhancer interactions formed by the co-regulating enhancers and the promoter are extensively supported by SPRITE.

### 4.2.5 Superior accuracy in predicting multi-enhancer regulations

In addition to evaluating the accuracy of the enhancer-gene links and enhancer-enhancer links, we further evaluated the multi-enhancer regulations based on the SPRITE dataset. Unlike the evaluations based on the pairwise interactions, the integration of the SPRITE dataset facilitates the evaluation of multi-way interactions, i.e. co-regulating enhancers and the target gene. We specifically focused on the three-way interactions, where two enhancers are simultaneously linked to one gene (Figure 4.5.a). By overlapping the predicted three-way enhancer-gene links with SPRITE three-way interactions under different resolutions, the enrichment of SPRITE interactions is calculated to quantify the accuracy of the predicted three-way enhancer-gene links. Strikingly, ComMUTE

demonstrates a significantly higher enrichment over JEME and IM-PET across all resolutions (Figure 4.5.b), supporting the high accuracy of the predicted three-way enhancer-gene links. As a representative example (Figure 4.5.c), 13 enhancers are linked to the PDE4DIP gene promoter in a ~114kb genomic window. The possible three-way interactions between two enhancers and the PDE4DIP gene promoter are extensively supported by 15 SPRITE three-way interactions in 1kb resolution. By combining proximal enhancers within the 1kb genomic window into 18 1kb genomic bins, 83% of possible three-way interactions are supported by SPRITE interactions.

### 4.2.6 ComMUTE decodes the TF regulatory grammars of gene expression

One of the significant contributions of ComMUTE is discovering TF modules that can synergistically regulate the target gene expressions, which represent the underlying gene regulatory grammars. Based on the predictions of ComMUTE, diverse TF combinations are captured across different gene groups (Figure 4.6.a), suggesting that the gene regulatory grammars are highly complex. Such TF combinations cannot be observed from the traditional co-binding analyses for three reasons. First, the flexible Bayesian framework of ComMUTE allows the TFs to bind to some, but not all, of the co-regulating enhancers. Therefore, instead of co-binding in the 1D space, these TFs are interacting in the 3D space, which can not be captured based on the correlations of TF ChIP-seq signals. Second, the prioritized TF modules are predicted for clusters of genes. For each gene cluster, the members are not required to have spatial proximity and could be far away from each other or even located in different chromosomes. Thus, the long-range functional gene-gene relationships can not be captured by the co-binding analyses based

on the sequential TF ChIP-seqs. Third, ComMUTE only prioritized the functional TF combinations. Although TF motifs are frequently observed within co-regulating

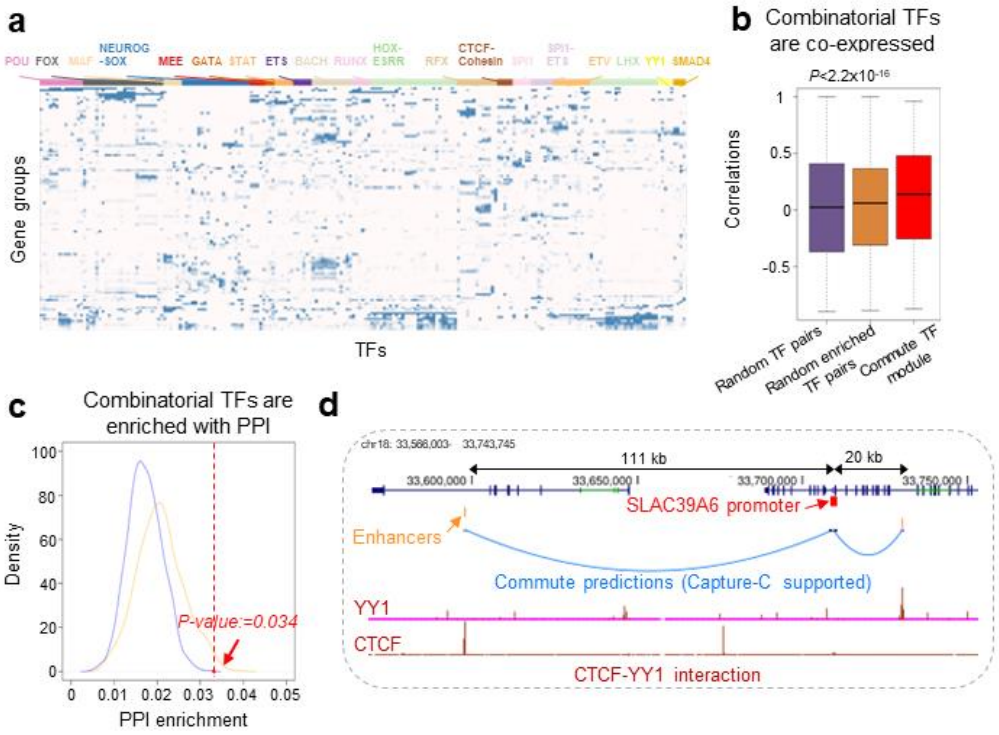



**Figure 4.6 Accurate predictions of cooperative TF modules.** (**a**) Gene-group-specific TF combinations predicted by ComMUTE. The heatmap shows the TF enrichments (columns) across all gene groups (rows). Twenty clusters of TFs are observed. (**b**) TF module predicted by ComMute are significantly co-expressed across cell-types compared with random TF combinations (purple) and (**c**) The TF modules predicted by ComMUTE are enriched with PPI compared (x-axis) with controls (blue and orange curves). (**d**) Example of ComMUTE predicted enhancer-gene links with CTCF binding sites in one enhancer and YY1 binding sites in another enhancer. The regulatory function of the PPI between CTCF and YY1 is well-studied.

enhancers, only a few of them are functional and contribute to the gene regulations. By modeling the TF grammar of gene regulations, only master TF regulators are used for predicting enhancer-gene links. Compared with the TF groups captured by hierarchical

clustering analysis (Figure C.14), the TF profiles predicted by ComMUTE show clear TF enrichments.

To evaluate the predicted TF regulatory grammar, we calculated the pairwise correlations of the TF gene expressions within TF modules predicted by ComMUTE (Figure 4.6.b). As a comparison, the correlations based on the randomly paired TFs are calculated. To further control the potential bias caused by the different occurrence of motifs across TFs, we generated a more stringent control by randomly pairing TFs that are captured by at least one TF module. Compared with both controls, the TF modules prioritized by ComMUTE shows the highest correlations of activities, supporting the functional interactions of the TF grammar. We further validated the predicted TF modules using the PPI datasets. Compared with controls, the predicted TF modules are highly enriched with PPIs, suggesting these TFs are not only functionally associated but also physically interacted (Figure 4.6.c). As a representative example, Figure 3.6.d shows an example of the two co-regulating enhancers which are predicted from the YY1-CTCF TF combinations. Interestingly, YY1 only binds to the right most enhancer and CTCF only binds to the left most enhancer, suggesting the YY1-CTCF combinations are co-localized in the 3D space, rather than in the 1D genome. The PPI between YY1 and CTCF are well-known and plays an important role in establishing chromatin interactions and gene regulations, which is consistent with the observed long-range multi-enhancer regulation of the SLAC39A6 gene (Figure 4.6.c).

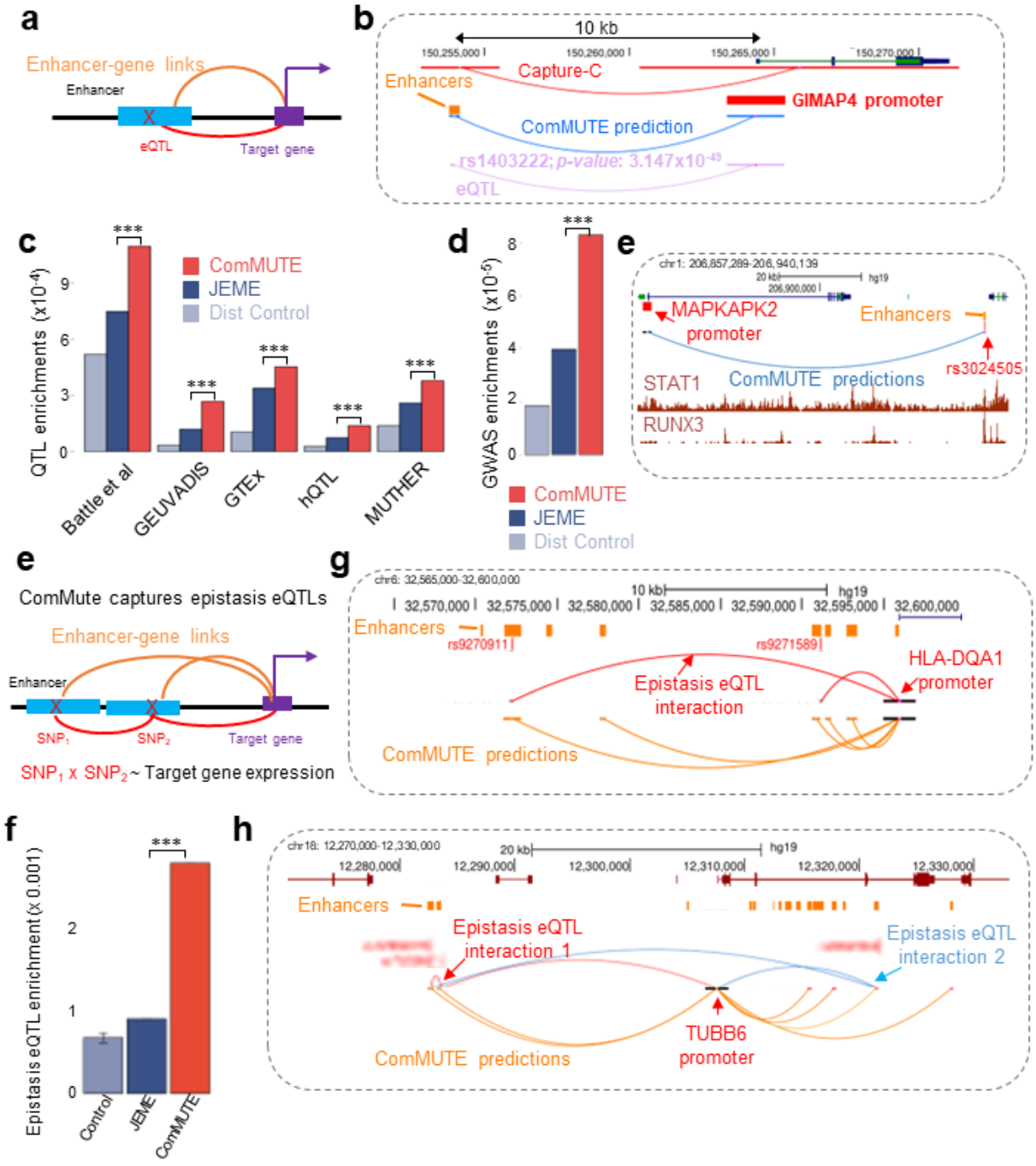



**Figure 4.7 Predicted enhancer-gene interactions are enriched with eQTLs.** (**a**) Schematic figure of eQTL SNPs located in the enhancers, whose functional interactions with the target genes are mediated by enhancer-gene interactions. (**b**) Example of the ComMUTE predicted enhancer-gene link (blue curve) mediating the SNP-gene interaction of eQTL(purple curve, rs1403222, p-value: $3.147x10^{-49}$). The predicted enhancer-gene interaction is supported by the Capture-C interaction (red curve). (**c**) Global enrichment of eQTLs from different resources (x-axis) in predicted enhancer-gene links of ComMUTE, JEME and random controls. In all groups, ComMUTE predictions

**Figure 4.7 (cont'd)**

show significantly higher enrichment compared with JEME predictions (p-value < $2.2x10^{-16}$, Binomial test). (**d**) Enrichment of GWAS SNPs in predicted enhancer-gene links. (**e**) Multi-enhancer regulations capture epistasis eQTLs. The interactions between SNPs in regulating the common gene expression are mediated by three-way enhancer-promoter-enhancer interactions. (**f**) Enrichment of epistasis eQTLs (y-axis) in the multi-enhancer regulatory networks. Multi-enhancer regulatory networks predicted by ComMUTE are significantly enriched with epistasis eQTLs compared with JEME (p-val=$1.73x10^{-5}$, Binomial test). (**g**) Example of predicted multi-enhancer regulatory networks (orange curves) mediating epi-stasis eQTLs. The SNPs of two pairs of epistasis eQTLs of TUBB6 gene, i.e rs12966726-rs7229921 pair and rs12966726-rs8092506 pair, are located in the interacting enhancers and regulate the gene through enhancer-gene links.

### 4.2.7 Predicted enhancer-gene links are enriched with QTLs and GWAS SNPs

To further support the superior accuracy of ComMUTE, we utilized the functional genomic datasets of eQTLs and hQTLs by calculating the enrichment of QTLs in predicted long-range enhancer-gene links. The enhancer-gene links are supported by the QTL datasets if the enhancers harbor the SNPs and are linked to the same target genes or histone peaks (Figure 4.7.a). As a representative example, ComMUTE predicted one enhancer gene link at the GIMAP4 gene locus, which is supported by the Capture-C interactions (Figure 4.7.b). Interestingly, the linked enhancer harbors a significant eQTL of the GIMAP4 gene (rs1403222, p-value=$3.15x10^{-49}$). A similar example at the IL6R gene locus is also shown in Figure C.15, where four eQTLs of the IL6R gene are precisely captured by four predicted enhancer-gene links. Globally, we compared the QTL enrichment of ComMUTE with JEME and randomly linked enhancer-gene pairs with 1D genomic distance controlled and observed significantly higher enrichments of ComMUTE (p-value<$2.2x10^{-16}$) across five QTL datasets (Figure 4.7.c). These results not only supports the superior performance of ComMUTE in discovering the functional interactions

between enhancer-gene links but also suggest that the SNP-gene associations are mediated by the physical enhancer-gene links.

We further interpreted the GWAS SNPs based on the predicted enhancer-gene links. Compared with JEME and distance-controlled random links, ComMUTE predictions are significantly enriched with GWAS SNPs, suggesting the interacting enhancers are also strongly associated with disease phenotypes (Figure 4.7.d). Figure 4.7.e shows an example of Leukemia-associated SNP rs3024505. Based on the predictions of ComMUTE, the enhancer that contains the SNP rs3024505 is linked to the MAPKAPK4 gene, which is a well-known leukemia-related gene based on the TF grammar of STAT1 and RUNX3. By overlapping the SNP location with the TF ChIP-seq signals, we found the SNP is precisely located within the summits of the TF ChIP-seq peaks, which further supports the predicted regulatory effect of the combination of STAT1 and RUNX3. These examples, together with the global enrichment of QTLs and GWAS SNPs, provide a mechanistic interpretation of the disease association of the SNP, where the SNP disrupts the TF binding sites within enhancers and dysregulate the disease-related genes.

### 4.2.8 Multi-enhancer regulations unravel the regulatory basis of epistasis-QTLs

The key feature that distinguishes ComMUTE from existing methods is the prediction of multi-enhancer regulations with close spatial proximity and strong functional associations, which delineates the high-order chromatin interaction landscape. In the eQTL analyses, the high-order interactions between SNPs, i.e. epistasis eQTLs, are predicted to be associated with the disease phenotypes. However, the traditional analyses require a large number of tests to prioritize all possible combinations of SNPs, thus hampering their

applications in the whole genome. Here, we showed that the multi-way enhancer-gene interactions predicted by ComMUTE could help to discover the high-order interactions of SNPs within co-regulating enhancers.

To demonstrate that the predicted multi-enhancer regulation can precisely capture the epistasis eQTLs, we overlapped the SNPs of the epistasis eQTLs to the co-regulating enhancers with the same target genes and calculated the enrichment of epistasis eQTLs. Compared with the predictions of JEME and random links, ComMUTE achieves a significant higher enrichment, suggesting the predicted multi-enhancer regulations are not only physically interacted but also functionally associated (Figure 4.7.f). Take the HLA-DQA1 gene as an example, the interaction of two SNPs (rs9270911 and rs9271589) is predicted to be associated with the gene expression. Strikingly, the two epistasis eQTLs are captured by the multi-enhancer regulations predicted by ComMUTE, where the co-regulating enhancers captured the SNPs. Another example at the TUBB6 gene locus is also shown in Figure 4.7.h. Two pairs of epistasis eQTLs, i.e. rs12966726-rs7229921 and rs12966726-rs8092506, are identified as statistically significant epistasis eQTLs of the TUBB6 gene from the previous analysis. Interestingly, both sets of epistasis eQTLs are captured by the co-regulating module of six enhancers based on the predictions of ComMUTE.

## 4.3 DISCUSSION

In this study, we developed a Bayesian Graphical model, ComMUTE, to predict the multi-enhancer regulations across 127 cell types/tissues. By jointly modeling the TF bindings of all co-regulating enhancers, ComMUTE captures the synergistic effect of multiple enhancers in regulating the target gene expressions and unravels the complex high-order

gene regulatory landscape. As an unsupervised learning algorithm, ComMUTE does not require the experimental chromatin interaction datasets for training, thus showing strong generalizability compared with existing supervised learning algorithms. Furthermore, the unsupervised framework of ComMUTE fully addresses the overfitting risks of the existing algorithms and facilitates rigorous evaluation of the model performance. By extensively comparing the performance with existing cutting-edge algorithms based on 19 experimental chromatin interaction datasets, ComMUTE demonstrates consistently improved performance in predicting long-range enhancer-gene interactions. Based on the permutation analyses, we show that the integration of the TF binding sites and gene regulatory grammars can significantly improve the performance of ComMUTE, suggesting the cooperation of TFs is important in decoding the complex enhancer-gene regulatory networks.

The genome-wide application of ComMUTE in 127 cell-types/tissues based on four sets of imputed and non-imputed epigenomic and transcriptomic datasets to delineate the multi-enhancer regulations (Figure C.4). We show that the co-regulating enhancers predicted by ComMUTE demonstrate strong partial correlations, multi-correlations, and enrichment of Hi-C-supported enhancer-enhancer interactions, supporting the direct functional and physical interactions between these enhancers. We highlighted several examples where the long-range co-regulating enhancers are directly linked by Hi-C and ChIA-PET interactions. These results strongly support the utility of ComMUTE in predicting the large-scale cell-type-specific multi-enhancer regulatory landscape. Beyond predicting enhancer-gene links, the predicted TF grammars are also supported by the

PPIs and co-activate patterns, which suggest new biological innovations in gene regulations.

The accurate prediction of the multi-enhancer regulatory landscape provides new avenues to mechanistically interpret QTLs, GWAS SNPs, and epistasis eQTLs. The high consistency between the functional genomic datasets and predicted enhancer-gene links further supports the accuracy of the predicted enhancer-gene links. More importantly, these results suggest that the SNP-gene associations and SNP-disease association are mediated by the enhancer-gene links, which bring the SNPs to the proximal of the gene promoters and indirectly control the phenotypes. As a unique contribution of ComMUTE, the predicted multi-enhancer regulations facilitate the interpretation of the epistasis eQTLs, where the co-regulating enhancers bring SNPs to the proximal 3D neighborhoods of the target gene promoters. Together with the global enrichment analyses, we highlighted several examples where the SNPs of epistasis eQTLs are accurately captured by the co-regulating enhancers. Therefore, the predicted multi-enhancer regulations can help explain the discovered QTLs and GWAS SNPs and provide a mechanistic approach to significantly reduce the required number of tests in predicting epistasis eQTLs.

# CHAPTER 5

# PREDICT LONG-RANGE ENHANCER REGULATION BASED ON PROTEIN-PROTEIN INTERACTIONS BETWEEN TRANSCRIPTION FACTORS

A modified version of this chapter was previously published (Wang H. et al, 2021): Wang H.*, Huang B*., and Wang. J. (2021) Predict long-range enhancer regulation based on protein-protein interactions between transcription factors. Nucleic Acids Research.

## 5.1 INTRODUCTION

Cell-type specific transcriptional regulation plays important roles in differentiation and development [90-102]. In addition to proximal regulatory elements, e.g. promoters, which are located around transcriptional start sites (TSS) of genes, distal enhancers provide complex and precise controls on gene expression through long-range regulation [103, 104]. Based on recent genome-wide enhancer annotations from ENCODE and Roadmap Epigenomics projects [50, 86], hundreds of thousands of putative enhancers across the whole human genome have been identified, especially in non-coding regions, highlighting the biological impacts of enhancer regulation. Although a series of computational algorithms have been developed to predict the genomic locations of cell-type specific enhancers [105, 106], it remains challenging to identify the specific target genes regulated by enhancers in different cell-types or tissues. Unlike promoters, enhancers are usually located far away from their target genes along the genome [107] and the nearest genes may not be regulated by a proximal enhancer [108]. In three-dimensional (3D) space, an

enhancer and its target genes are placed close to each other through long-range chromatin interactions, i.e. enhancer-promoter interactions [109].

The discoveries of tissue-specific long-range enhancer regulation have the potential to enable novel insights in a wide range of different biological studies. As one of the canonical examples, long-range regulation by distal enhancers play pivotal roles in controlling the tissue and condition-specific expression of the mouse $\beta$-globin (Hbb) gene expression [90, 94, 95]. As another well-known example, the expression of the Shh gene in mouse limb bud is precisely regulated by a distal enhancer located 850kb away, which is critical for the proper limb development [96-98, 110]. In addition to normal tissue development, the annotation of long-range enhancer regulation has also facilitated the interpretation of genetic variants underlying complex diseases. A non-coding genetic variant associated with obesity is located in an intron of the FTO gene but regulates the IRX3 and IRX5 genes that are located >400kb away [91, 99, 111]. Similar examples of long-range interactions linking disease-associated genetic variants to distal genes have also been found in studies of autoimmune diseases [92, 93, 100-102].

Given the functional importance of long-range enhancer regulation, experimental techniques have been developed to identify chromatin interactions linking distal enhancers to promoters of their target genes. Based on the pioneering chromosome conformation capture (3C) technology [112], along with its derivatives of 4C and 5C [113, 114], the genome-wide version, i.e. Hi-C [7], has been applied to several human cell-types and tissues [10, 45, 50]. Furthermore, the promoter-enriched genome conformation assay, Capture Hi-C [115], improves the resolution and cell-type specificity of the identified chromatin interactions for gene promoters [116]. On the other hand, the method of chromatin

interaction analysis with paired-end-tag sequencing (ChIA-PET) [117] was developed to capture long-range chromatin interactions associated with a protein of interest, such as a specific transcription factor (TF), with high-resolution and cell-type specificity [118]. These cutting-edge technologies have generated large-scale chromatin contact maps for a number of cell-types or tissues in the human genome and other model species [10, 45, 50, 118].

Although experimental techniques have substantially expanded the catalog of annotations for long-range chromatin interactions, there are several limitations that hinder in-depth analysis on cell-type specific enhancer-promoter interactions. First, the resolution of interacting genomic anchors profiled by Hi-C and Capture Hi-C is relatively low (~5-10kb genomic fragments) [10, 115], which makes it difficult to pinpoint the specific enhancers involved in long-range regulation. Second, while Capture Hi-C and ChIA-PET experiments can discover cell-type or tissue-specific enhancer regulation, data generated by Hi-C experiments have been found to be largely invariant across different cell-types or tissues [119]. Third, the background noise levels of Hi-C and Capture Hi-C datasets are high, leading to many false positive discoveries [71]. Fourth, due to the dependency on specific protein antibodies, such as CTCF or RNA Pol II [118], each ChIA-PET experiment can only profile a subset of long-range interactions, resulting in large numbers of false negative interactions that are not identified [120].

Because of these limitations, computational models are needed to predict cell-type specific long-range enhancer regulation, based on integration of multi-omics signatures, e.g. genomics, transcriptomics, and epigenomics. Large-scale multi-omics data resources collected by the ENCODE and Roadmap Epigenomics projects contain the

multi-view information of gene regulation [50], including gene expression, transcription factor binding and histone modifications. They can help to overcome the limitations of experimental techniques because they are cell-type or tissue specific [121], provide high-resolution signal landscape along the genome [85, 122], have high signal-to-noise ratio [122], and cover the genomic binding sites for diverse transcription factors [50]. The existing computational models of long-range enhancer-promoter interaction prediction can be grouped into two classes. For the first class, i.e. supervised algorithms, 3D chromatin interactions profiled by experimental techniques are used as labels for enhancer-promoter pairs. The commonly used features include: 1) cell-type specific gene expression based on RNA-seq data; 2) enhancer activity based on specific epigenetic signals, such as H3K4me1, H3K27ac or DNase hypersensitivity; 3) genomic separation distance between enhancers and gene promoters; and 4) correlations between gene expression and enhancer activity. Supervised methods incorporating some or all of these features include RIPPLE [123], FOCS [124], EAGLE [125] and JEME [126]. As one of the most recently developed supervised methods, JEME [126] employs a combined approach of regression and random forest to predict long-range regulatory links between enhancers and genes. But it requires multi-omics datasets from a large panel of diverse cell-types and tissues as inputs, which is usually not available for users. The other two top-performing methods are IM-PET [127] and TargetFinder [128]. These two algorithms not only integrate the features described above but also leverage additional features of transcription factor binding in promoters, enhancers, or genomic windows between enhancers and promoters. With respect to machine learning techniques, IM-PET employs a random forest model, and TargetFinder implements a boosting tree approach. For the

second class, i.e. unsupervised algorithms, every enhancer-promoter pair is assigned with a score and then ranked based on the scores. Top-ranking enhancer-promoter pairs are predicted to interact with each other. The scores are generally based on genomic separation distance and co-activity patterns, e.g. correlations, between enhancers and genes [129-131]. Based on a systematic performance evaluation analysis [132], supervised

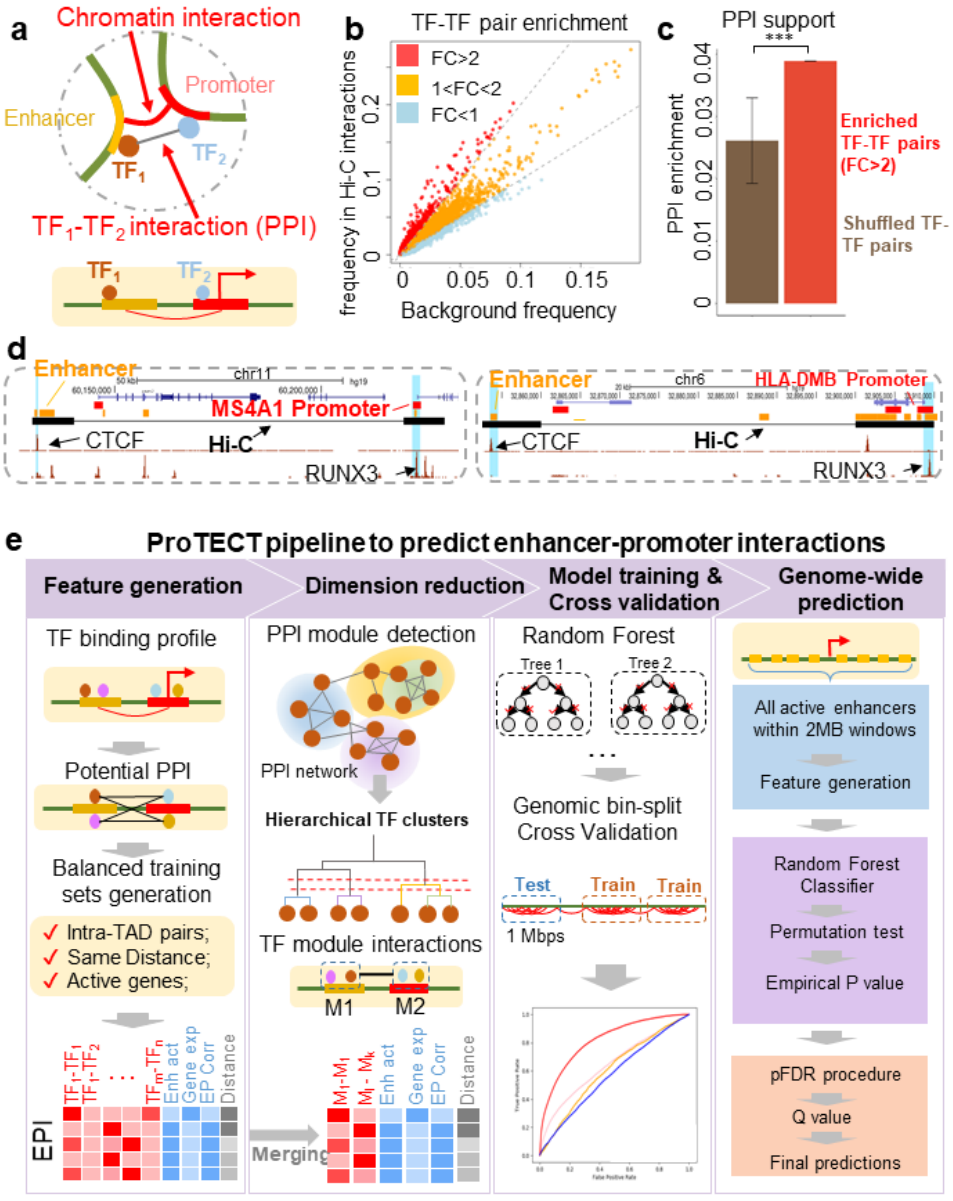



**Figure 5.1 Schema of ProTECT in predicting PPI mediated enhancer-gene links.** (**a**) The enhancer-promoter interactions are regulated by PPIs between enhancer-binding

**Figure 5.1 (cont'd)**

TFs (brown) and promoter-binding TFs (blue), which link distal enhancers (orange) to the proximity of promoters (red) in 3D chromatin structure. (**b**) Enrichment of TF-TF pairs in Hi-C interactions (y-axis) compared to background (x-axis). Points represent TF-TF pairs. Frequency is calculated as the fraction of enhancer-gene pairs containing the specific TF-TF pairs. Fold-change (FC) is the ratio of the frequency in Hi-C interactions over the frequency in background. TF-TF pairs are colored by the FC (red: FC>2; orange: 1<FC<2; blue: FC<1). (**c**) Enriched TF-TF pairs are supported by PPIs. The fraction of pairs supported by PPIs are calculated for the set of enriched TF-TF pairs (red). As controls, the TF members from the enriched TF-TF pairs are randomly paired (brown). Statistical test is done based on 1,000 random repeats of controls (***: p-value=$10^{-3}$). Error bar represents sd. (**d**) Examples of Hi-C interactions linking enhancers (orange) and promoters (red) showing enhancer-binding CTCF ChIP-seq peaks and promoter-binding RUNX3 ChIP-seq peaks in GM12878 cells. (**e**) The workflow of ProTECT algorithm. A balanced training dataset is generated with confounding factors controlled. A feature matrix summarizing cell-type specific TF PPI features, activity-based features (enhancer activity, gene expression, enhancer-gene activity correlation), and genomic distances is then constructed. A novel hierarchical network community detection-based approach is applied for feature dimension reduction. Based on the reduced feature matrix, a random forest model is trained, and rigorous genomic-bin split cross-validations are used for performance evaluations and comparisons. Using the trained predictive model, genome-wide high-confidence enhancer-promoter interactions are predicted based on stringent permutation statistical tests.

methods overall demonstrate better performance than unsupervised methods, but many of the supervised methods suffer from overfitting issues due to high model complexity [132] or excessively high-dimensional features that are often shared across training and testing sets [133]. Furthermore, existing methods provide limited mechanistic insights on how specific long-range chromatin interactions are established to link distal enhancers with promoters of target genes [134].

Interestingly, as shown by recent experimental studies [91, 135-140], in addition to the binding of individual TFs on enhancers or promoters, the protein-protein interactions (PPIs) between TFs have been found to participate in the process of long-range chromatin interaction formation and thus, mediate distal enhancer to the proximity of target gene

promoters (Figure 5.1.a-D). For example, the PPI between the enhancer-binding and promoter-binding YY1s (i.e. YY1 dimerization) has been found to mediate enhancer-promoter contacts [141]. The ChIA-PET data from mESCs suggests that the YY1-YY1 interactions largely participate in the connections between active enhancers and gene promoters [141]. In a chromatin structure engineering study, based on a CRISPR-dCas9 system, two proteins (PYL1 and ABL1) are fused to dCas9 and are guided to bind on different genomic locations [142]. Remarkably, the PYL1-ABL1 dimerization can establish novel long-range chromatin interactions, highlighting the mechanistic importance of PPIs in orchestrating chromatin loops. In addition, a couple of genome-wide analyses have also found that specific groups of transcription factors are enriched in cell-type specific long-range chromatin interactions [143-145]. Within each group, some TF members can interact with each other and form protein complexes. As a representative example, a group of CTCF, RAD21, SMC3 and ZNF143 is found to be enriched in chromatin interactions [143], consistent with the chromatin loop extrusion model that CTCF and cohesin can interact with each other and regulate chromatin loops [16, 146].

These observations strongly support the mechanistic hypothesis that specific TF PPIs may mediate long-range enhancer regulation. Therefore, incorporation of TF PPIs as a new set of features into a machine learning model is expected to improve the accuracy of long-range enhancer-promoter interaction predictions. Moreover, the prioritized TF PPIs from the predictive model can further indicate the important transcription factors that facilitate long-range enhancer regulation, leading to novel understandings of enhancer biology. However, unlike basic enrichment analysis of candidate TF-TF pairs that are over-represented in enhancer-promoter interactions [143-145], building a predictive model

based on TF PPI features is computationally challenging. First, the number of candidate TF PPIs is large (~200,000). By filtering the features using cell-type specific TF expression, there are still large amounts of potential TF PPI features. Take the human GM12878 cell-line as an example, by only considering TFs that are expressed [86], the number of PPIs between expressed TFs is ~1,900. The excessively high-dimensional TF PPI features easily render predictive models with high overfitting risks. Second, individual TF PPIs are not independent features because of 1) co-binding TF modules along the 1D genome [50]); and 2) protein complexes consisting of multiple interacting TFs [147, 148]. Both challenges require advanced feature dimension reduction approaches to efficiently handle the non-linear dependencies in features. In addition, as highlighted by recent benchmark studies [132, 133], rigorous settings of cross-validation need to be designed for unbiased performance evaluation and interpretation.

In this study, we developed a new predictive model, ProTECT, to infer long-range enhancer-promoter interactions with substantially improved accuracy. A unique novelty of the model is designing a graph-based dimension reduction algorithm, which can efficiently incorporate combinatorial TF PPI features into the model and, in the meantime, control the overfitting risks. By setting rigorous genomic bin-split cross-validations and controlling various confounding factors, we systematically demonstrated the superior performance of our model compared to existing algorithms. Furthermore, we analyzed the relative importance of TF PPI features in different cell-types and prioritized the key TF PPIs that may participate in the regulation of long-range enhancer-promoter interactions, leading to new mechanistic insights on enhancer regulation. Accordingly, we further classified genes into specific subsets, where enhancer-gene interactions are predicted to

be mediated by different TF PPIs. Interestingly, genes in different subsets are enriched with distinct biological pathways, suggesting the specific functional impacts of TF PPIs. Genome-wide implementation of ProTECT in human GM12878 and K562 cell-lines results in 134,792 long-range enhancer-promoter interactions, which are significantly enriched with cis-eQTLs. In addition, by analyzing enhancer-promoter interactions mediated by different TF PPIs, we were able to assign specific TFs as upstream trans-factors to downstream target genes through distal enhancers. Strikingly, the prioritized TF-gene pairs are significantly supported by trans-eQTLs, leading to new mechanistic interpretations of trans-genetic effects propagated through the combined regulatory pathways of TF bindings, TF PPIs and long-range chromatin interactions.

## 5.2 MATERIALS AND METHODS

To predict cell-type specific long-range enhancer-promoter interactions and obtain understandings of the underlying mechanisms, we have developed a new algorithm ProTECT (i.e. PROtein-protein interactions of Transcription factors predicting Enhancer Contacts with Target genes). In addition to cell-type specific multi-omics data, ProTECT (https://github.com/wangjr03/PPI-based_prediction_enh_gene_links) further integrates the information of PPIs between transcription factors as new features, because TF PPIs have been found to be functionally associated with the regulation of chromatin loops [90-94, 99, 101, 102, 110]. The major steps of ProTECT are summarized in Figure 5.1.e. By creating balanced training sets with confounding factors systematically controlled, ProTECT is trained on cell-type specific chromatin interactions linking distal enhancers and gene promoters. The high-dimensional TF PPI features are hierarchically grouped into feature modules based on a novel graph-based dimension reduction approach. This approach

can simultaneously control the overfitting risk and also reveal the cooperative complexes of TF interactions. Our model demonstrated substantially improved accuracy based on a series of rigorous performance evaluations. Along with genome-wide enhancer-promoter interaction predictions, ProTECT also identifies the key TF PPIs involved in chromatin interaction mediation and prioritizes specific gene sets whose expressions are regulated by distinct TF PPIs.

### 5.2.1 Chromatin contact maps and multi-omics datasets

ProTECT can take different types of chromatin contact maps as input data (Figure 5.1.e), such as Hi-C [10], Capture Hi-C [45] and ChIA-PET [117]. In this study, we used the significant high-resolution Hi-C interactions from human GM12878 and K562 (GEO: GSE63525) [10] to train models for the two cell-lines separately. Enhancer-promoter pairs are labeled as positive samples if overlapping with Hi-C interactions, or are labeled as negative samples otherwise.

Enhancer coordinates are based on Roadmap and ENCODE enhancer annotations [50, 86]. Cell-type specific enhancer activities in GM12878 and K562 cell-lines are quantified using the cell-type specific DNase-seq signals [86]. Other enhancer-associated histone marks, such as H3K27ac or H3K4me1 ChIP-seq data, can also be used to represent enhancer activities and have been found to produce similar predictions in our testing (see Results). Promoters of genes are defined as +/-1kb around transcriptional start sites (TSS), based on gene annotations from GENCODE v17 [149]. Cell-type specific gene expressions are measured by RPKM values of RNA-seq dataset from Roadmap Epigenomics project [86]. Correlation coefficients are calculated for enhancer-gene pairs across diverse cell-types

[50, 86] based on the same set of RNA-seq data for genes and DNase-seq data for enhancers.

The ChIP-seq datasets of transcription factor (TF) bindings in GM12878 and K562 are collected from ENCODE separately [50]. For each TF, if multiple datasets exist, one ChIP-seq dataset is selected based on data quality evaluations (Supplementary Methods). In total, 129 TFs in GM12878 and 270 TFs in K562 cell-lines are included in the analysis (Figure D.1.A). The significant narrow peaks identified by MACS2 [150] are used to label whether a TF binds to a specific genomic location (Figure 5.1.e). Detailed information of all datasets (i.e. TF ChIP-seq, epigenomic signals, transcriptomic data, and chromatin contact maps) are summarized in Supplementary Table 1.

The protein-protein interaction dataset is collected from the STRING database v11 [148]. To remove low-quality PPIs, only PPIs with confidence scores greater than 100 in the 'Experiments' category are included into the analysis. Multiple PPI confidence score thresholds (e.g. 200 and 300) are also tested, which produce similar predictive performance (see Results). The high-quality PPIs are then summarized into a matrix and represented as a PPI network, where every node corresponds to a protein and every edge corresponds to a protein-protein interaction. To account for the intratypic dimerizations of TFs from the Nuclear Receptor (NR), bHLH, and bZIP families, these PPI edges are removed from the PPI network [151] (Supplementary Table 2), because they can only bind locally as dimers. The nodes are further classified into two types: 1) TF protein nodes and 2) non-TF protein nodes. For edges connecting two TF nodes, i.e. TF-TF PPIs, if both TFs are expressed in the specific cell-type, then the TF-TF PPI is considered as active. Therefore, cell-type specificity is assigned for every TF-TF PPI.

non-TF protein nodes are maintained in the PPI network because they are useful to identify indirect TF-TF interactions mediated by non-TF proteins, leading to the discovery of TF PPI modules in subsequent steps.

### 5.2.2 Generation of the training dataset and the matrix of features

In a specific cell-type, enhancer-promoter pairs that overlap with significant Hi-C interactions [10], i.e. the enhancer of the pair overlaps with one of the Hi-C interaction anchors and the promoter overlaps with the other anchor, are labeled as positive samples of enhancer-promoter interactions. As reported by previous studies [119, 152, 153], the data quality of Hi-C interactions whose anchors are located in different topologically associated domains (TADs) are substantially reduced. Therefore, we remove cross-TAD interactions from the analysis, and only use intra-TAD enhancer-promoter interactions, i.e. the interacting enhancer and promoter are located in the same TAD, to train the model.

To avoid biased model training and inflated performance evaluations, we generate a balanced negative set of training samples by randomly selecting the same number of enhancer-promoter pairs that do not overlap with Hi-C interactions. In addition, as pointed out by recent benchmark studies [132], predictions of enhancer-promoter interactions can be substantially biased due to uncontrolled confounding factors. Thus, in the process of generating the balanced random set of negative samples, we strictly control three key confounding factors that have been found to influence the model (Figure 5.1.e): 1) The negative samples of enhancer-promoter pairs should be intra-TAD pairs (Figure D.1.B); 2) The genomic separation distances between the enhancers and promoters follow the same distance distribution of the positive training set. Uncontrolled genomic distances have been found to substantially dominate the models and result in simple short-range

predictions, leading to inflated performance [132, 133]. Using the positive training set of enhancer-promoter pairs, we group them into different genomic distance bins. For each distance bin (bin-size=50kb), we sample the same number of negative enhancer-promoter pairs as observed from the positive set. Therefore, the genomic distance is controlled and the final predictions will not be driven by genomic distances alone (Figure D.1.C-D). 3) The negative enhancer-promoter pairs are sampled for genes which are actively transcribed (Figure D.1.E-F). As demonstrated by previous studies [154], the false negative rates of Hi-C datasets are substantially lower in actively transcribed genomic regions, i.e. more enhancer-promoter interactions can be mapped by Hi-C in active regions compared to repressive genomic regions. To account for this intrinsic bias of Hi-C data, we restrict the sampling of negative enhancer-promoter pairs only from genes whose cell-type specific expression is nonzero (RPKM > 0). By controlling these three key sets of confounding factors, we thus construct the rigorous balanced training dataset for robust model training and performance evaluation. In total, the balanced training dataset contains 5,348 enhancer-promoter pairs in GM12878 and 8,650 enhancer-promoter pairs in K562.

Based on the cell-type specific multi-omics datasets, the matrix of features are then constructed for enhancer-promoter pairs in the training dataset (Figure 5.1.e). There are three types of features incorporated into the model: 1) activity-based features; 2) genomic distance; and 3) TF PPI features. Activity-based features include (i) cell-type specific enhancer activity measured by DNase-seq signals as described above [86]; (ii) cell-type specific gene expression measured by RNA-seq [86]; and (iii) the activity correlations between enhancers and their paired genes calculated from diverse cell-types profiled in

the ENCODE and Roadmap Epigenomics projects [50, 86]. All these activity-based features are differentially distributed across positive and negative training sets, suggesting they are informative to make predictions (Figure D.2.A-C). For each enhancer-gene pair, the genomic distance is calculated as the distance between the center of the enhancer and the gene's TSS. Although they have been controlled in the positive and negative training sets based on genomic bins, there might be residue distance bias within bins. Therefore, the inclusion of genomic distances into the feature matrix captures the residue effects of genomic distances, leading to robust feature prioritization in subsequent analyses.

TF PPIs are the most important set of features for the model because of both the mechanistic relationship with long-range regulation [140, 141, 155] and their significant enrichment in enhancer-promoter interactions (Figure 5.1.b, 5.1.c and Figure D.2.D). In each specific cell-type (i.e. GM12878 or K562 cells), all TFs with available ChIP-seq datasets are collected as described above and compared with the PPI database [148]. From the pool of all candidate pairs, the TF-TF pairs that are capable of forming direct PPIs are considered as TF PPIs. Considering the differences of binding sites in enhancers or promoters, each TF PPI pair is allocated with two directional features. For example, $TF_a$-$TF_b$ represents the PPI between enhancer-binding $TF_a$ and promoter-binding $TF_b$, while $TF_b$-$TF_a$ represents the PPI between enhancer-binding $TF_b$ and promoter-binding $TF_a$. Thus, a set of directional TF PPI features is generated. Because the features are generated only for TFs with cell-type specific ChIP-seq signals, PPIs between TFs that are not active in the specific cell-type do not participate in the predictions. Enhancer-promoter pairs are scanned for TF binding peaks in enhancers and promoters. For each enhancer-promoter pair, if $TF_a$ binds to the enhancer and $TF_b$ binds to the promoter, then

the directional PPI feature $TF_a$-$TF_b$ is labeled as 1. Therefore, a matrix of TF PPI features is constructed for all enhancer-promoter pairs. Combining with the activity-based features and genomic distances, the full matrix of features is then built (Figure 5.1.e).

### 5.2.3 Hierarchical TF community detection on the PPI network

Due to the large number of TF PPI features, dimension reduction is fundamentally important for the construction of robust predictive models. Without dimension reduction, there are 1,888 TF PPI features in GM12878 and 7,066 TF PPI features in K562 cells. Although a number of TF PPIs are enriched in enhancer-promoter interactions (Figure 5.1.b and 5.1.c), direct incorporation of these TF PPI features makes the model to be over-complicated, leading to poor generalization of predictions. To illustrate the significant overfitting issues of direct incorporation of high-dimensional TF PPI features, a basic random forest model is used to test the performance in GM12878 [10]. The features include the activity correlations between enhancers and genes, genomic distances, and 1,888 active TF PPI features. Although the regular 5-fold cross-validation shows an AUC of 0.89, a rigorous genomic-bin split cross-validation (see subsequent sections on cross-validation) shows the unbiased AUC as 0.55, suggesting strong overfitting problems without advanced feature dimension reductions (Figure D.3). Thus, a novel predictive model is needed for predicting long-range enhancer-promoter interactions based on PPI features among transcription factors.

To address the over-fitting problem, we substantially reduce the feature dimensions by hierarchically grouping individual TF PPIs into TF PPI modules based on the topology of the PPI network, while maintaining the predictability of the model (Figure 5.1E). TF PPI modules represent densely connected groups of TFs in the PPI network, and they are

hierarchically organized where smaller PPI modules merge together to form larger modules (Figure D.4). Biologically, using TF PPI modules as features is consistent with the regulatory mechanisms of long-range chromatin loops, because multiple TFs usually interact with each other as protein complexes. Empirically, the biological relevance of TF PPI modules is also supported by the data. As can be seen in Figure D.5, similar to individual TF-TF pairs, a specific subset of TF modules are strongly enriched in enhancer-promoter Hi-C interactions and are strongly supported by PPI connections (p-value=1.39x10$^{-2}$, permutation test).

TF PPI modules are computationally identified from the PPI network [148] using a random-walk based network-community detection approach. The PPI network, including non-TF protein nodes, is modeled as an undirected weighted graph, where the weights on edges are the 'Experiment' PPI scores from the STRING database [148]. Define $W$ as the adjacency matrix of the PPI network, and define the diagonal degree matrix $D$ as $D_{ii} = \sum_j W_{ij}$. Hence, based on the stochastic model of random-walks on graphs [156], the 1-step transition probability from node $i$ to node $j$ is $\frac{W_{ij}}{D_{ii}}$ , and the p-step transition matrix $Trans_p$ can be calculated as $Trans_p = (D^{-1} * W)^P$. Based on the p-step transition matrix, the pairwise distance matrix between TFs (denoted as $R$) can be further calculated as: $R = diag(G)^t * \mathbf{1} + \mathbf{1^t} * diag(G) - 2G$ , where $G = Trans_p * {Trans_p}^t$ . Each entry in the matrix $R$ quantifies the distance between a pair of TFs based on the PPI network structure. Hierarchical clustering is then applied to the pairwise distance matrix $R$ to identify hierarchical PPI modules of TFs (Figure 5.1.e). "wald" method is used in the hierarchical clustering as suggested by previous studies of network-community detections [157]. By testing multiple values (Figure D.4.A and 5.4.b), $p$ is set to be 20 in order to balance the

detection of both local (i.e. small-size) and global (i.e. large-size) modules (Supplementary Methods).

In the constructed hierarchical clustering tree, the leaf nodes are individual TF PPIs. By applying the bottom-up merging strategy on the tree, individual TF PPIs are first grouped into small-size PPI modules, i.e. S-modules, with the maximum size of $S_{max}$. S-modules represent densely connected TFs in the PPI network, corresponding to candidate protein complexes. S-modules are further merged to form large-size PPI modules, i.e. L-modules, with the maximum size of $L_{max}$. L-modules represent larger PPI network components that cover multiple densely connected S-modules. Biologically, L-modules represent candidate groups of highly interacting protein complexes. The maximum sizes for S-modules ($S_{max}$) and L-modules ($L_{max}$) are selected based on the modularity score of the clustering [158] (Figure D.4, Supplementary Methods). The modularity score $Q$ is defined as $Q = \frac{1}{2m} * \sum_{ij} \left(W_{ij} - \frac{k_i k_j}{2m}\right) * \delta(c_i, c_j)$ where $W$ is the adjacency matrix, $k_i$ is the degree of node $i$, $m$ is the total number of edges in the PPI network ($m = \frac{1}{2}\sum_i k_i$), and $c_i$ is the membership assignment to modules for node $i$. Modularity scores are extensively calculated for different choices of maximum module sizes (Figure D.4.C and 5.4.d), because the choice of specific maximum module sizes automatically determines the total number of modules and results in the final module membership assignments. The optimal size of S-modules is selected as the one yielding the maximum modularity score, which guarantees that the generated S-modules represent densely connected TF groups. The optimal size of L-modules is selected as the one corresponding to the elbow point of modularity score curves, leading to the delineation of large-scale PPI components without significant loss of modularity. Compared to Markov Cluster Algorithm, the PPI modules

from our approach demonstrate higher modularity scores and larger module sizes (Figure D.6), which is desired for feature dimension reductions. Using this procedure, a two-layer hierarchical modular structure is finally built and each individual TF PPI is assigned with the memberships belonging to a specific S-module and a specific L-module.

Based on the TF PPI module assignments, individual TF PPI features (i.e. direct TF-TF PPIs) are merged into module-level PPI features, and, therefore, the feature matrix of TF PPIs are restructured accordingly (Figure 5.1.e). There are two types of module-level PPI features: (i) intra-module features, which include all S-modules and L-modules. The intra-module features cover PPIs between TFs within the same modules. (ii) inter-module features, which include inter S-module features and inter L-module features. The inter-module features cover PPIs linking TFs from two different modules. Given a pair of S-modules, e.g. S-module $a$ and S-module $b$, if there exists a TF member from S-module $a$ that has PPI with a TF member from S-module $b$, then the pair of S-modules $a$ and $b$ is included into the feature matrix as one inter S-module PPI feature. The inter L-module PPI features are defined in the same way by checking PPIs of TF members from two L-modules. Each inter-module feature is further split into two directional features, depending on the binding sites of TF members in enhancers and promoters. Using this approach, the PPI features are substantially reduced. For example, the 1,888 individual TF PPI features are reduced to only 78 module-level PPI features in GM12878 and the 7,066 individual TF PPI features are reduced to only 238 module-level PPI features in K562 cells.

The training set of enhancer-promoter pairs are then scanned for module-level PPI features. For each specific enhancer-promoter pair, based on the counts of individual TF

PPI features calculated in the previous step, the counts of module-level PPI features are generated depending on the module memberships of TFs (Figure 5.1.e). For each module-level PPI feature, if multiple TF PPI features are found for an enhancer-promoter pair, the maximum count is used for the module-level feature. Although the number of features is substantially reduced after using module-level PPIs, the specific PPI information is still maintained in this procedure, as shown in Figure D.5. It suggests that the module-based dimension reduction does not cause the loss of information, while substantially reducing the risk of over-fitting.

### 5.2.4 Predictive model of long-range enhancer-promoter interactions

Random forest model is used to predict cell-type specific long-range enhancer-promoter interactions based on the feature matrix constructed above, after module-based dimension reduction (Figure 5.1.e). Random forest model is selected due to its superior performance of handling non-linear feature dependency and its capability of prioritizing the key set of important features for subsequent biological interpretations. As a free model parameter, the number of decision trees in the model is extensively tested with different values, and the accuracy of predictions is found to be robust (Figure D.7).

Additionally, to quantitatively demonstrate the contributions from TF PPIs, we train random forest models based on two versions of input features: 1) the model is trained using only activity-based features and genomic distances; and 2) the full set of features including module-level TF PPI features. The Area Under Curve (AUC) values of cross-validations are calculated for the two versions. The increased AUC from version 2 is the quantitative measurement of the additional information contributed from TF PPIs that is not encoded in activity-based or genomic distance features.

### 5.2.5 Feature selection

In the random forest model, the backward feature elimination approach is used to select useful module-level TF PPI features, where the features with the minimum importance are recursively eliminated from the model. Furthermore, the statistical significance of the directions of TF PPI features are evaluated. As described in the previous section, every module-level PPI feature is split into a pair of two directional features, based on the binding sites of TFs in enhancers or promoters. For example, the feature module $a$ - module $b$ represents the PPI between an enhancer-binding TF member from module $a$ and a promoter-binding TF member from module $b$. Reversely, the feature module $b$ - module $a$ represents the PPI between an enhancer-binding TF member from module $b$ and a promoter-binding TF member from module $a$. Based on the statistical evaluation of the feature directions, insignificant directional features are merged into un-directional features. This feature merging procedure not only reduces the number of features but also reveals the biological roles of TF bindings in the context of different binding orientations.

The determination of whether a pair of directional TF PPI features to be merged into an un-directional feature is a model selection problem. While Akaike Information Criterion (AIC) has been a widely used metric for parametric models, it can not be applied to random forest models, which are non-parametric. Instead, we use the Generalized Degrees of Freedom (GDF) method to calculate a relaxed AIC [159] for the random forest model. GDF is a metric to evaluate the degree of freedoms for Bernoulli distributed data, e.g. the binary labels for enhancer-promoter interactions. And it is defined as $GDF \approx \sum_i(\widehat{y_i}' - \hat{y}_i)/(y'_i - y_i)$, where $y_i$ is the observed label for data point $i$, $y'_i$ is the perturbed

label by inverting $y_i$, i.e. $y_i' = 1 - y_i$, $\widehat{y_i}$ is the predicted label from the model using the unperturbed $y_i$, and $\widehat{y_i'}$ is the predicted label from the model using the perturbed $y_i'$. As suggested by previous studies [159], to calculate GDF, 20% samples are simultaneously perturbed. The relaxed AIC of random forest models are then estimated as $AIC = -2l_m + 2GDF + GDF(GDF+1)/(N - GDF - 1)$, where $N$ represents the total number of data points and $l_m$ represents the goodness-of-fit of the random forest model. As suggested by previous analyses [159], $l_m$ is calculated as the averaged $R^2$ value from 5-fold cross-validations.

For each pair of directional TF PPI features, the relaxed AIC metrics are calculated before and after they are merged into an un-directional feature. If a smaller AIC is observed by merging the two directional features, the model with the merged un-directional feature is then selected, because the reduced AIC suggests the directions of the pair are not statistically important. This procedure is conducted for all pairs of directional TF PPI features, and a final random forest model with the selected features is built. In GM12878 cells, the number of module-level TF PPI features is reduced to 53 from 78. In K562 cells, the number is reduced to 139 from 238. This feature selection process further boosts the generalizability of our model and improves the biological interpretations of the learned TF PPI features (i.e. directional or un-directional).

### 5.2.6 Cross-validation and performance comparison

To evaluate the performance of our model, i.e. Area Under Curve (AUC), we designed a stringent strategy of 5-fold cross-validation. As highlighted by previous studies [132, 133], multiple factors have been found to substantially inflate the performance evaluations and cause overfitting problems. First, the confounding factors (i.e. TAD domain structures,

genomic distances between enhancers and promoters, and gene expression levels) need to be controlled. Otherwise, the performance will be biased and dominated by confounding factors. We addressed this issue in the step of data generation as described in previous sections. Negative samples are randomly generated with the confounding factors controlled to have the same distributions as seen from the positive samples. Second, inflated cross-validation AUC can be found due to the spatially proximal enhancer-promoter pairs across the training and testing datasets [132, 133]. Because TF binding profiles are highly correlated among enhancers and promoters in neighboring genomic regions, proximal enhancer-promoter interactions that are allocated in the testing set will substantially inflate the accuracy. Therefore, random splits of samples based on typical cross-validation may suffer from the dependency of spatially proximal samples allocated in both training and testing sets, as has been noted in previous studies [132, 133]. To address this issue, we developed a genomic bin-split cross-validation approach (Figure 5.1.e). In this approach, the human genome is first divided into consecutive 1Mb bins. In each of the 5-fold cross-validation steps, 80% of the genomic bins are selected as training bins. And the balanced and confounding factor controlled samples of enhancer-promoter pairs from the training bins are used to train the random forest model. The remaining 20% bins are selected as testing bins, and the samples of enhancer-promoter pairs from the testing bins are used to test the model. Using this genomic bin-split cross-validation method, the dependency between training and testing samples are broken and the model performance can be rigorously quantified.

The performance of our model, ProTECT, is compared with two most recent supervised methods that also leverage TF information: IM-PET [127] and TargetFinder [128]. In addition

to activity-based features and genomic distances, IM-PET and TargetFinder also includes the TF binding features in enhancers and promoters, while TargetFinder further incorporates TF binding information in the genomic windows between enhancers and promoters. By comparing with these two algorithms, we can further demonstrate the improved accuracy is obtained purely from the unique features of our model, i.e. the PPIs between TFs.

The stand-alone package of IM-PET (https://github.com/tanlabcode/IM-PET) is applied to the same dataset. Since IM-PET automatically makes predictions for all enhancer-gene pairs with distances <2Mb, only the enhancer-gene pairs overlapping with the dataset are used for performance evaluation, leading to a fair comparison for IM-PET. The TargetFinder software (https://github.com/shwhalen/targetfinder) is also implemented to the same training and testing dataset. The same set of TF ChIP-seq peaks are used to generate the window related features for TargetFinder. 5-fold cross-validation with the same genomic bin-split strategy is applied to remove the potential issues of inflated performance evaluations.

In addition, to quantitatively demonstrate that the improved accuracy of ProTECT is indeed contributed by TF PPI features, we randomly permute the PPIs between TFs, with the degree of each TF in the PPI network unchanged. Furthermore, for every TF, the specific binding sites in enhancers and promoters are also maintained. Therefore, only the TF PPI features are shuffled across enhancer-promoter pairs. The same model training and evaluation procedure are then applied on the permuted dataset. The resulting AUC is then compared to the model trained on the original dataset. This comparison

provides direct evidence on the contributions of TF PPIs to chromatin interaction regulation.

### 5.2.7 Genome-wide prediction of long-range enhancer-promoter interactions

The trained ProTECT algorithm is applied to all enhancer-promoter pairs with genomic distances <2Mb across the whole human genome to make genome-wide predictions of cell-type specific enhancer-promoter interactions (Figure 5.1.e). The features for each candidate enhancer-promoter pair are generated in the same way as described in previous sections. By applying the trained random forest classifier, every candidate enhancer-promoter pair is assigned with a predicted score of interacting with each other. To derive unbiased estimates of the statistical significance for the scores, i.e. p-values, a null distribution of the scores is generated by permuting the feature matrix across enhancer-promoter pairs. This permutation approach effectively maintains the overall abundances of different features in the shuffled dataset. Based on the null distribution, the p-value for each enhancer-promoter pair is then calculated.

Unlike the phase of model training, where the genomic distances are controlled in order to learn specific TF PPI signatures, the phase of genome-wide predictions requires the incorporation of genomic distance information. As shown by chromatin contact maps, e.g. Hi-C datasets, enhancer-promoter pairs with shorter genomic separation distances have higher probability to interact and the probabilities decay as the distances increase (Figure D.1.C). To statistically incorporate the genomic distances based on this prior knowledge, we use the pFDR algorithm [160] to transform p-values into distance-aware q-values. In pFDR, the distribution of distances between Hi-C linked enhancers and promoters is treated as prior probabilities of interactions for enhancer-promoter pairs. Based on Hi-C

data, ProTECT divides the range of distances into consecutive 20kb bins, and the prior probability of interactions for each distance bin is calculated as:

$\pi_i = 5\% * (number\ of\ significant\ Hi-C\ in\ bin_i)/(\text{number of significant Hi} - \text{C in bin}_1)$ , where $\pi_i$ is the prior probability for distance-bin $i$. The prior probability for bin 1 (i.e. the shortest distance bin) is set to be the default 0.05. The pFDR under rejection region $[0, \gamma]$ in distance-bin $i$ is then calculated as $\text{pFDR}(\gamma) = \pi_i \Pr(P \leq \gamma | H = 0)/\Pr(P \leq \gamma) = \pi_i \gamma / \Pr(P \leq \gamma)$, where $P$ represents the p-value for each enhancer-promoter interaction. $P$ follows the uniform distribution under the null hypothesis, i.e. H=0, so that $\Pr(P \leq \gamma | H = 0) = \gamma$. $\Pr(P \leq \gamma)$ can be estimated by $\widehat{\Pr}(P \leq \gamma) = (\sum_{j=1}^{N} \delta(P_j \leq \gamma))/N$, where $P_j$ is the p-value for the enhancer-promoter interaction $j$, $N$ represents the total number of p-values, and $\delta(x)$ equals to 1 if x is true and equals to 0 otherwise. Therefore, the q-values can be calculated as $Q(P) = \inf_{\gamma > P}(\pi_i \gamma / \widehat{\Pr}(P \leq \gamma))$, which combines the information from both the distance-aware prior probabilities ($\pi_i$) and the p-values from the random forest model ($P$). Based on the q-value threshold of 0.05, the final genome-wide predictions of significant enhancer-promoter interactions are obtained.

### 5.2.8 Feature interpretation for mechanistic insights

Using the trained random forest model of ProTECT, we evaluate and rank the importance of features, i.e. the module-level PPI features in the model. The top-ranking module-level PPIs are considered as important features, which represent putative protein complexes that may regulate chromatin interactions. Furthermore, in order to obtain detailed mechanistic understandings of important PPIs between specific TFs, we decode the module-level PPI feature importance into TF-level PPI feature importance. For each prioritized module-level PPI feature, we decompose it into individual TF-TF PPI features,

i.e. specific PPIs between an individual enhancer-binding TF and an individual promoter-binding TF. Then the genome-wide predictions of enhancer-promoter interactions are scanned, and the fractions of predictions that contain the specific TF-level PPI features are calculated. The fractions scanned from genome-wide predictions are highly correlated with the fractions calculated from the cross-validation samples in model training, and are more robust given the larger pool of genome-wide enhancer-promoter pairs (see Results). Using the fractions, the top-ranking TF-level PPI features are thus identified for each important module-level PPI feature. The prioritized features, both module-level and TF-level, shed light on new biological insights on long-range enhancer regulation.

### 5.2.9 Pathway enrichment analysis for genes regulated by specific TF PPIs

To investigate whether chromatin interactions mediated by different TF PPIs may participate in distinct biological pathways, we classify genes based on the specific TF PPI features involved in their interactions with enhancers. For each top-ranking module-level PPI feature, we first identify the top five TF-level PPI features using the method described above. Then, we scan the genome-wide predictions of enhancer-promoter interactions and collect the subset of interactions that contain at least one of the top five TF-level PPI features. Finally, the subset of interactions are ranked by their q-values, and the top 1,000 genes regulated by these interactions are selected. In this way, the prioritized subset of genes represent strong targets of long-range enhancer regulation mediated by the important TF PPIs. Gene Ontology enrichment analyses are performed on different gene sets using DAVID [161] to check whether they are enriched with specific biological pathways.

### 5.2.10 cis-eQTL enrichment analysis for predicted long-range enhancer-promoter interactions

As the orthogonal information to validate the accuracy of genome-wide predictions made by ProTECT, cis-eQTL datasets from the matched human tissues and cell-types are compared with the predicted enhancer-promoter interactions. Because our genome-wide predictions are made in human GM12878 and K562 cells, we selected four eQTL datasets [58-60, 162] which were profiled from either whole blood tissues or lymphoblastoid cells. A predicted enhancer-promoter interaction is considered to be supported by a cis-eQTL (i.e. a significantly associated SNP-gene pair), if the enhancer contains the SNP and the promoter matches with the gene. For each eQTL dataset, the fraction of predicted enhancer-promoter interactions that are supported by cis-eQTLs is calculated, and is compared to two versions of negative controls. The first version of negative control is based on random pairing enhancers with promoters that are within 2Mb distances. The second version of negative control further requires the genomic distances of random enhancer-promoter pairs follow the same distribution from our predicted enhancer-promoter interactions. Therefore, the second version is a more stringent control. For each version, 1,000 random samples are generated. And the statistical significance, i.e. p-values, of the observed overlapping fractions from our predictions is calculated as the portion of random samples showing a higher overlapping fraction than the real observed one.

In addition to cis-eQTLs, we also use cis-hQTLs, i.e. histone QTLs, to evaluate the accuracy of our predictions. The hQTL dataset was also profiled from the human GM12878 cells [62]. Similarly, a predicted enhancer-promoter interaction is considered to

be supported by a cis-hQTL (i.e. a significantly associated SNP-histone pair), if the enhancer contains the SNP and the promoter overlaps with the histone modification peak. The overlapping fraction is also compared with the two versions of negative controls to justify the enrichment of cis-hQTLs in support of our predictions.

### 5.2.11 cis-eQTL enrichment around TF binding sites

For cis-eQTLs that overlap with predicted enhancer-promoter interactions, the genomic locations of the SNPs from cis-eQTLs are further compared with TF binding sites within enhancers. Here, the TF binding sites are defined as the ChIP-seq peak summits. For each enhancer included in this analysis, the TFs involved in important PPI features prioritized from the previous steps are selected. The genomic distances between the SNPs and the binding sites of these TFs are calculated. To statistically test whether the SNPs are closer to these important PPI-related TFs, two versions of random controls are generated. The first version is generated by randomly sampling binding sites of any TFs within the same set of enhancers. And the second version is generated by randomly sampling binding sites of TFs that are members of bottom-ranking PPI features, based on feature importance calculations from the previous sections. For each version of negative controls, p-values are calculated using Kolmogorov-Smirnov tests by comparing the cumulative distributions of distances.

### 5.2.12 trans-eQTL enrichment analysis for enhancer-mediated TF-gene pairs

Compared to cis-eQTLs, trans-eQTLs can provide additional evidence to support the functional associations between the prioritized TFs and specific genes, where the TF's PPIs are predicted to mediate enhancer-promoter interactions of the target genes. For enhancer-binding TFs that are members of the important PPI features, we first collect the

predicted enhancer-promoter interactions mediated by the corresponding PPI features. Genes regulated by these predicted interactions are thus considered as the downstream target genes of the specific enhancer-binding TFs. We define this relationship as enhancer-mediated TF-gene pairs. To exclude the possibility of promoter-mediated effects, we remove the genes whose promoters are also bound by the specific TF.

Using the trans-eQTLs from the published database [163], we identify a subset of trans-eQTLs whose SNPs are located within TF's gene bodies (plus -10kb from TSS) and target genes are covered in our input dataset. For this specific subset of trans-eQTLs, the SNPs are likely to disrupt the transcription of the TF genes, which in turn affects the TF's regulation on the downstream target gene's expression (Supplementary Methods).

Hypergeometric test is used to statistically test whether the enhancer-mediated TF-gene pairs significantly overlap with the subset of trans-eQTLs described above. A TF-gene pair is considered to overlap with a trans-eQTL if the SNP is located within the TF's gene body and the gene is the same as the trans-eQTL's target gene. As comparisons, two versions of controls are generated based on the same set of TFs and enhancers. The first version uses the nearest genes to the enhancers as target genes, instead of using ProTECT's predictions. The second version randomly selects genes within 2Mb distances as target genes. In each version, the same number of enhancer-promoter interactions are generated as seen from the foreground for each sample, and totally 1,000 random samples are created, along with the hypergeometric p-values.

## 5.3 RESULTS

### 5.3.1 Long-range enhancer-promoter interaction prediction based on PPIs among TFs

As discovered by recent experimental studies [93-95, 97-102, 140, 141], the protein-protein interactions between specific transcription factors have been found to participate in the regulation of long-range chromatin loops, where the TFs bind to enhancers and promoters respectively (Figure 5.1.a). The PPIs between the enhancer-binding TFs and promoter-binding TFs facilitate the 3D proximity of enhancers and the target gene's promoters. By analyzing the Hi-C interactions between enhancers and promoters in human GM12878 cells, a specific set of TF-TF pairs are found to be enriched in enhancer-promoter interactions (Figure 5.1.b), compared to their frequencies in distance-controlled random enhancer-promoter pairs. Interestingly, these TF-TF pairs are also enriched with known PPIs (Figure 5.1.c, p-value=$10^{-3}$), suggesting that the TFs within each pair can establish interactions at the protein level. Figure 5.1.d shows two examples, where both enhancer-promoter Hi-C interactions contain enhancer-binding CTCF peaks and promoter-binding RUNX3 peaks. And the physical interaction between RUNX3 and CTCF is validated by the PPI database STRING [148], suggesting the RUNX3-CTCF interaction as a putative mechanism linking the enhancers with specific promoters. These observed enrichments strongly indicate the functional importance of TF PPIs in long-range chromatin loops and the possibility of predicting cell-type specific enhancer-promoter interactions using TF PPI features.

Due to the large number of TF PPI features, i.e. PPIs between enhancer-binding TFs and promoter-binding TFs, basic predictive models significantly suffer from overfitting

problems, as shown in Figure D.3. Therefore, to efficiently leverage the information of TF PPIs from the high-dimensional feature space and overcome the overfitting risks, we developed a new machine learning classifier, ProTECT, to predict cell-type specific long-range enhancer-promoter interactions (Figure 5.1.e). Detailed algorithmic designs have been described in Materials and Methods. Overall, there are four main steps to achieve the final predictions: 1) Generation of the balanced Hi-C based training dataset, along with cell-type specific TF PPI features; 2) Dimension reduction of features based on hierarchical network community detection; 3) Predictive model construction using random forest; and 4) Genome-wide predictions of cell-type specific enhancer-promoter interactions.

As a new predictive model, here we highlight a series of key novelties of ProTECT (see Materials and Methods for details). First, a rigorous method of controlling confounding factors, such as TAD domains, genomic separation distances and gene expression levels, is designed in the steps of data and feature generations. This method efficiently removes the impacts of confounding factors, which are fundamentally important to control as discussed by recent benchmark analyses [132, 133]. Second, the graph-based dimension reduction approach not only addresses the potential risk of overfitting but also facilitates the prioritization of functionally important TF PPIs and TF complexes. Third, a generalized degree of freedom (GDF) technique [159] is incorporated to improve feature selections, leading to new biological understandings of specific TFs. Fourth, a stringent genomic bin-split cross-validation strategy is developed for unbiased and robust performance evaluation. This stringent strategy thoroughly breaks the dependency between the training and testing datasets and avoids the inflated performance estimations that have

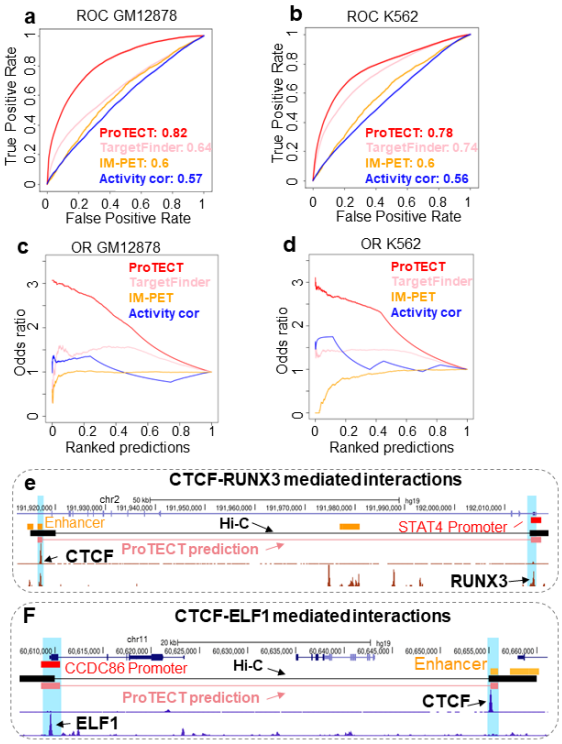



**Figure 5.2 Performance comparison in GM12878 and K562.** (**a**) ProTECT, TargetFinder, and IM-PET are applied on the same input datasets and are evaluated based on the averaged performance of 5-fold genomic-bin split cross-validation. As a baseline comparison, a random forest model using only enhancer-gene activity correlations is also included in the analysis. (**a-b**) ROC curves in GM12878 (A) and K562 (B). (**c-d**) The enrichment of Hi-C interactions in top-ranking predictions. Cumulative odds ratios of true positives (y-axis), i.e. overlapping Hi-C interactions, are calculated across the ranked lists of predictions where predictions with stronger scores are ranked at the top (x-axis), in GM12878 (C) and K562 (D). (**e-f**) Examples of enhancer-promoter interactions predicted by ProTECT (pink paired lines) in GM12878 (E) and K562 (F). In each example, the highlighted enhancer (orange) is predicted to interact with the highlighted promoter (red) by ProTECT. Both predictions are supported by cell-type specific Hi-C interactions (black paired lines). The prioritized TF PPIs mediating the interactions are CTCF-RUNX3 (E) and CTCF-ELF1 (F) respectively, both of which are top-ranking PPI features from the random forest model.

been commonly found in existing methods [132, 133]. Fifth, a genomic distance-aware pFDR procedure [160] is implemented to identify statistically significant enhancer-promoter interactions along the whole human genome. We trained ProTECT using the high-resolution Hi-C datasets from the human GM12878 and K562 cell-lines separately [10]. The

balanced and confounding factor-controlled training dataset contains 5,348 long-range enhancer-promoter interactions in GM12878 and 8,650 interactions in K562 cells. The trained classifiers were further applied to make genome-wide cell-type specific predictions of enhancer-promoter interactions. As shown in subsequent sections, the ProTECT algorithm not only improves the prediction accuracy substantially, but also reveals novel mechanistic insights on the functional roles of TF PPIs in the regulation of long-range chromatin loops. The prioritized TFs and their specific PPIs provide a new platform to understand the complex interplay among TFs, enhancers and genes, and remarkably, open a new avenue to systematically interpret both cis- and trans-eQTLs in human genetics analyses.

### 5.3.2 Boosted performance based on features of TF PPIs

Using the genomic bin-split cross-validation strategy (see Materials and Methods), we rigorously tested the accuracy of ProTECT and compared with the other two supervised methods, i.e. IM-PET[127] and TargetFinder [128]. In both GM12878 and K562 cell-lines, ProTECT achieves the highest performance (Figure 5.2.a and 5.2.b): AUC=0.82 in GM12878 and AUC=0.78 in K562 cells. And the accuracy of ProTECT is robust with respect to the number of trees used in the random forest models (Figure D.7). As comparison, TargetFinder is ranked as the second algorithm with AUC values below 0.74, while the AUC metrics of IM-PET is around 0.6. As a baseline comparison, a random forest model using only activity correlations between enhancers and genes, without using TF PPI features, shows AUC values around 0.57. Because we systematically controlled confounding factors in the training dataset, the AUC estimates are not dominated or biased by those factors, especially the genomic separation distances. Therefore, these

comparisons strongly support that the ProTECT model substantially boosts the prediction accuracy over existing algorithms.

In addition to the overall AUC metrics, to demonstrate that ProTECT has better capabilities of pinpointing true enhancer-promoter interactions in top-ranking predictions, we calculated the cumulative Odds Ratio (OR) of true positives along the ranked list of predictions. As shown in Figure 5.2.c and 5.2.d, ProTECT achieves much higher OR curves than other algorithms, especially in the zone of top-ranking predictions. Because top-ranking predictions are the main de novo discoveries used for experimental studies in practice, this observation further exemplifies the superior precision of ProTECT.

Moreover, we further evaluated the robustness of ProTECT's superior performance with respect to different settings of input features and data. As shown in Figure D.8, by setting different confidence score cutoffs on PPIs to be included as input features (i.e. 100, 200 and 300), ProTECT robustly achieves the highest accuracy (AUC>0.78) compared to other methods. In addition, using different epigenetic signals to represent cell-type specific enhancer activity levels, such as DNase-seq, H3K27ac and H3K4me1, ProTECT demonstrates highly similar accuracy, with DNase-seq and H3K27ac based versions slightly better than the H3K4me1 based version (Figure D.8). Furthermore, we also tested the performance on imbalanced dataset, where the ratio of positive-to-negative samples is 0.1, as suggested by previous studies [127, 128]. ProTECT consistently shows the best ROC and Precision-Recall curves (Figure D.9). To obtain orthogonal evidence on ProTECT's accuracy, we also used a diverse panel of Hi-ChIP [108, 164, 165] and ChIA-PET [50] datasets from the matched cell-types as gold-standards for enhancer-promoter interactions. Remarkably, ProTECT maintains the highest accuracy across all

comparisons based on different gold-standard datasets (Figure D.10 and 5.11). Across the five Hi-ChIP evaluations, ProTECT achieves AUC>0.78, while TargetFinder and IM-PET only show AUC<0.66. Using ChIP-PET datasets as gold-standards, ProTECT achieves AUC>0.84 while other methods demonstrate AUC<0.76. These tests systematically support the robustness of ProTECT's performance advantages.

Figure 5.2.e shows one example predicted by ProTECT in human GM12878 cells. The distal enhancer is located 99.4kb from the predicted target gene's promoter, and this long-range prediction is supported by a cell-type specific Hi-C interaction [10]. Based on the trained random forest model, this enhancer-promoter interaction is mediated by the PPI between the enhancer-binding CTCF and the promoter-binding RUNX3 (Figure 5.2.e). Interestingly, the correlation between the enhancer's activity and the target gene's expression across different cell-types is only 0.28, which strongly suggests the importance of incorporating TF PPI features in predicting enhancer-promoter interactions. A similar example from K562 is shown in Figure 5.2.f, where the distal enhancer is located 46kb from the predicted target gene's promoter, and is also supported by a cell-type specific Hi-C interaction (Figure 5.2.f). This enhancer-promoter interaction, which only shows an activity correlation of 0.261, is successfully predicted based on the PPI between enhancer-binding CTCF and promoter-binding ELF1. Overall, these results demonstrate that TF PPI features can improve the delineation of specific interacting

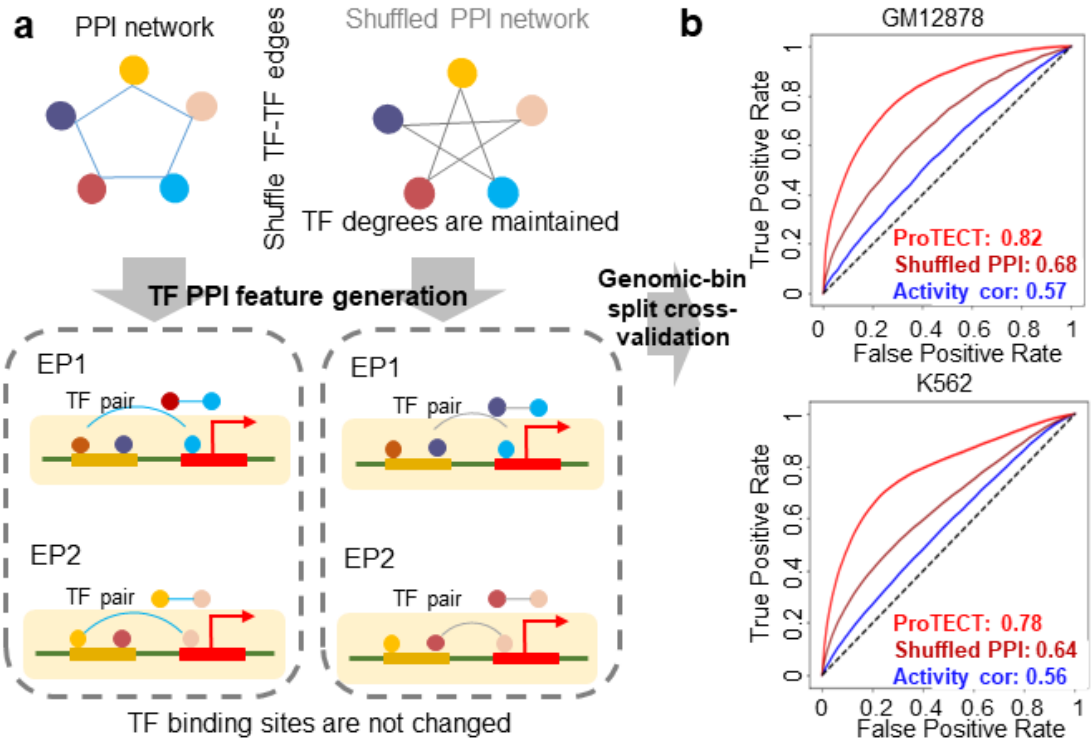



**Figure 5.3 TF PPI features provide additional information beyond TF bindings and activity-based features.** (**a**) Schematic figure of the permutation test on TF PPI features. The shuffled PPIs are generated by randomly pairing two interacting TFs from the original pool of TF PPIs, while the degrees of PPI partners and TF binding sites in enhancers and promoters are maintained. Based on the shuffled PPI features, a new random forest model is trained and then evaluated by the same cross-validation procedure. (**b**) ROC plots for the models based on the original TF PPI features (red), the models based on the shuffled TF PPI features (salmon), and the baseline models based on activity-correlation features alone (blue), in GM12878 and K562 cells.

enhancer-promoter pairs from neighboring non-interacting pairs, beyond the information of activity-related features. In addition, specific hypotheses of the mechanisms mediating chromatin interactions, i.e. the functional TF PPIs linking enhancers and promoters, are derived from the model simultaneously.

To further justify that the superior performance of ProTECT is indeed due to the information from TF PPI features, we randomly shuffled the TF-TF connections in the PPI network (Figure 5.3.a). Therefore, the specific TF binding sites in enhancers and

promoters are strictly maintained (see Materials and Methods), while the PPI features across enhancer-promoter pairs are randomized. This shuffling strategy also controls the degree of PPI partners for each TF, i.e. the number of protein neighbors in the PPI network. By training the ProTECT model on the shuffled data, we found that the accuracy is substantially reduced. The AUC based on PPI-shuffled data is only 0.68, while the original AUC of ProTECT is 0.82 in human GM12878 cells (Figure 5.3.b). Similar decrease of performance is also observed in human K562 cells (Figure 5.3.b). The striking differences of prediction accuracy suggest that the performance improvement of ProTECT is mainly induced by TF PPI features, instead of TF binding information, consistent with previous biological studies of the functional roles of PPIs in chromatin loop regulation [146].

To evaluate the model's dependence on the cell-type specificity of TF bindings, we swapped the TF ChIP-seq data across GM12878 and K562, and run ProTECT based on the swapped data. As expected, the prediction accuracy decreased in both cell-types (Figure D.12.A and 5.12.B), suggesting the necessity of using TF datasets from the matched cell-types. Interestingly, ProTECT still maintains the highest prediction accuracy when other algorithms are also trained on the swapped TF data. In addition, to test the model's dependence on the number of TFs included as features, we obtained the intersection subset of TFs whose ChIP-seq are available in both GM12878 and K562, and trained ProTECT based on features derived from this subset. The cell-type specific predictions in GM12878 and K562 demonstrate similar accuracy (AUC=0.74 and 0.70, Figure D.12.C), suggesting additional TFs are needed in each cell-type beyond the intersection subset.

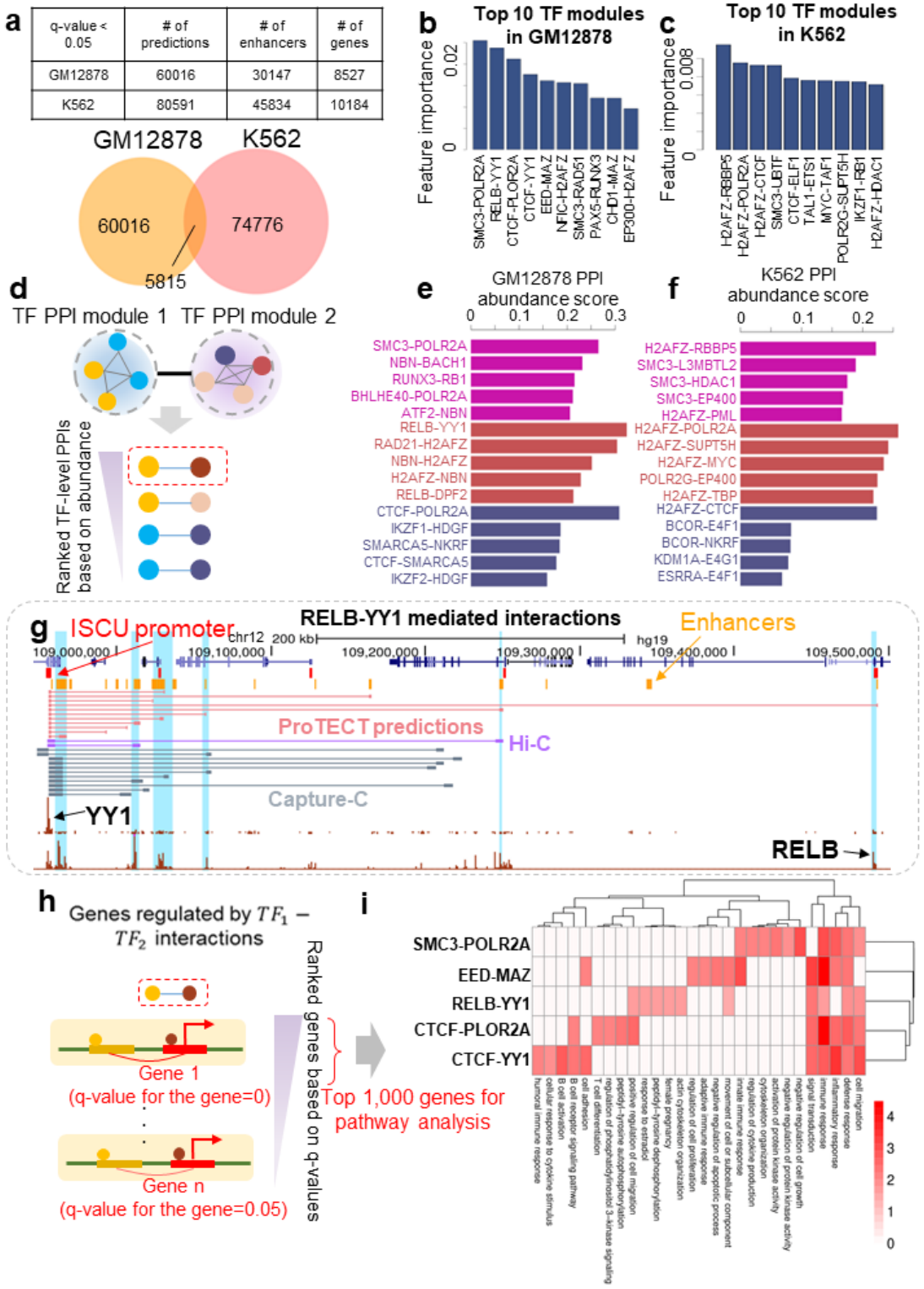



**Figure 5.4 Genome-wide prediction of enhancer-promoter interactions reveals functional roles of TF PPIs in gene regulation.** (**a**) Summary of genome-wide predictions in GM12878 and K562. The venn-diagram shows the overlap between predicted enhancer-promoter interactions in GM12878 (yellow) and K562 (salmon). (**b-c**)

**Figure 5.4 (cont'd)**

Feature importance (y-axis) of top 10 module-level TF PPI features based on the random forest models in GM12878 (B) and K562 (C). Each module-level PPI feature is named by the most abundant TF-level PPIs between the modules as axis-labels (x-axis). (**d**) Schematic figure of ranking specific TF-level PPIs in each PPI module. For each module-level PPI feature, all TF-level PPIs linking two TFs from the pair of two modules (the pair of modules can be the same to represent intra-module TF-level PPIs) are ranked by their occurrences in the predicted long-range enhancer-promoter interactions (abundance scores). (**e-f**) Examples of top 5 TF-level PPIs for three representative module-level features in GM12878 (E) and K562 (F). (**g**) Examples of predicted enhancer-promoter interactions regulated by RELB-YY1 in the ISCU locus. Predicted enhancer-promoter interactions for the ISCU gene are shown as the pink paired lines. Totally 11 enhancers are predicted to interact with the promoter of ISCU, and 5 predictions are supported by Hi-C (purple paired lines) or Capture Hi-C (grey paired lines). ChIP-seq signal tracks of RELB and YY1 (brown signal peaks) are consistent with predictions. (**h**) Schematic figure of ranking enhancer-promoter interactions regulated by specific TF PPIs. For each prioritized TF PPI feature, enhancer-promoter interactions are ranked based on the q-values inferred by ProTECT. Top 1,000 genes are then selected by following the ranked list of interactions for pathway enrichment analysis. (**i**) Pathway enrichments of genes regulated by five different TF PPIs in GM12878. The top 10 most enriched pathways for each TF PPI feature are shown. The heatmap is colored based on the $-\log_{10}$(p value) of pathway enrichments.

### 5.3.3 Genome-wide prediction of long-range enhancer-promoter interactions

The trained random forest model is then applied to the genome-wide dataset in GM12878 and K562 cell-lines separately to predict novel enhancer-promoter interactions (Figure D.13.A-D). All enhancer-promoter pairs within 2Mb distance windows are included into genome-wide predictions (see Materials and Methods), as suggested by observations from experimental Hi-C datasets [10]. For each enhancer-promoter pair, a p-value from the permutation test is generated, which is further used to derive a q-value based on the pFDR approach [160] (see Materials and Methods). Using the q-value threshold of 0.05, there are totally 60,016 significant enhancer-promoter interactions predicted in GM12878, and 80,591 significant enhancer-promoter interactions predicted in K562 (Figure 5.4.a). The median separation genomic distance between linked enhancers and promoters is

243kb in GM12878 (Figure D.13.E), consistent with enhancer's function of long-range regulation. In the predicted GM12878 enhancer-promoter network, >37% of enhancers regulate multiple genes (Figure D.13.F), whose accuracy is consistent with the overall performance (Figure D.14) and 24% of these multi-gene enhancer links are supported by experimental chromatin interactions. On average, every gene is regulated by 6.9 enhancers (Figure D.13.G), suggesting combinations of multiple enhancers are recruited for precise transcriptional regulation. Similar patterns are also observed in the predicted K562 enhancer-promoter network (Figure D.13.H-J). Furthermore, the predicted enhancer-promoter interactions are highly cell-type specific. By comparing the predictions in GM12878 and K562, only 5,815 (~4.2%) enhancer-promoter interactions are shared by the two cell-types (Figure 5.4.a). Compared to the recent activity-by-contact (ABC) model [166], our genome-wide predictions demonstrate higher accuracy, as quantified by both ROC and Precision-Recall curves, using Hi-ChIP data as gold-standards (Figure D.15).

### 5.3.4 Important protein-protein interactions regulating chromatin interactions

To gain insights of the underlying mechanisms of linking distal enhancers to target gene's promoters, we analyzed the feature importance of module-level PPI features inferred by the random forest model and further prioritize the representative TF-level PPI features. We first identified the top-ranking module-level PPI features, which represent the protein complexes of interacting TFs involved in chromatin loops (Figure 5.4.b and 5.4.c). For example, in GM12878 cells, module(CTCF)-module(POLR2A) is ranked as the top 3rd feature (here the module-level features are named by the most abundant TF-level PPIs linking the modules). Interestingly, this is consistent with a recent experimental study [167]

which also found that the enhancer-binding CTCF interacts with the promoter-binding Pol II and participates in the regulation of long-range chromatin loops. As another interesting example, the module-level PPI feature module(IKZF1)-module(RB1) is one of the top-ranking features in K562, consistent with their critical functions in leukemia cells and their impacts on chromatin structure [168, 169]. Additional examples of the prioritized module-level TF PPIs are visualized as PPI networks in Figure D.16, showing the complex PPI connectivity between TF modules binding to enhancers and promoters.

In order to characterize the key PPI features between individual TFs, instead of TF modules, we further decode the module-level PPI features into ranked TF-level PPI features (Figure 5.4.d), based on their occurrences across genome-wide predictions of enhancer-promoter interactions (see Materials and Methods). Genome-wide predictions are used to calculate the abundance scores for TF level PPIs because they provide a large pool of enhancer-promoter links, and the abundance scores are found to be highly correlated with the observations from cross-validation samples (Figure D.17, Spearman Correlation=0.95). For each module-level feature, the top 5 most abundant PPI features between specific enhancer-binding and promoter-binding TFs are identified. For example (Figure 5.4.e), RELB-YY1 is predicted to be a key TF-level PPI feature in long-range enhancer regulation. In support of this new discovery, RELB has recently been found to promote gene expression by interacting with YY1 [170]. As another example, SMC3-HDAC1 is one of the top-ranking features in K562 (Figure 5.4.f), consistent with the reported regulatory roles of HDAC1 on chromatin structure by interacting with SMC3 [171]. The discoveries of these key TFs and their PPIs as candidate functional factors in chromatin

loop formation may lead to new biological hypotheses of enhancer regulation for in-depth experimental investigations.

As a demonstration of the potential importance of TF PPIs in linking distal enhancers to promoters, Figure 5.4.g shows the predicted long-range enhancer-promoter interactions for the gene ISCU. There are totally 11 enhancers predicted by ProTECT to interact with ISCU's promoter, and 5 of them are supported by experimental data of chromatin interactions based on Hi-C or Capture Hi-C (Figure 5.4.g), indicating the high accuracy of the predictive model. The inferred top-ranking feature is the PPI between enhancer-binding RELB and promoter-binding YY1. Consistent with this prediction, YY1 has a strong ChIP-seq binding site at the promoter of ISCU, and almost all linked enhancers have ChIP-seq signals of RELB binding. Importantly, 4 out of the 5 validated enhancers show the strongest RELB ChIP-seq binding signals (Figure 5.4.g), indicating the shared mechanism of these enhancer-promoter interactions for the gene ISCU. In this region, the longest interaction predicted by ProTECT is from a distal enhancer located >547kb from ISCU's promoter. Although not captured by chromatin contact map experiments, this specific enhancer contains a sharp ChIP-seq peak of RELB binding (Figure 5.4.g), suggesting this novel prediction as a strong candidate of enhancer-promoter interactions. It also implies the capability of ProTECT to discover long-range enhancer regulation that might be missed by experimental approaches.

To investigate whether the orientations of PPI features between enhancer-binding and promoter-binding TFs have impacts in chromatin interactions, we designed a systematic model selection strategy to test whether a pair of two TF PPI features with opposite directions can be merged into one un-directional PPI feature without reducing the

predictive accuracy (see Materials and Methods). Using this approach, 32 pairs of directional PPI features in GM12878 are merged into 16 un-directional features, suggesting there is no statistical preference of binding sites (i.e. enhancers vs. promoters) between interacting TFs involved in these PPIs. For example, the features ATF2-SMARCA5 and SMARCA5-ATF2 are merged into an un-directional feature by the model, consistent with the observation that the two directional PPI features have similar abundance in enhancer-promoter interactions (Figure D.18.A). A similar example involves the merge of IKZF1-CREM and CREM-IKZF1 features (Figure D.18.A). In spite of these un-directional PPI features, there are 37 features remaining to be directional in GM12878. For example, there is a significant preference of SMC3-MXI1 feature over the MXI1-SMC3 feature (fold-enrichment=7.80, Figure D.18.B). This is an interesting observation considering the function of SMC3 (a subunit of cohesin [172]) in chromatin structural maintenance, and the reported regulatory function of MXI1 binding in promoter regions [173]. Another example corresponds to the preference of EP300-POL2R2A over POL2R2A-EP300 (fold-enrichment=9.19, Figure D.18.B), consistent with the well-known enhancer binding activities of EP300 [174] and the transcriptional initiation function of POL2R2A [175]. Similarly, 184 pairs of directional PPI features in K562 are merged into 92 un-directional features, while 47 PPI features remain to be directional.

### 5.3.5 Genes regulated by different TF PPIs are enriched in distinct pathways

To evaluate the downstream impacts of chromatin interactions mediated by different TF PPIs, we focused on the top 5 module-level PPI features (Figure 5.4.b and 5.4.c). We identified the strongest enhancer-promoter interactions mediated by each feature separately based on the ranked q-values of predictions (see Materials and Methods).

Genes that are regulated by the top-ranking enhancer-promoter interactions are therefore collected for pathway enrichment analysis (Figure 5.4.h). Overall, these prioritized genes are enriched with immune-related or B-cell-related pathways (Figure D.19.A-B), which is expected since the predictions are inferred from GM12878 and K562 cell-lines. Strikingly, for each specific PPI feature, the gene sets are strongly enriched with distinct groups of pathways (Figure D.19.A-B). Figure 5.4.i shows the most enriched pathways for each TF PPI feature discovered in the GM12878 cell-line. Clearly, the enhancer-promoter interactions mediated by different TF PPIs are enriched with diverse biological processes. For example, the CTCF-YY1 feature is found to be associated with long-range regulation of genes in the B cell receptor signaling pathway, while the SMC3-POLR2A feature is associated with genes of the innate immune response pathway (Figure 5.4.i). To exclude the potential bias caused by gene background, we carried out pathway enrichment analysis based on two additional gene backgrounds, respectively: 1) genes with the same set of promoter-binding TFs; and 2) genes with the same set of enhancer-binding TFs (Figure D.19.C-D). Based on these two rigorous gene backgrounds, the majority (>67%) of enriched pathways are still discovered. These differentially enriched pathways further highlight the functional roles of TF PPIs in regulating gene expression and maintaining the specific cellular states.

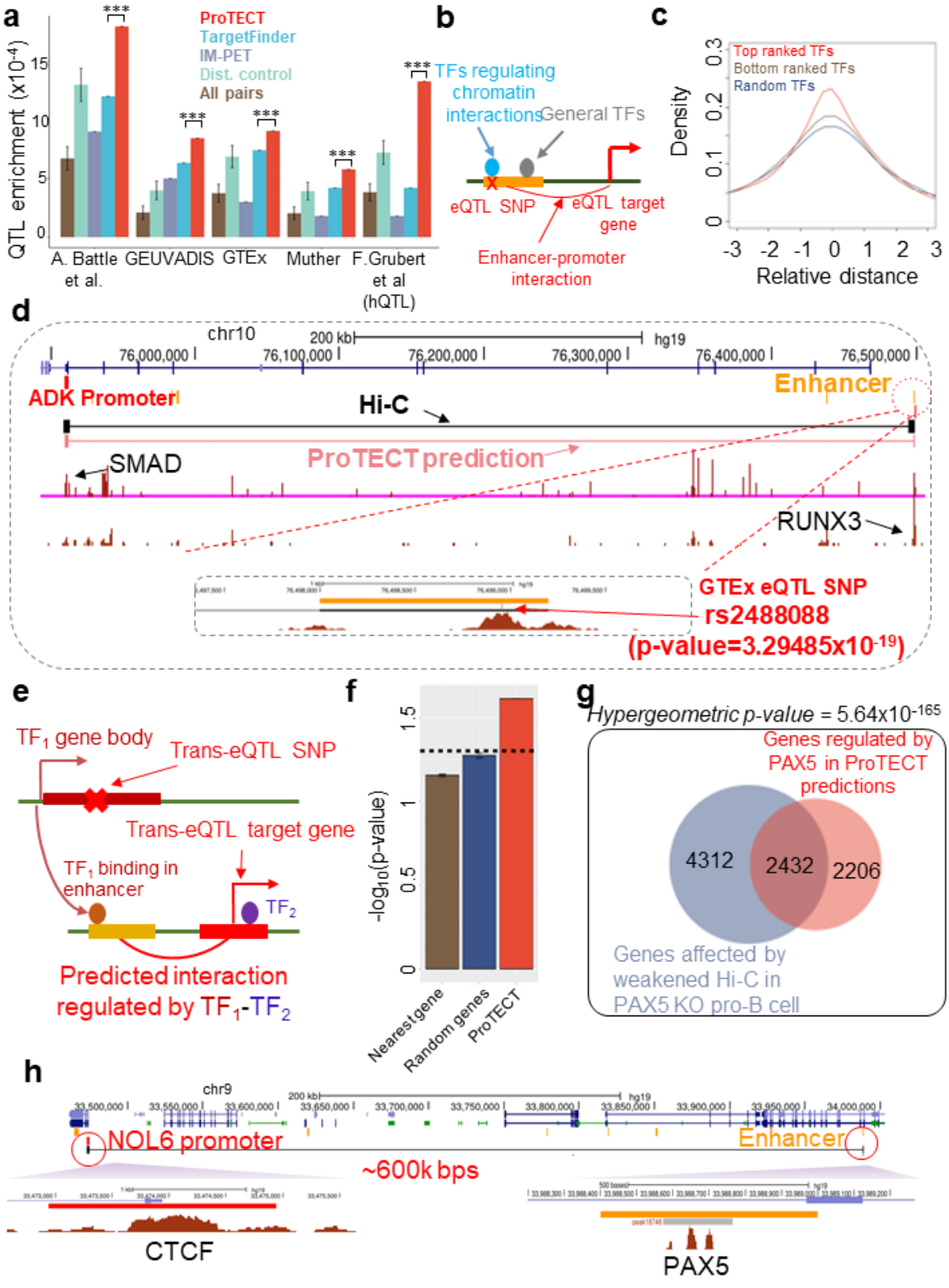



**Figure 5.5 Predicted enhancer-promoter interactions are enriched with cis-QTLs and trans-eQTLs.** (**a**) cis-eQTLs and cis-hQTLs from multiple datasets (x-axis) are

**Figure 5.5 (cont'd)**

significantly enriched in predicted enhancer-promoter interactions in GM12878 (red). The fractions of enhancer-promoter interactions overlapping with cis-QTLs (y-axis) are compared with other methods and two versions of controls: (1) random enhancer-promoter pairs (brown) and (2) distance-controlled random enhancer-promoter pairs (blue). 1,000 samples are generated for both versions to calculate p-values (***: p-value<$1.04 \times 10^{-4}$). Error bars represent sd. (**b**) Schematic figure of cis-eQTL SNPs located in the binding sites of functionally important TFs (blue) of chromatin interactions, compared to general enhancer-binding TFs (grey), as a mechanistic hypothesis of cis-regulatory effects on target gene expression. (**c**) Distributions of relative distances between cis-eQTL SNPs and binding sites of different enhancer-binding TFs. Relative distances (x-axis) are genomic distances between SNPs and TF ChIP-seq peak summits normalized by the sizes of TF peaks. Binding sites of top-ranking TFs inferred by ProTECT (red) significantly overlap with cis-eQTL SNPs, compared with bottom-ranking TFs (grey, p-value=$3.02 \times 10^{-4}$) and random enhancer-binding TFs (blue, p-value=$4.17 \times 10^{-18}$). (**d**) Example of a cis-eQTL, i.e. the rs2488088-ADK pair, overlapping with a predicted enhancer-promoter interaction (pink paired lines). The predicted interaction is supported by Hi-C (black paired lines). The prioritized PPI feature is RUNX3-SMAD, consistent with the ChIP-seq signal tracks (brown signals). Zoom-in view of the distal enhancer (orange) shows the cis-eQTL SNP rs2488088 is located at the peak summit of RUNX3 binding site. (**e**) Schematic figure of trans-eQTL SNPs located in specific TF genes, whose binding to enhancers are predicted to mediate long-range enhancer-promoter interactions of trans-eQTL target genes. (**f**) Hypergeometric test on the overlaps between trans-eQTLs (i.e. trans- SNP-gene pairs) and enhancer-mediated TF-gene pairs, if the SNP is located in the TF's gene body and the trans-eQTL's target gene is the same as the TF's target gene (red, p-value=0.014). The $-\log_{10}$(p-value) (y-axis) from the hypergeometric test is compared to two versions of controls: 1) nearest genes to the enhancers (brown); and 2) random target genes (blue). Each control is generated 1,000 times and the error bars show the sd. The black dash line corresponds to $-\log_{10}(0.05)$. (**g**) Venn diagram comparing genes affected by weakened Hi-C interactions in PAX5 KO pro-B cells and genes regulated by PAX5 in ProTECT predictions (Hypergeometric test, p-value=$5.64 \times 10^{-165}$). (**h**) Example of a trans-eQTL, i.e. rs10973104-NOL6 pair, supported by the predicted enhancer-mediated PAX5-NOL6 pair. The predicted enhancer-promoter interaction for NOL6 (black paired lines) is based on the prioritized TF PPI feature PAX5-CTCF. ChIP-seq signals (brown signal tracks) show a strong CTCF peak in the NOL6 promoter (red) and strong PAX5 peaks in the linked enhancer (orange). The trans-eQTL SNP rs10973104 is located in the gene body of PAX5, which is 3.6Mb away from this locus.

### 5.3.6 Predicted enhancer-promoter interactions are enriched with cis-eQTLs

Because the predictive model is trained on Hi-C datasets, we use cis-eQTLs as orthogonal evidence to quantitatively evaluate the accuracy of the genome-wide

predictions of enhancer-promoter interactions. By comparing the predictions with the SNP-gene pairs of significant eQTLs, we calculated the overlapping enrichment scores (see Materials and Methods). Using four eQTL datasets generated from matched cell-types or tissues (e.g. whole blood tissues or lymphoblastoid cell-lines) [58-60, 162], the predicted enhancer-promoter interactions in GM12878 cell-line show significantly higher fractions overlapping with eQTLs, compared to stringent distance-controlled random interactions and other algorithms (p-value<$1.04 \times 10^{-4}$, Figure 5.5.a). Similar, but relatively weaker, enrichment with eQTLs is found for predictions in K562 cell-line (Figure D.20.A). In addition to cis-eQTLs, we compared our predictions in GM12878 with histone-QTLs from the same cell-line [62] and also observed strong enrichment (p-value=$3.27 \times 10^{-5}$) compared to distance-controlled random samples and other algorithms (Figure 5.5.a). These observations not only support the high accuracy of genome-wide predictions but also suggest the putative mechanisms of cis-eQTLs mediated by chromatin interactions between regulatory elements and target genes.

### 5.3.7 cis-eQTLs are enriched in binding sites of prioritized TFs

The prioritized TF PPI features by the ProTECT model provides a new metric of delineating functionally important TFs for enhancer regulation against general enhancer-binding TFs, which is complicated due to the large array of TFs binding to enhancers. For a typical enhancer, it contains 10 different TF binding sites on average, based on the counts of TF ChIP-seq peaks in GM12878 from the ENCODE project [50]. However, binding itself is not sufficient to assign functional importance for TFs. As found by previous studies, TFs binding in enhancer regions are not equally important for the function of enhancers, with many enhancer-binding TFs lacking evidence of regulatory impacts on gene

expression [176]. This ambiguity hinders the understanding of enhancer activation and downstream effects. We hypothesized the TFs involved with top prioritized PPI features are more likely to be functional for enhancers. We tested this hypothesis by checking the enrichment of cis-eQTL SNPs within the binding sites of the prioritized TFs in enhancers (Figure 5.5.b, see Materials and Methods). The cis-eQTLs are called in whole blood tissues from the GTEx project [162]. Interestingly, the SNPs of cis-eQTLs are located significantly closer to the binding sites of prioritized TFs in GM12878 (p-value=$4.17 \times 10^{-18}$, Kolmogorov-Smirnov test), compared to the binding sites of other adjacent enhancer-binding TFs (Figure 5.5.c). To control the potential bias caused by data availability, we also generated a more stringent background only using TFs included in the model but inferred with low feature importance (see Materials and Methods). Compared with this new background, the prioritized TFs are still significantly enriched with cis-eQTL SNPs (p-value=$3.02 \times 10^{-4}$, Kolmogorov-Smirnov test, Figure 5.5.c). In the K562 cell-line, cis-eQTL SNPs are also closer to the binding sites of the prioritized TFs but not statistically significant (Figure D.20.B). Overall, this analysis supports the stronger regulatory effects of prioritized TFs whose PPIs may mediate long-range enhancer-promoter interactions. Additionally, the prioritized TF binding sites provide a new layer of information to pinpoint regulatory SNPs at a higher resolution, by dissecting the ambiguity of numerous TF bindings within enhancers.

As a representative example, a distal enhancer located >589kb away is predicted by ProTECT to interact with the promoter of the ADK gene in GM12878 (Figure 5.5.d), which is supported by experimental Hi-C data [10]. This long-range interaction is also supported by a significant eQTL, i.e. rs2488088-ADK (p-value=$3.29 \times 10^{-19}$) [162]. The prioritized TF

PPI feature for this interaction is RUNX3-SMAD, where RUNX3 binds to the enhancer and SMAD binds to the promoter. By zooming into the enhancer element, which is 1.2kb long and contains binding sites of 5 different TFs, the SNP rs2488088 is found to be precisely located at the ChIP-seq peak summit of RUNX3 (Figure 5.5.d), consistent with our prioritization of RUNX3 as the important TF for this enhancer. This observation also implies the mechanistic interpretation of this non-coding SNP, whose disruptive effect on the RUNX3 binding causes the loss of RUNX3-SMAD mediated long-range interaction to ADK.

### 5.3.8 trans-eQTLs are enriched in enhancer-mediated TF-gene pairs

As one of the advantages of the ProTECT algorithm, both cis-regulatory elements (i.e. enhancers) and trans-regulatory factors (i.e. TFs) are jointly modeled in long-range chromatin interactions. In traditional studies of trans-regulation of gene expression, analyses have been mainly limited to promoter-binding TFs as candidate trans-regulatory factors [177, 178]. Based on the functional impacts of the predicted important TF PPI features (Figure 5.4.b-l) and the observed enrichment of cis-eQTL SNPs in prioritized enhancer-binding TFs (Figure 5.5.b-D), we hypothesized that there is an enhancer-mediated pathway of trans-regulation, i.e. the enhancer-binding TFs associated with top-ranking PPI features for long-range chromatin interactions are trans-regulatory factors for the expression of distal target genes (Figure 5.5.e). To quantitatively validate this hypothesis, we compared the enhancer-mediated TF-gene pairs with significant trans-eQTLs [163], and the significance of overlaps are statistically tested using Hypergeometric tests (see Materials and Methods). Interestingly, the enhancer-mediated TF-gene pairs are found to be strongly supported by trans-eQTLs (p-value=0.014, Figure 5.5.f, Figure D.20.C),

suggesting that the SNPs of trans-eQTLs are associated with target gene's expression via the disruption of the TF gene's activity (Figure 5.5.e), although the SNPs may be located far away from the target genes or even located in different chromosomes. The observed statistical significance is also stronger than two versions of controls, excluding the potential confounding effects of biased enhancer activity and genomic distances (Figure 5.5.f, see Materials and Methods).

To obtain additional experimental evidence on the predicted enhancer-mediated TF-gene regulation, we leveraged a differential Hi-C interaction dataset in mouse pro-B cells where 7,810 weakened Hi-C interactions were identified following PAX5 knock-out [179]. The top-ranking PAX5 related PPI feature predicted by ProTECT is PAX5-CTCF, consistent with their collaborative roles in B cells [180, 181]. Based on our genome-wide predictions in GM12878, we identified the subset of PAX5-CTCF mediated enhancer-promoter interactions (see Materials and Methods), and thus collected the enhancer-mediated target genes of PAX5. To purify the subsequent analysis, genes whose promoters are also bound by PAX5 are removed from the list. If PAX5 is a true trans-regulatory factor for these genes, the genes are expected to be targeted by the weakened long-range interactions following PAX5 knock-out. By mapping the genes to their homology in the mouse genome [182], 6,744 enhancer-mediated target genes of PAX5 are conserved. Strikingly, these genes are found to significantly overlap with the genes of weakened Hi-C interactions in PAX5-/- pro-B cells [179] (hypergeometric p-value=$5.64 \times 10^{-165}$, Figure 5.5.g). To control the potentially biased enhancer activity and TF bindings, we generated two versions of controls. The first version randomly selects genes as enhancer-mediated target genes of PAX5. And the second version randomly chooses target genes of other

TFs. 1,000 random samples are generated for each version and the same number of genes are selected for each sample. Both versions of negative controls show decreased overlap with genes of weakened Hi-C interactions in PAX5-/- pro-B cells (p-value=$10^{-3}$), supporting the predicted trans-regulatory links between PAX5 and target genes by ProTECT. Figure 5.5.h shows one representative example of PAX5-CTCF mediated long-range enhancer-promoter interaction (~600kb), where the enhancer contains multiple PAX5 binding sites and the promoter of the target gene, i.e. NOL6, contains a strong CTCF binding site. Interestingly, NOL6 is linked with weakened Hi-C interactions in PAX5-/- pro-B cells. These strong experimental validations, along with the enrichment of trans-eQTLs, suggest the biological validity of the predicted enhancer-mediated TF-gene pairs, and provide a new regulatory mechanism to discover and interpret trans- regulatory genetic variants.

## 5.4 DISCUSSION

In this study, we have developed a novel supervised algorithm, ProTECT (https://github.com/wangjr03/PPI-based_prediction_enh_gene_links), to predict long-range enhancer-promoter interactions. By incorporating new features of protein-protein interactions among transcription factors, the algorithm achieves superior performance compared to other methods, based on a rigorously designed genomic bin-split cross-validation procedure. Considering the overfitting risk of high-dimensional inter-dependent TF PPI features, a novel network-community based dimension reduction strategy is used to hierarchically organize TF PPIs into module-level features. This approach efficiently improves the generalizability of the predictive model to make robust predictions based on complex TF PPI patterns, while maintaining the detailed ranking of TF-level PPI features

for specific mechanistic understandings of long-range enhancer regulation. With the impacts of confounding factors strictly controlled, the relative contributions of different features are systematically evaluated, which shows that TF PPIs contain substantially additional information beyond activity-based features of enhancers and genes.

The genome-wide implementation of ProTECT in GM12878 and K562 cell-lines generated 60,016 and 80,591 new predictions of significant enhancer-promoter interactions, which will be useful resources of cell-type specific enhancer regulation for biologists. In addition, a set of prioritized TF PPIs, in both module-level and TF-level, are identified as the key PPIs mediating long-range chromatin loops. Different TF PPIs are found to mediate enhancer regulation for genes in distinct biological pathways, implying specific functional roles of complex TF cooperation. The TF members participating in these prioritized PPI features can be used as candidate targets for knock-out to investigate the changes of specific enhancer-promoter interactions, which will expand the insights on the underlying mechanisms of chromatin loop formation and long-range gene regulation.

To gain orthogonal evidence of the validity of genome-wide predictions, cis- and trans-eQTLs are compared with the predicted enhancer-promoter interactions in three ways, each of which supports one aspect of the interplay among TFs, enhancers and genes. First, the enrichment of overlaps between cis-eQTLs and enhancer-promoter interactions suggests the accuracy of predicted long-range cis-regulation by distal enhancers. Second, the enrichment of cis-eQTL SNPs located within the binding sites of prioritized TFs underscores the precise delineation of functionally important TFs for enhancer activities against other general enhancer-binding TFs. Third, the enrichment of overlaps between

trans-eQTLs and enhancer-mediated TF-gene pairs highlights the novel identification of trans-regulatory pathways from upstream TFs to downstream genes via distal enhancers. The promising enrichment analyses further indicate that the predictions from ProTECT can be used as a platform to interpret cis- and trans-eQTLs, i.e. characterize the non-coding SNP's disruptive effects propagated through long-range enhancer regulation on gene expression. Therefore, combined with eQTL datasets, the ProTECT model can also be a useful tool to generate testable hypotheses in statistical genetics studies.

To control the model complexity, only direct PPIs between TFs are included as features, while indirect PPIs between TFs may also participate in the regulation of chromatin loops. For example, an enhancer-binding TF and a promoter-binding TF may not be able to interact with each other but they both can interact with a third protein. The incorporation of module-level TF PPI features helps to capture the potential indirect PPIs to some degree, but does not explicitly address this problem. Due to the large number of indirect PPI features and the limited number of labeled samples for model training, more advanced designs of feature selection will be needed to achieve a balance between predictive accuracy and model generalizability.

As a major novelty of the ProTECT model, the efficient inclusion of TF PPIs as features not only improves the predictions but also reveals mechanistic insights on long-range enhancer regulation. In the meantime, the algorithm requires the availability of large panels of TF ChIP-seq data for the specific cell-types under study, which may be a practical challenge for users. As one of the directions to extend the ProTECT model, it is possible to leverage the combined information of chromatin accessibility data, e.g. DNase-seq or ATAC-seq data, and TF binding motif annotation datasets as

approximations for cell-type specific TF bindings. Several recent studies have demonstrated the reasonable accuracy of this approximation [50, 86]. Furthermore, multiple imputation algorithms have been recently developed for ENCODE cell-types or tissues to impute cell-type specific TF binding ChIP-seq signals [183, 184]. The imputed TF binding signals can be used as alternative inputs for the model to make cell-type specific predictions of enhancer-promoter interactions, for cell-types lacking ChIP-seq datasets. As an evaluation of this possibility, we generated the imputed TF bindings by overlapping TF motifs with cell-type specific DNase-seq peaks, and then derived TF PPI features based on the imputed data. Remarkably, applied on the imputation-based input features, ProTECT is able to achieve high accuracy (Figure D.21). This evaluation strongly supports the wide applicability of ProTECT on diverse cell-types even if TF ChIP-seq data is not directly available.

# CHAPTER 6

# DISCUSSION

Characterizing the high-order chromatin conformation and complex interplays between TFs, enhancers, and genes in the 3D space play an important role in understanding the complex gene regulations. In this dissertation, we showed two directions to fully delineate the interaction landscape based on the multi-omics datasets. We reconstructed the 3D chromosome structures from the chromatin contact maps and single-cell chromatin conformation capture datasets, which provide the structural basis of the long-range chromatin interactions. We also developed two computational algorithms to predict the long-range enhancer-gene regulations based on the TF bindings. Notably, we predicted the multi-enhancer regulations with high accuracy, which expanded the analyses of gene regulations from one enhancer to the cooperation of multiple enhancers. This chapter summarizes the results, biological innovations, and future directions of our work.

## 6.1 SUMMARY

We first reconstruct the 3D chromosome structures from the chromatin contact maps based on the completion of the low-rank matrix. Our developed algorithm, FLAMINGO, demonstrated high accuracy and scalability in reconstructing high-resolution 3D structures from sparse chromatin contact maps. Using FLAMINGO, we successfully predicted the 3D structures of all 23 chromosomes in 5kb and 1kb resolution, which is the highest resolution for now. Based on the extensive evaluation of the simulated data and orthogonal biological evidence, FLAMINGO demonstrated superior performance over existing algorithms. The 3D chromosome structures predicted by FLAMINGO innovate

the interpretation of the long-range QTLs and multi-way interactions, where chromatin loops bring anchors into proximal 3D neighborhoods and facilitate long-range functional interactions. An integrative variant of FLAMINGO, iFLAMINGO, is further developed to facilitate the cross-cell-type prediction of the 3D structures and refine the resolution. The development of FLAMINGO provides a powerful tool to delineate the interaction landscape in high resolution.

We further developed tFLAMINGO to predict the single-cell 3D chromosome structures. To mitigate the high missing rate of single-cell datasets, tFLAMINGO utilized a low-rank tensor completion method. Compared with existing algorithms, tFLAMINGO demonstrated superior performance in reconstructing single-cell 3D structures and imputing the chromatin contact maps. Given the complete single-cell 3D chromosome structures, we proved the 3D chromosome structures are robust in low-resolution but highly dynamic in terms of single-cell chromatin interactions. For example, TADs are overall robust but could be shifting, merging, and vanishing across single cells. We showed that the genomic loci with critical biological functions, e.g. open chromatin and active transcription, tend to be densely organized in the 3D space and less dynamic across single cells. The delineation of the single-cell 3D chromosome structures also provides a new approach to interpreting the somatic mutations, GWAS SNPs, and predicting the dynamic multi-way chromatin interactions.

To computationally predict the long-range enhancer-gene interactions, we developed an unsupervised learning method, ComMUTE, to integrate the gene regulatory grammar and enhancer-binding TF profiles. The unsupervised framework of ComMUTE largely expands its usability in cell types without experimental chromatin interactions and avoids

the overfitting risks. Compared with existing algorithms, ComMUTE simultaneously links multiple enhancers with synergistic regulatory functions to the same target gene, which captures the multi-enhancer regulations. By extensively benchmarked with existing algorithms, ComMUTE demonstrated consistently improved performance in predicting enhancer-gene links and multi-enhancer regulations. The decoded high-order regulatory landscape shed light on understanding the eQTLs and GWAS SNPs. Strikingly, the multi-enhancer regulations predicted by ComMUTE can help predict the epistasis eQTLs, whose discovery is important for gene regulations but highly challenging due to the unrealistically large searching space. We proposed that the SNPs within the co-regulating enhancers should have a higher probability of being the epistasis eQTLs and thus significantly reduce the number of tests.

In addition to ComMUTE, we also developed supervised learning, ProTECT, to predict the PPI-mediated enhancer-gene links. In addition to standard features used by other methods, we included a new set of features: the PPI between enhancer-binding TFs and promoter-binding TFs. Based on the permutation test, we proved that the new features can boost the accuracy in predicting the enhancer-gene links. We also developed a graph-based dimension reduction method and feature selection approach to avoid overfitting risks. Besides predicting enhancer-gene links, ProTECT also prioritized important TF-TF interactions to establish long-range regulatory interactions. Such predictions can help interpret the trans-eQTLs, where the SNPs and target genes are very far or even on the different chromosomes. Based on the global evaluations and case studies, SNPs may disrupt the important TF regulators and block the enhancer-gene links, thus indirectly controlling the distal target gene expressions.

In summary, the development of the four algorithms fully characterizes the high-resolution 3D chromosome structures in bulk tissue and single cells and depicts the more detailed enhancer-gene regulatory interactions across diverse cell types. The rich predictions and algorithmic advancements of these methods provide a solid foundation for future studies of the complex biological events in the 3D space.

## 6.2 FUTURE DIRECTION

Although we captured the dynamic chromosome structures across single cells, investigating the cell-type-specific structures in single cells is still challenging since the single-cell chromatin conformation capture datasets are few. Therefore, an important feature of the desired algorithm is building a connection between the cell-types-specific epigenomic signals and 3D spatial distances between genomic loci. We will continue the study of the underlying driving force in shaping the 3D chromosome structures.

Another important direction is utilizing the predictions of these methods to predict downstream biological events, for example, gene expressions, TF binding sites, and disease-associated genes. We will further integrate the 3D chromosome structures and enhancer-gene regulatory interactions into the following algorithms to improve the model performance and gain better mechanistic insights

Another important direction is utilizing the predictions of these methods to predict downstream biological events, for example, gene expressions, TF binding sites and disease associated genes. We will further integrate the 3D chromosome structures and enhancer-gene regulatory interactions into following algorithms to improve the model performance and gain better mechanistic insights

# APPENDICES

## APPENDIX A

## SUPPLEMENTARY FIGURES FOR CHAPTER 2

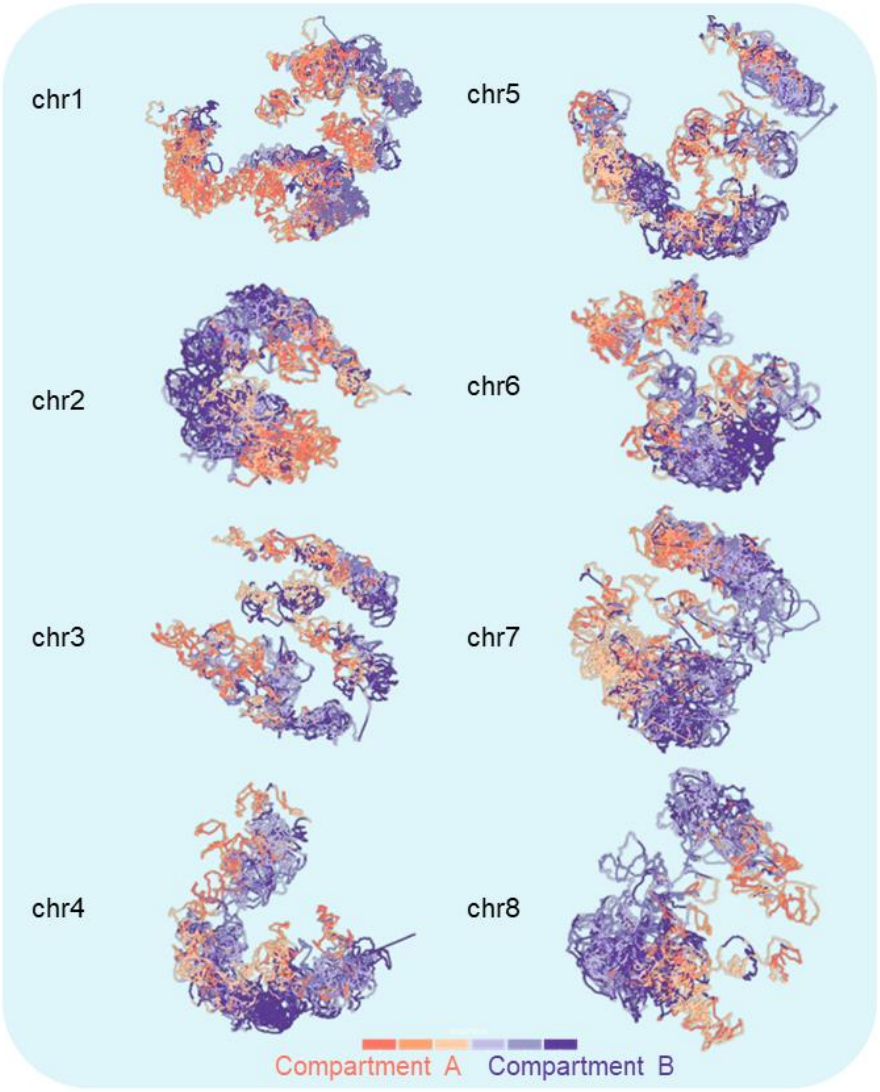



**Figure A.1 5kb-resolution 3D structures for 23 chromosomes predicted by FLAMINGO.**

**Figure A.1 (cont'd)**

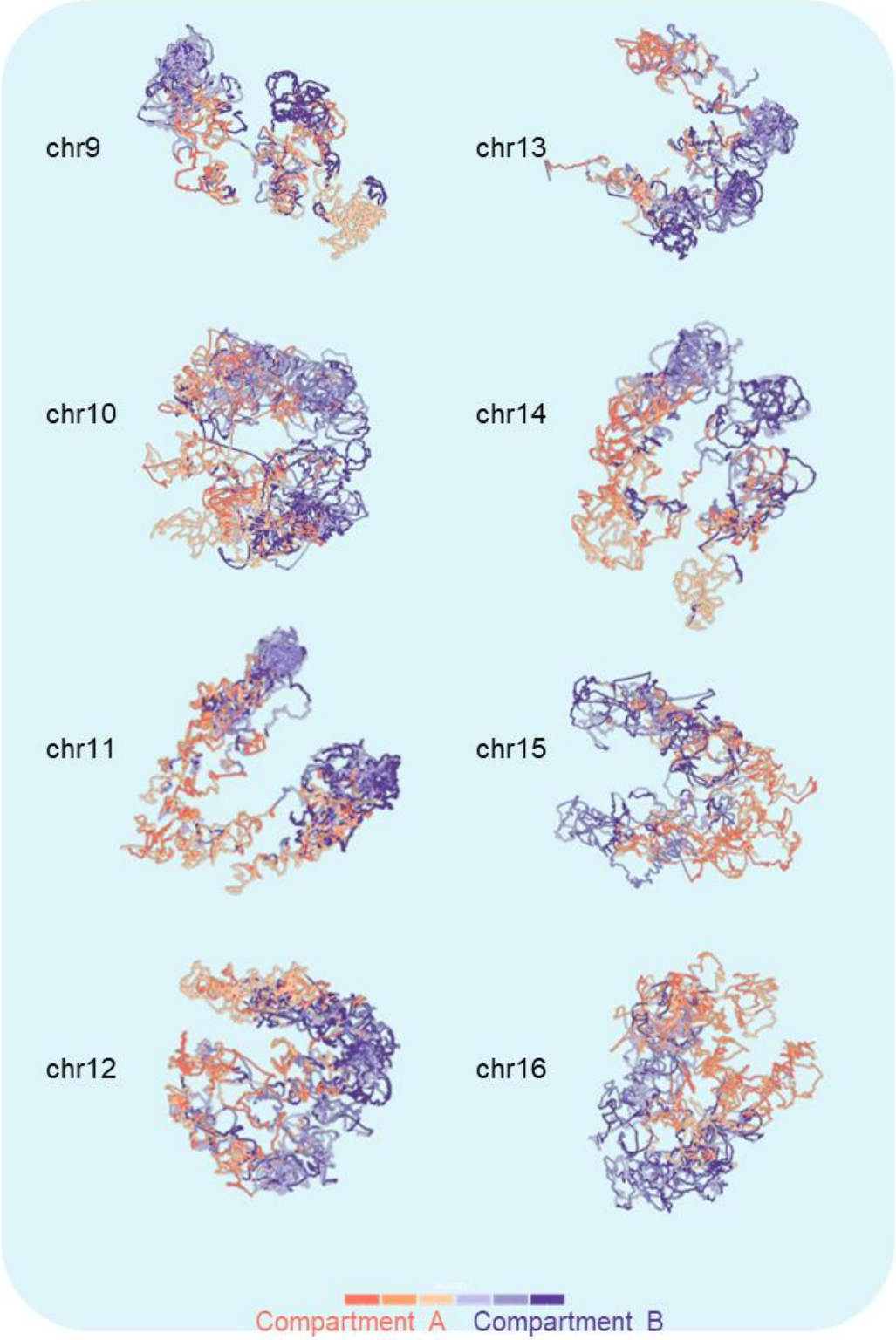
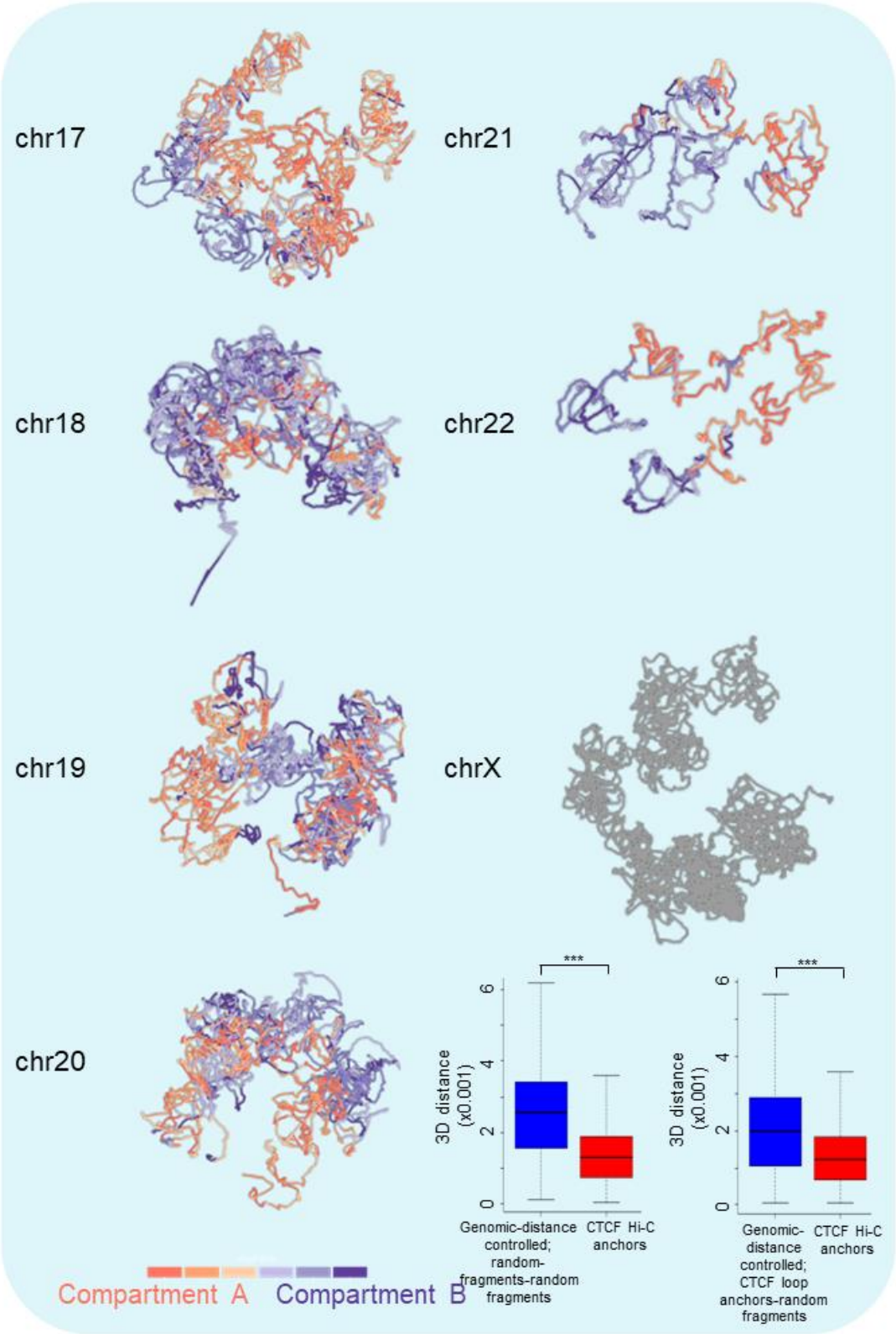


**Figure A.1 (cont'd)**



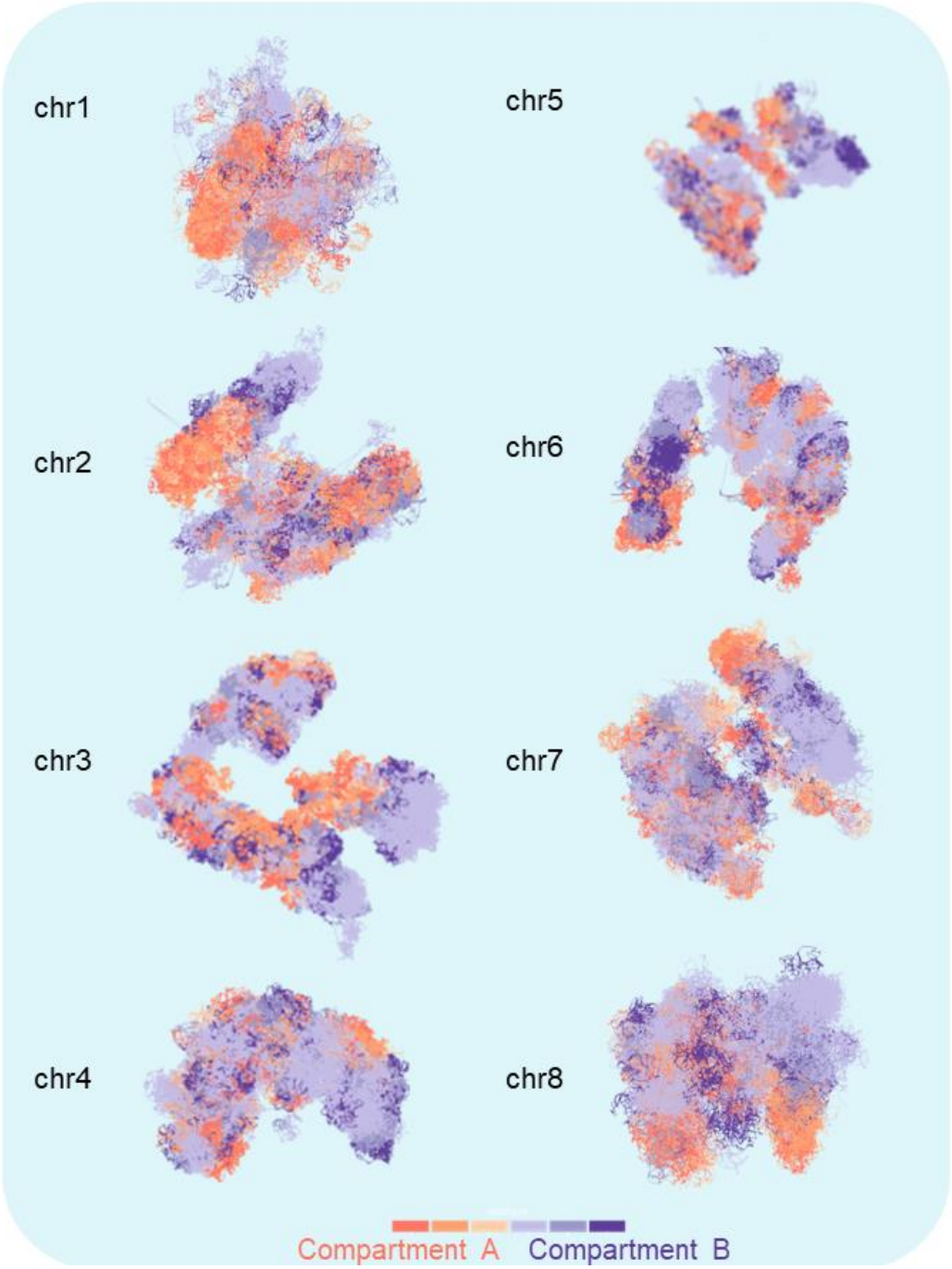



**Figure A.2 1kb-resolution 3D structures for 23 chromosomes predicted by FLAMINGO.**

**Figure A.2 (cont'd)**

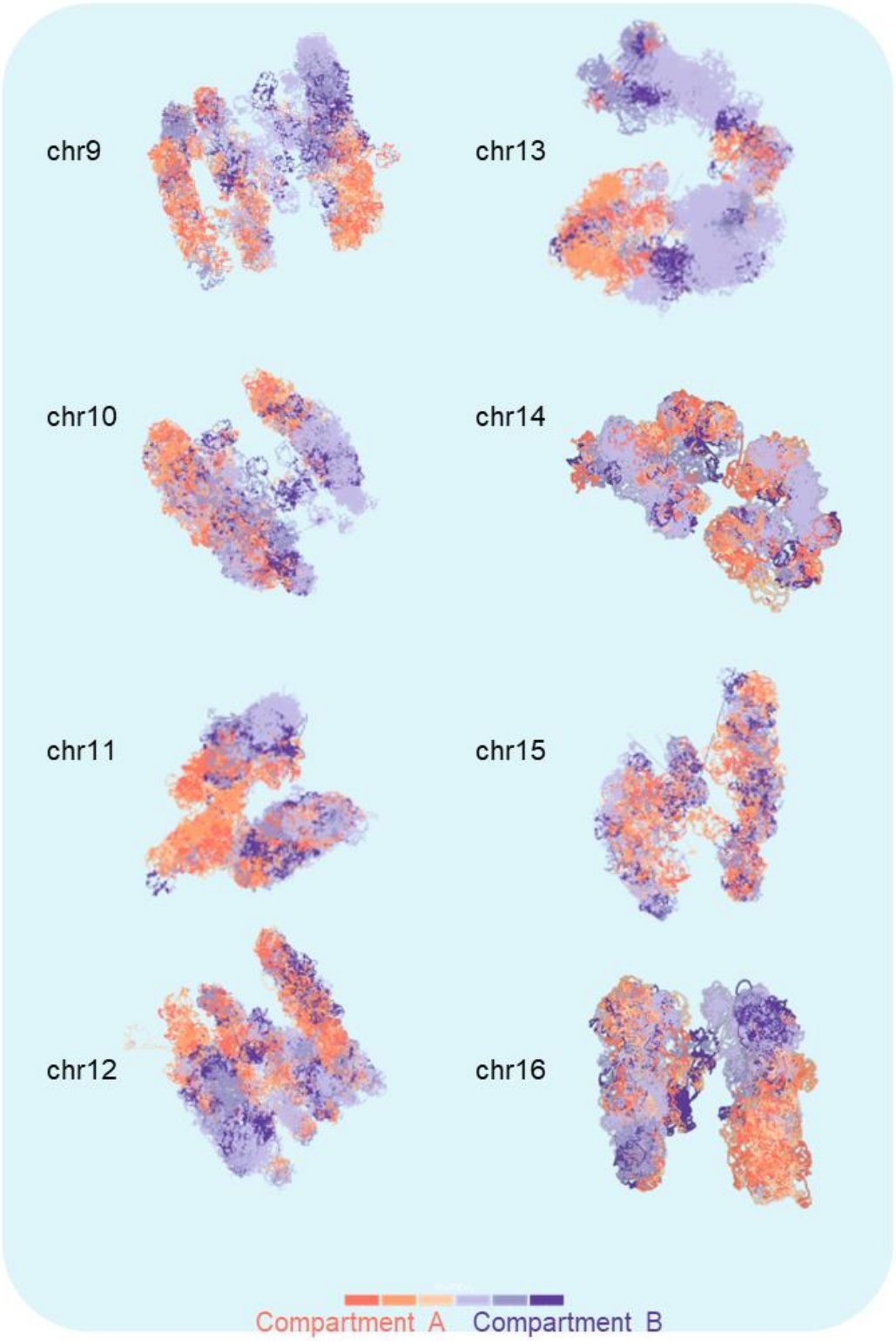
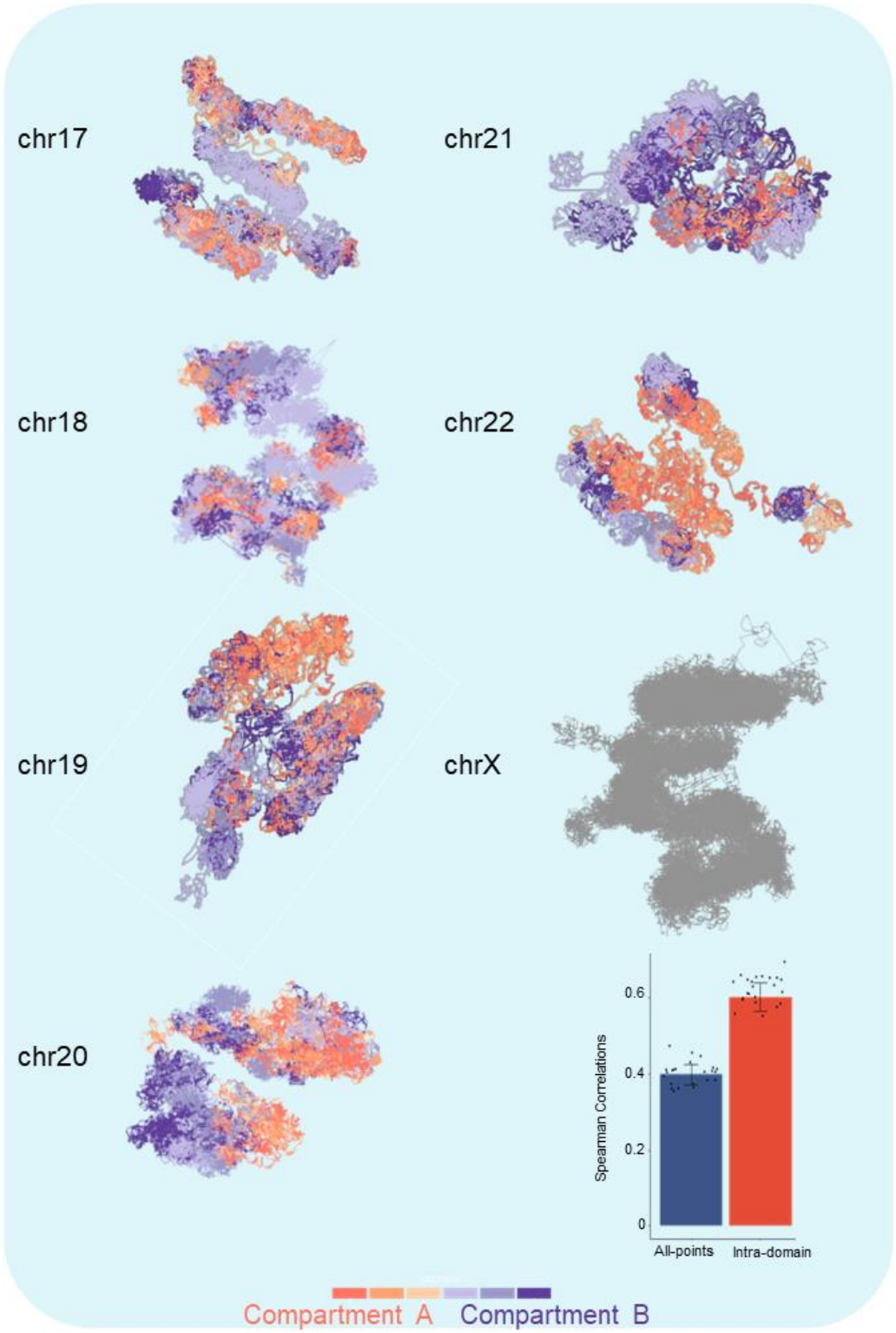


**Figure A.2 (cont'd)**



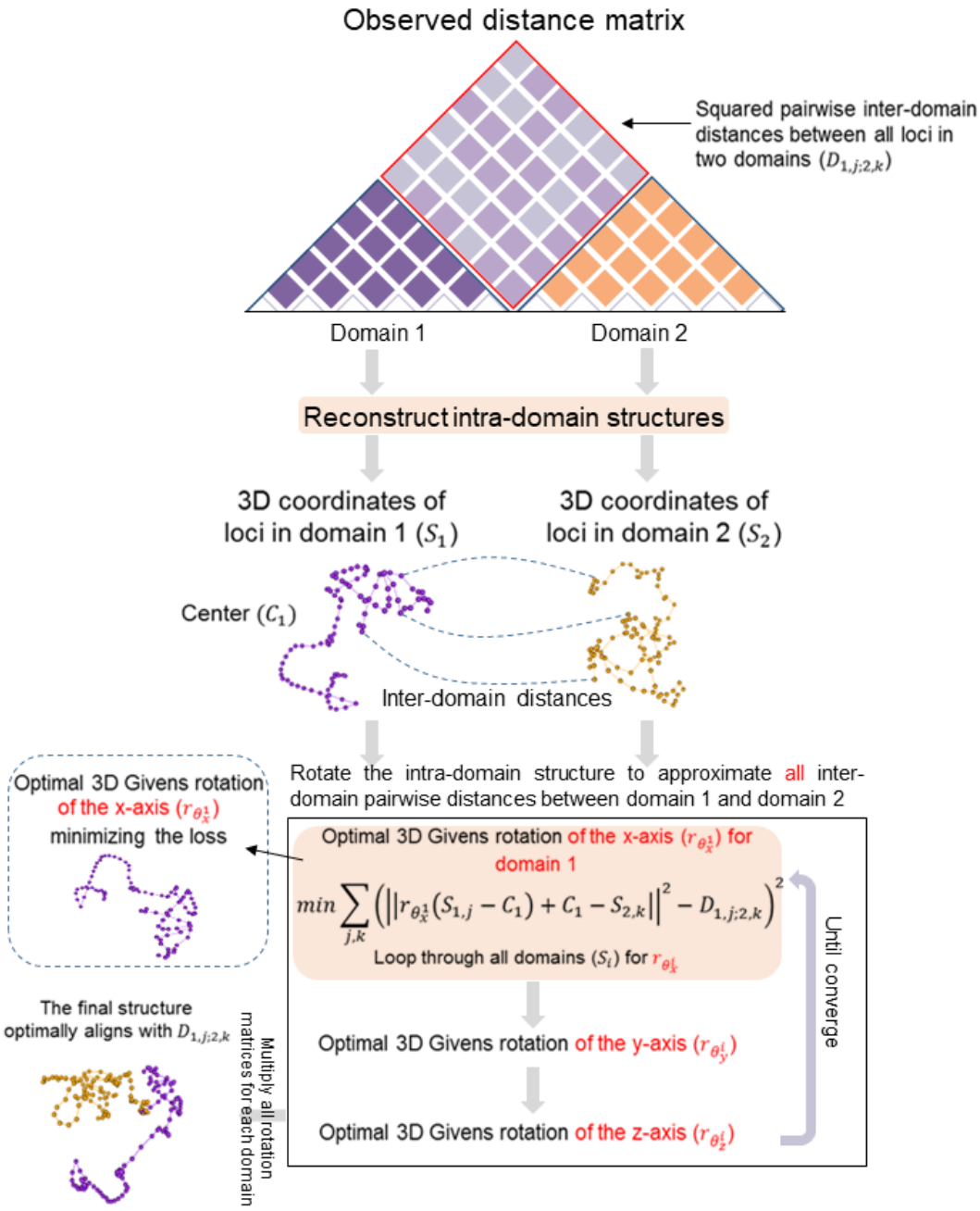



**Figure A.3 Overview of the assembly algorithm of FLAMINGO.**

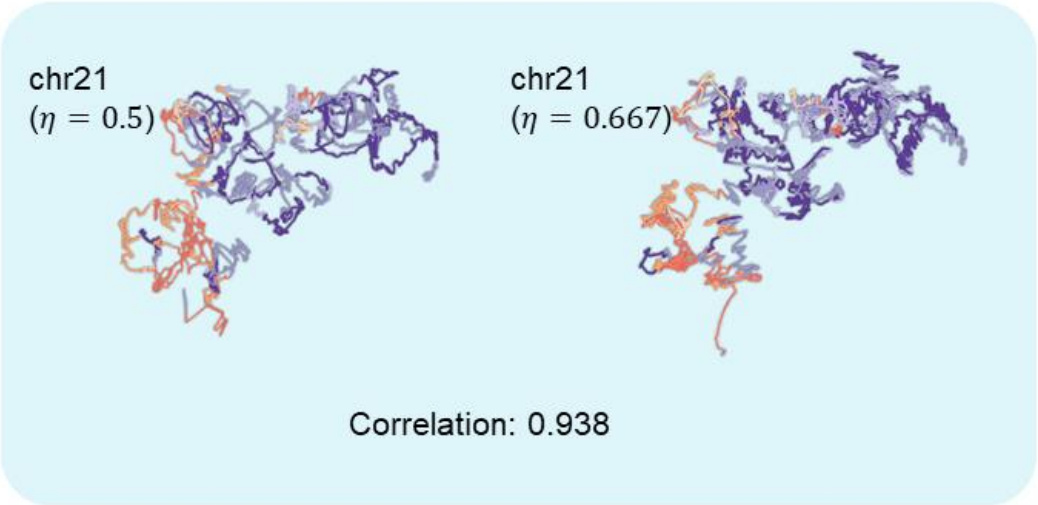



**Figure A.4 High similarity of predicted structures using different conversion factors.**

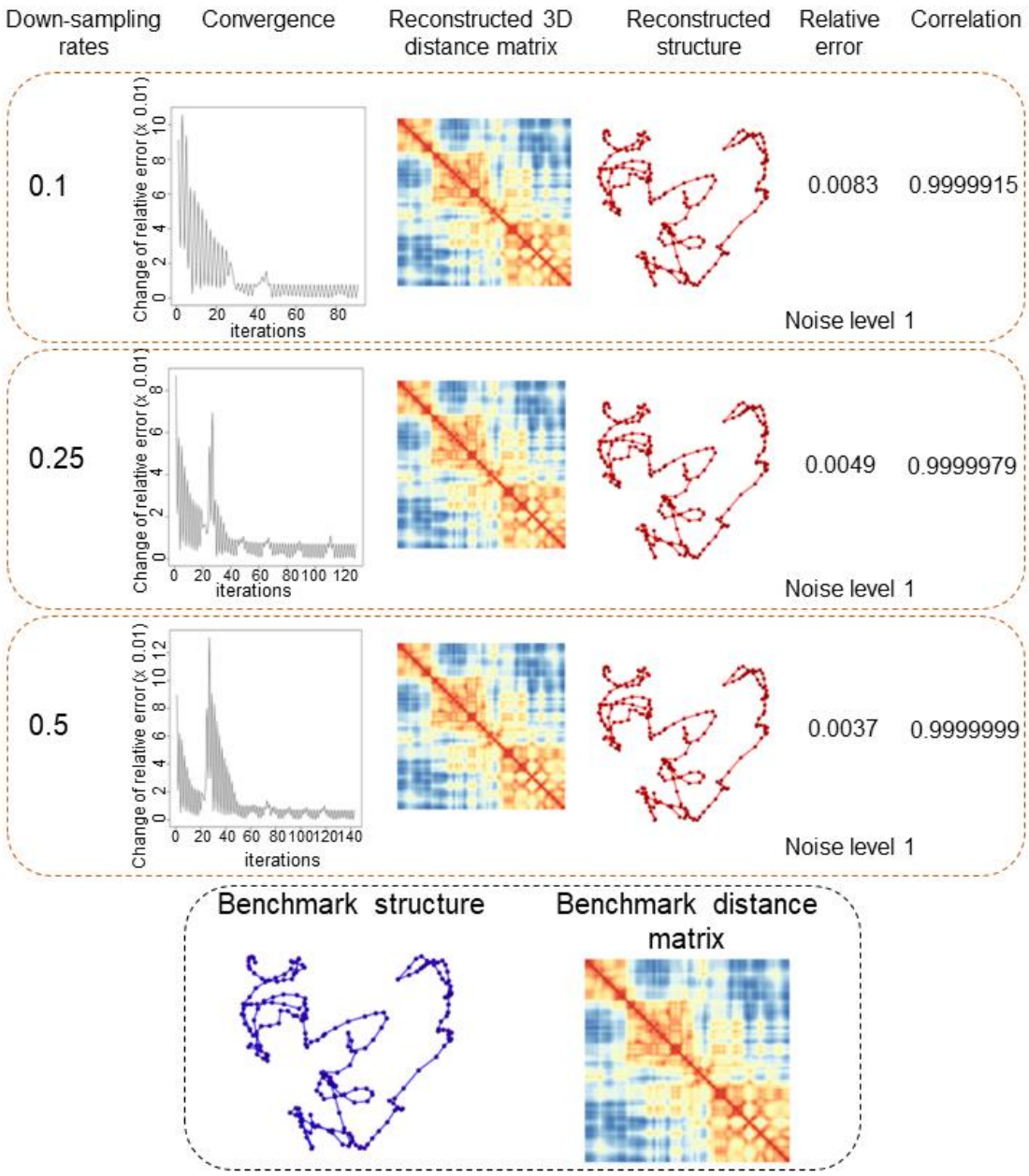



**Figure A.5 Convergence and model performance under different down-sampling rates based on simulated structures.**

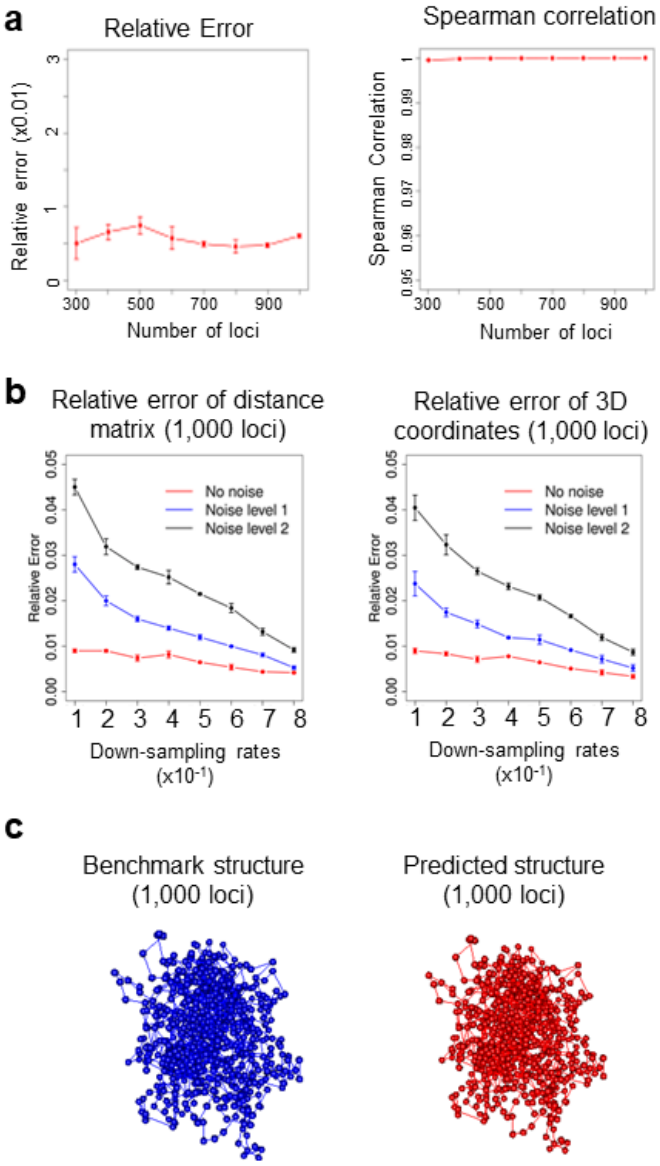



**Figure A.6 Model performance under different number of loci and down sampling rates based on simulated structures.**

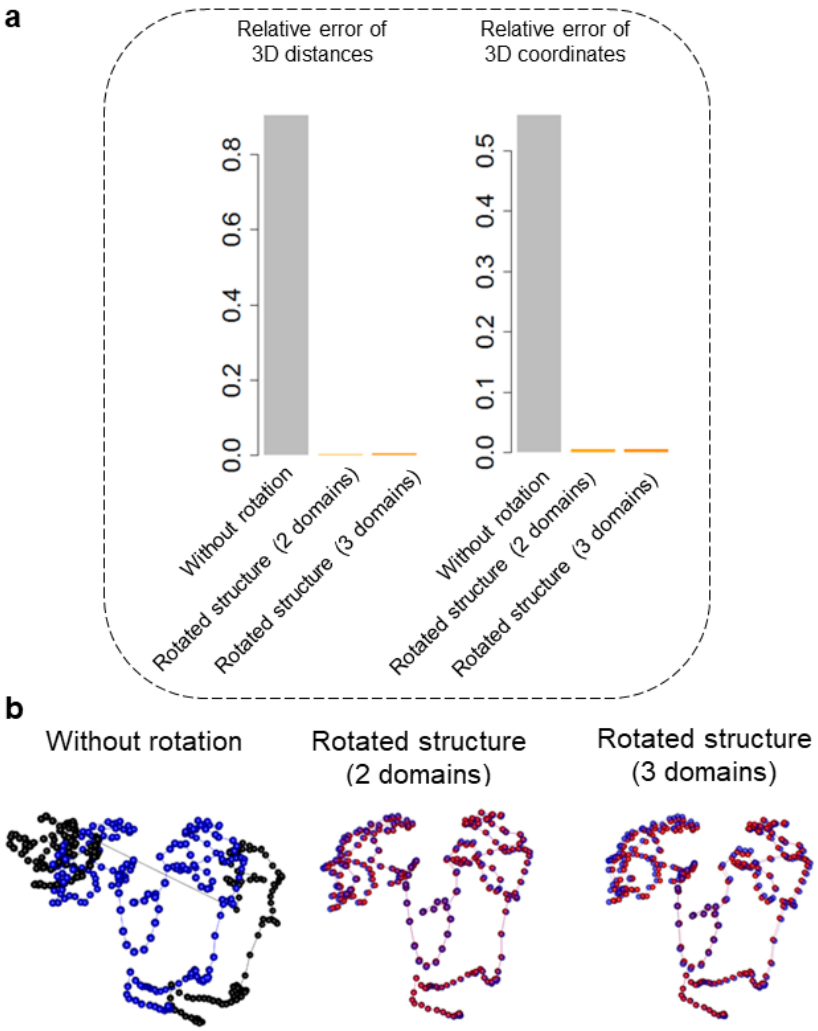



**Figure A.7 Validation of the assembly algorithm based on simulations.**

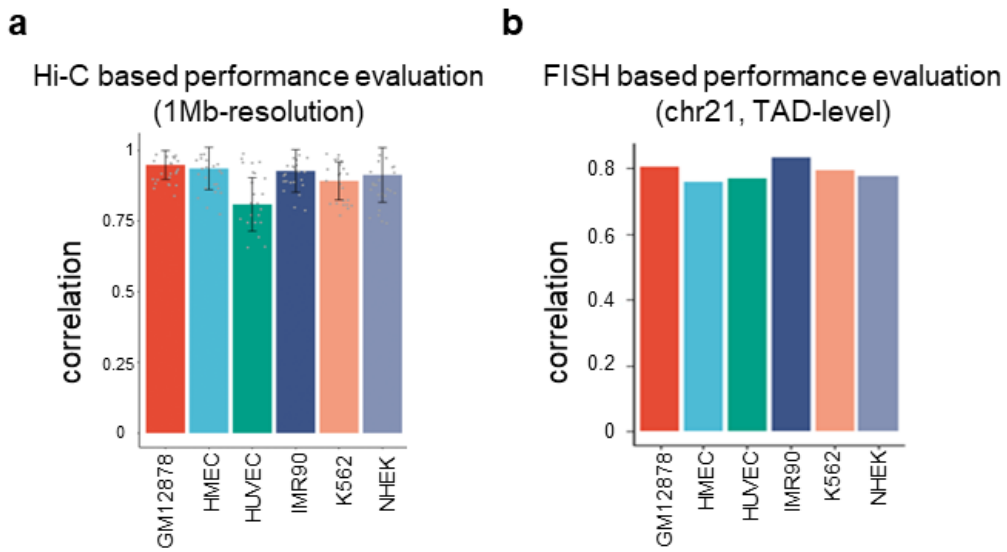



**Figure A.8 Performance validation using low-resolution Hi-C data and FISH data.**

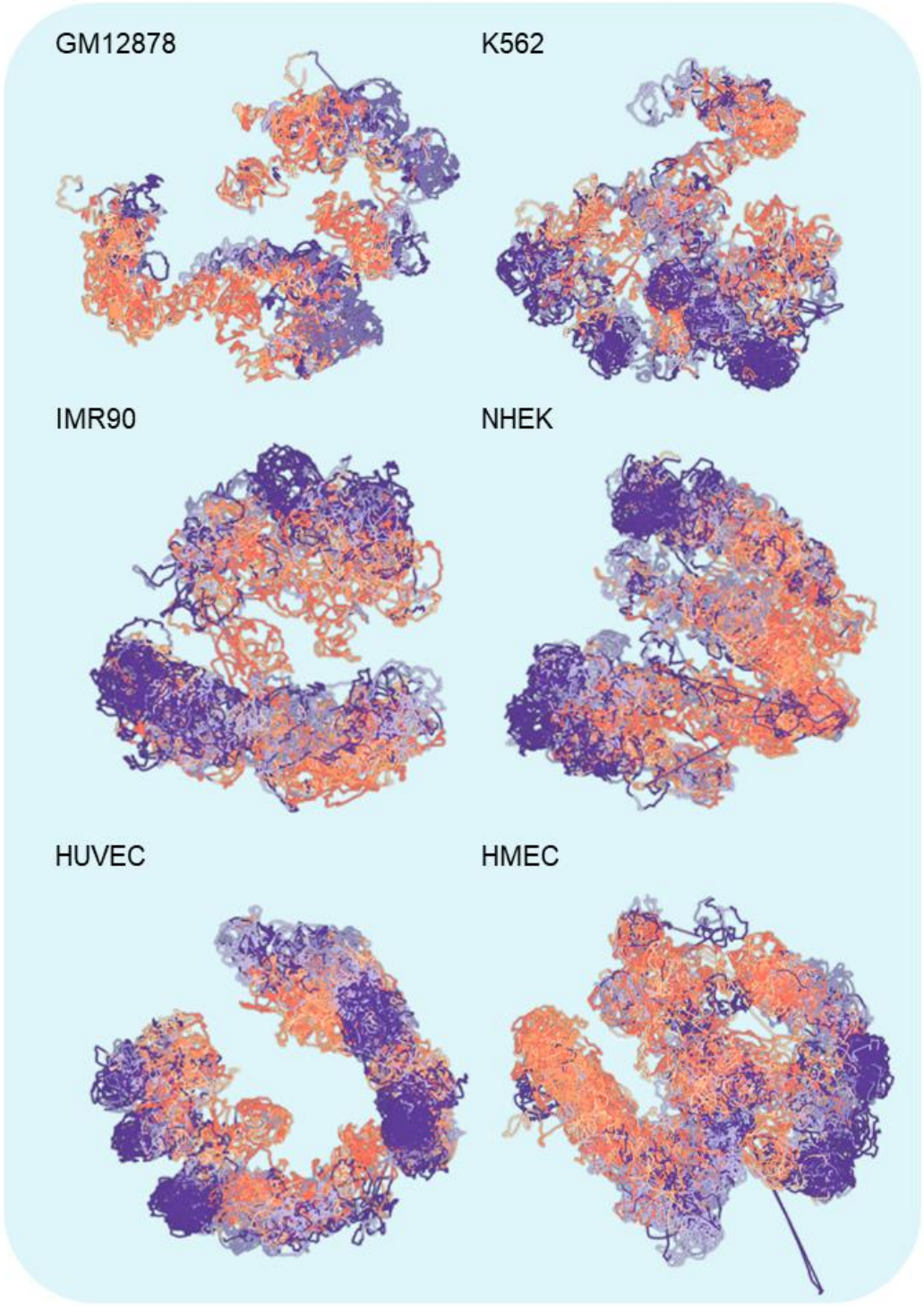



**Figure A.9 Predicted 3D structures of chr1 by FLAMINGO in six cell-types at 5-kb resolution.**

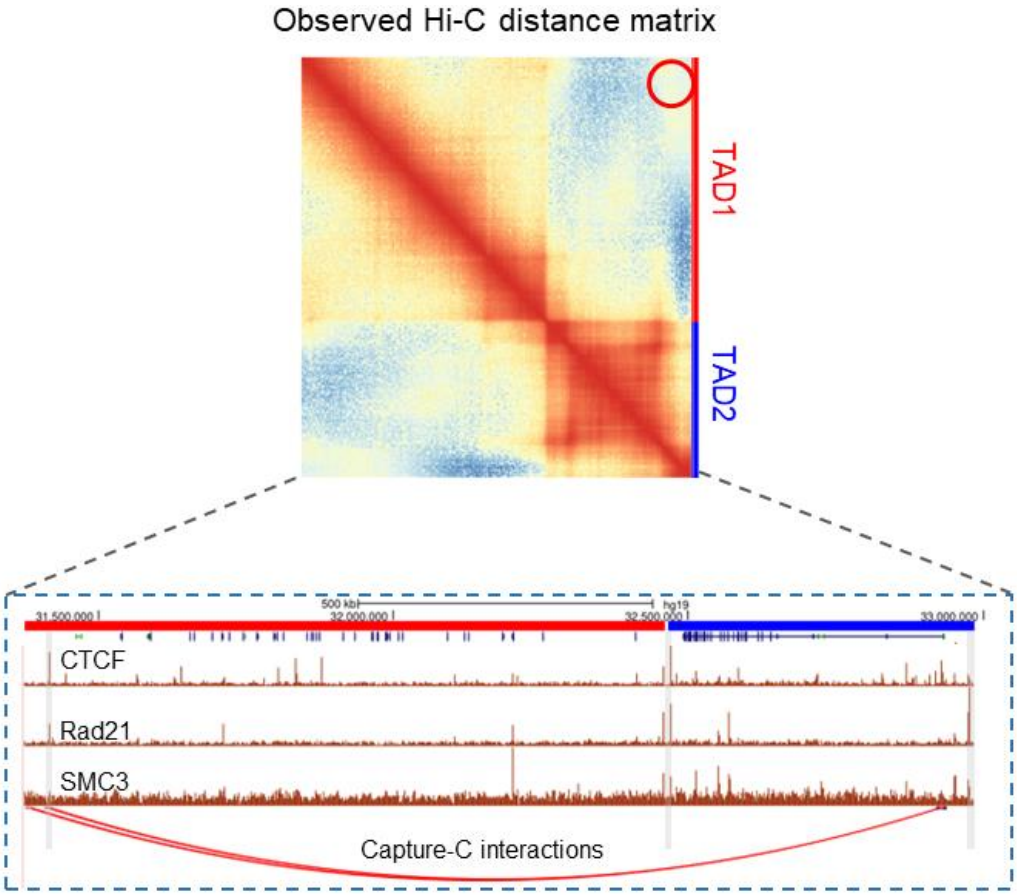



**Figure A.10 The observed long-range chromatin interactions are supported by TF ChIP-seq and Capture-C interactions.**

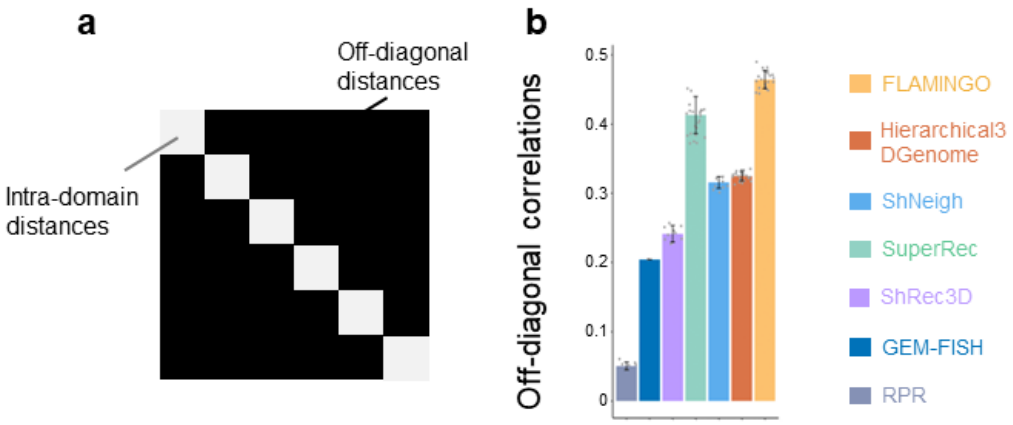



**Figure A.11 Performance comparison in GM12878 based on off-diagonal distances.**

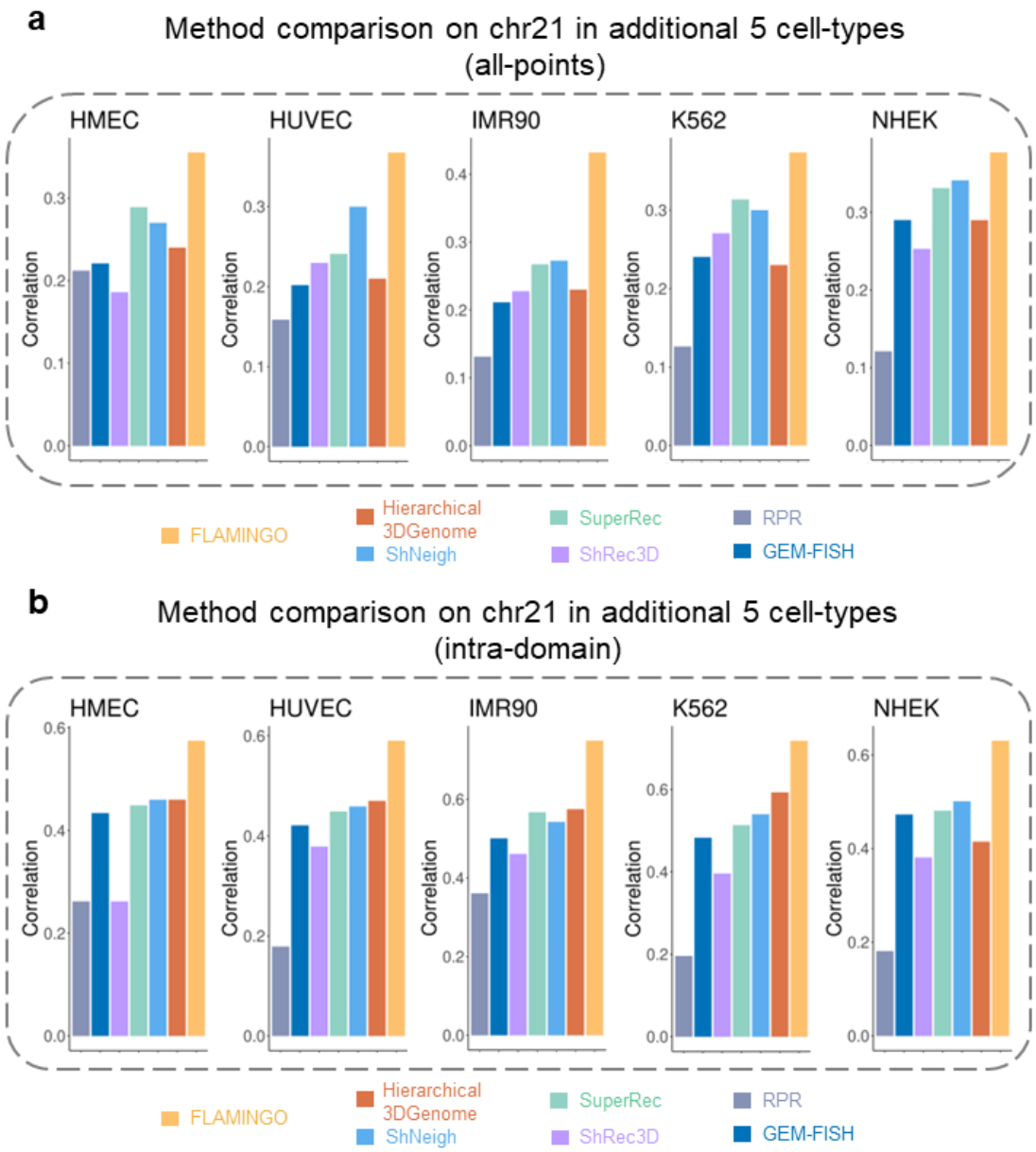



**Figure A.12 Performance comparison in the additional five cell-types.**

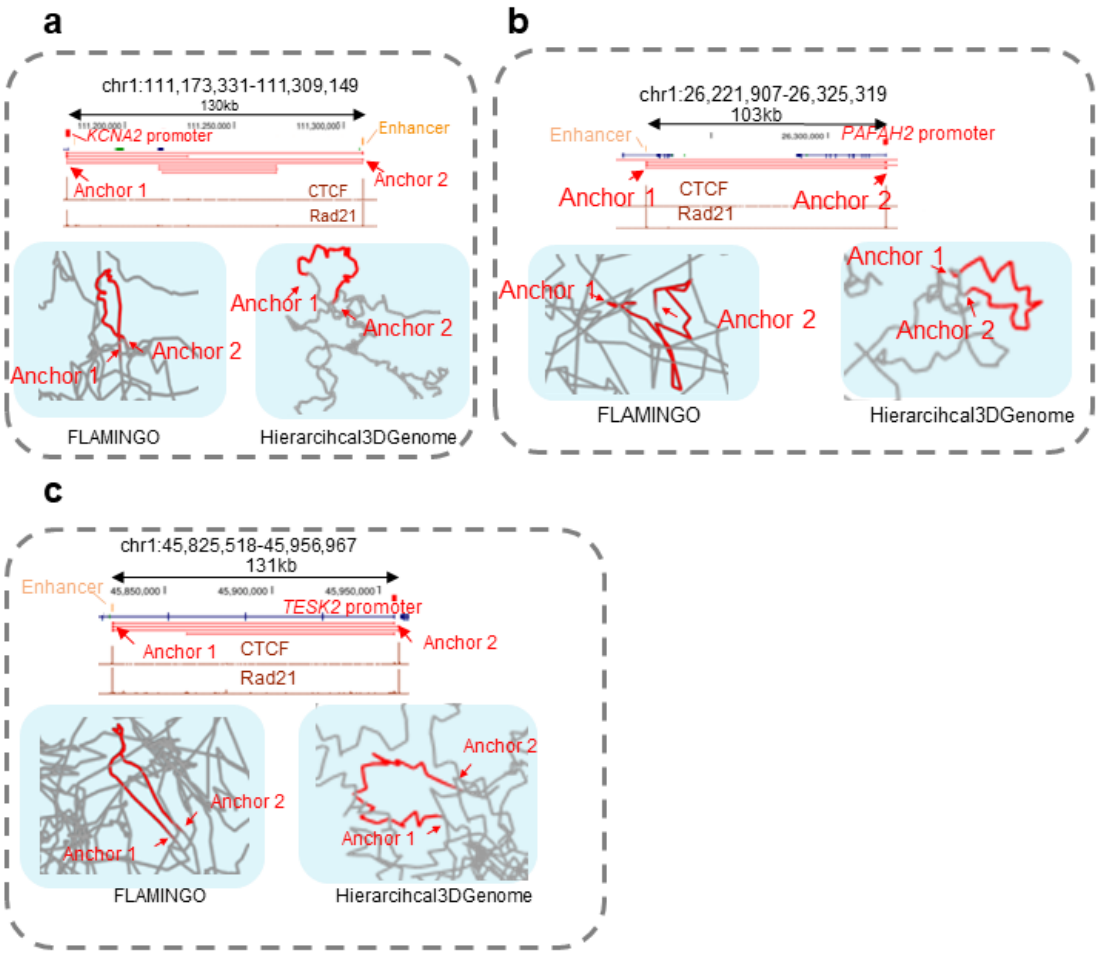



**Figure A.13 Example of 3D chromatin loops reconstructed by FLAMINGO.**

**a** The runtime and memory usage of FLAMINGO across all 23 chromosomes

| Chromosome ID | Chromosome size (5kb fragments) | Run time (mins) | Memory (GB) | Fraction of missing data |
|---|---|---|---|---|
| 1 | 44,027 | 42 | 2.2 | 0.945 |
| 2 | 44,872 | 41.2 | 2.3 | 0.937 |
| 3 | 38,583 | 38.6 | 1.8 | 0.925 |
| 4 | 35,105 | 39.6 | 1.7 | 0.934 |
| 5 | 32,167 | 37.5 | 1.6 | 0.925 |
| 6 | 31,977 | 38.4 | 1.6 | 0.92 |
| 7 | 30,352 | 38.1 | 1.5 | 0.919 |
| 8 | 26,246 | 37.6 | 1.3 | 0.908 |
| 9 | 17,608 | 35.7 | 1.2 | 0.933 |
| 10 | 25,363 | 39.5 | 1.3 | 0.9 |
| 11 | 25,826 | 40 | 1.2 | 0.895 |
| 12 | 25,390 | 38 | 1.2 | 0.904 |
| 13 | 15,304 | 33.9 | 1 | 0.917 |
| 14 | 13,845 | 35.2 | 1 | 0.907 |
| 15 | 12,284 | 33.4 | 1 | 0.911 |
| 16 | 13,126 | 31.6 | 1 | 0.889 |
| 17 | 14,170 | 31.6 | 1 | 0.866 |
| 18 | 14,427 | 30.1 | 1 | 0.852 |
| 19 | 10,791 | 31.6 | 1 | 0.815 |
| 20 | 11,585 | 28.2 | 1 | 0.815 |
| 21 | 6,896 | 26.7 | 1 | 0.875 |
| 22 | 6,684 | 27.2 | 1 | 0.875 |
| X | 28,818 | 39.1 | 1.8 | 0.911 |

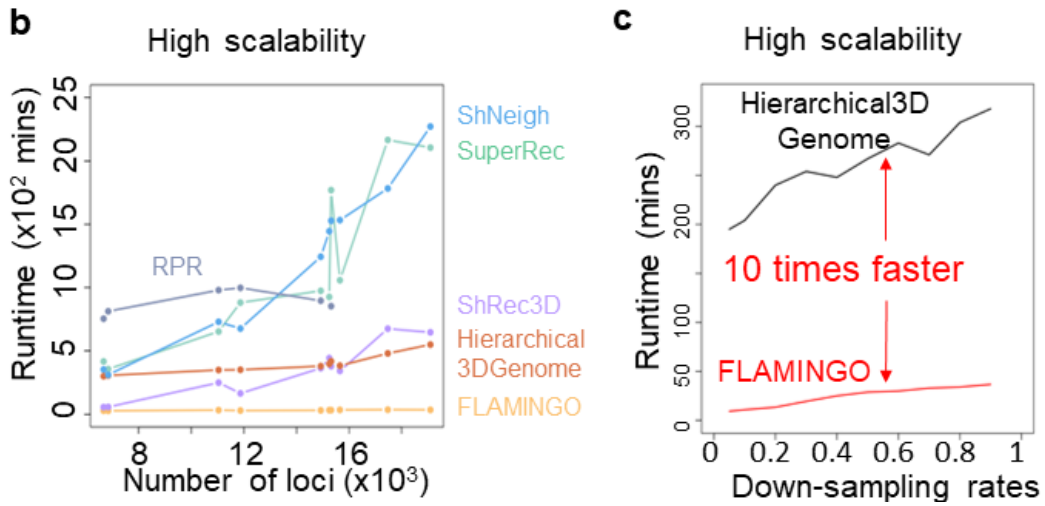



**Figure A.14 High scalability of FLAMINGO over existing algorithms.**

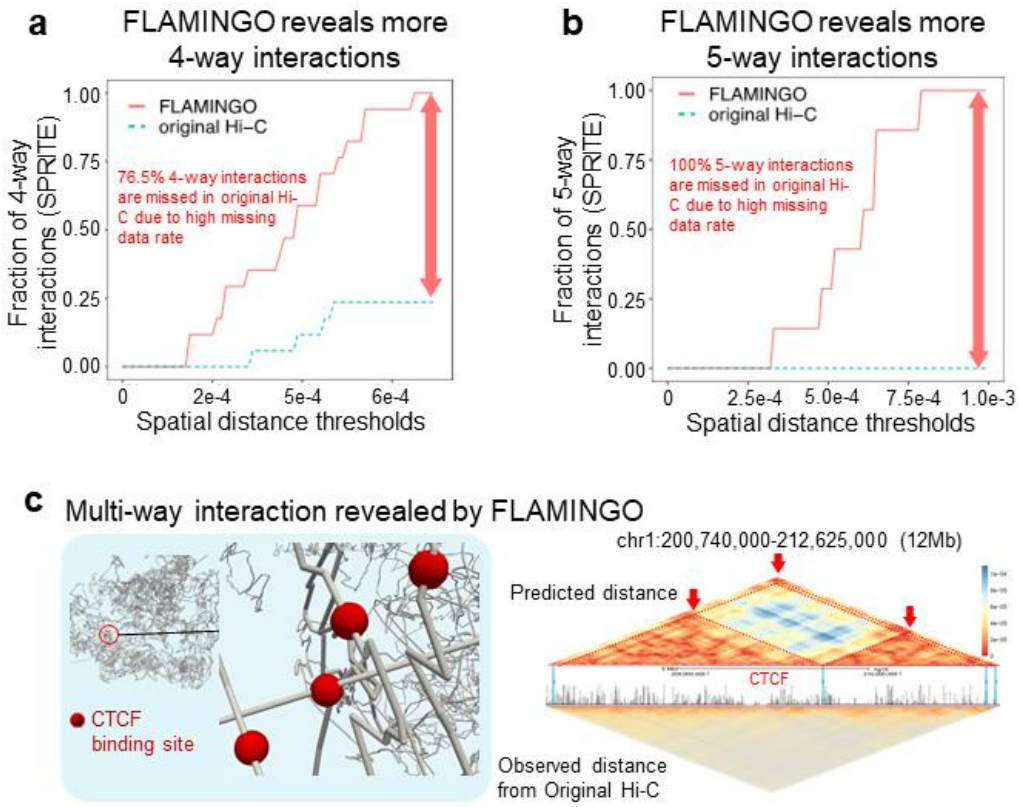



**Figure A.15 FLAMINGO leads to the discovery of multi-way chromatin interactions.**

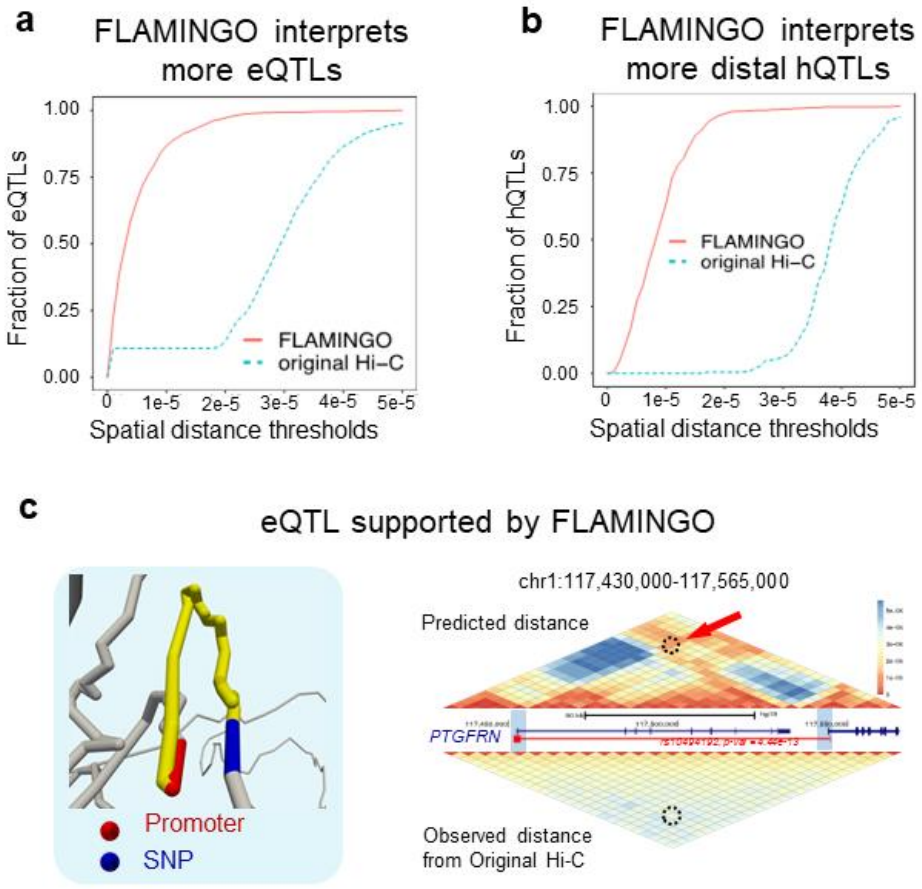



**Figure A.16 FLAMINGO provides structural basis of long-range QTLs.**

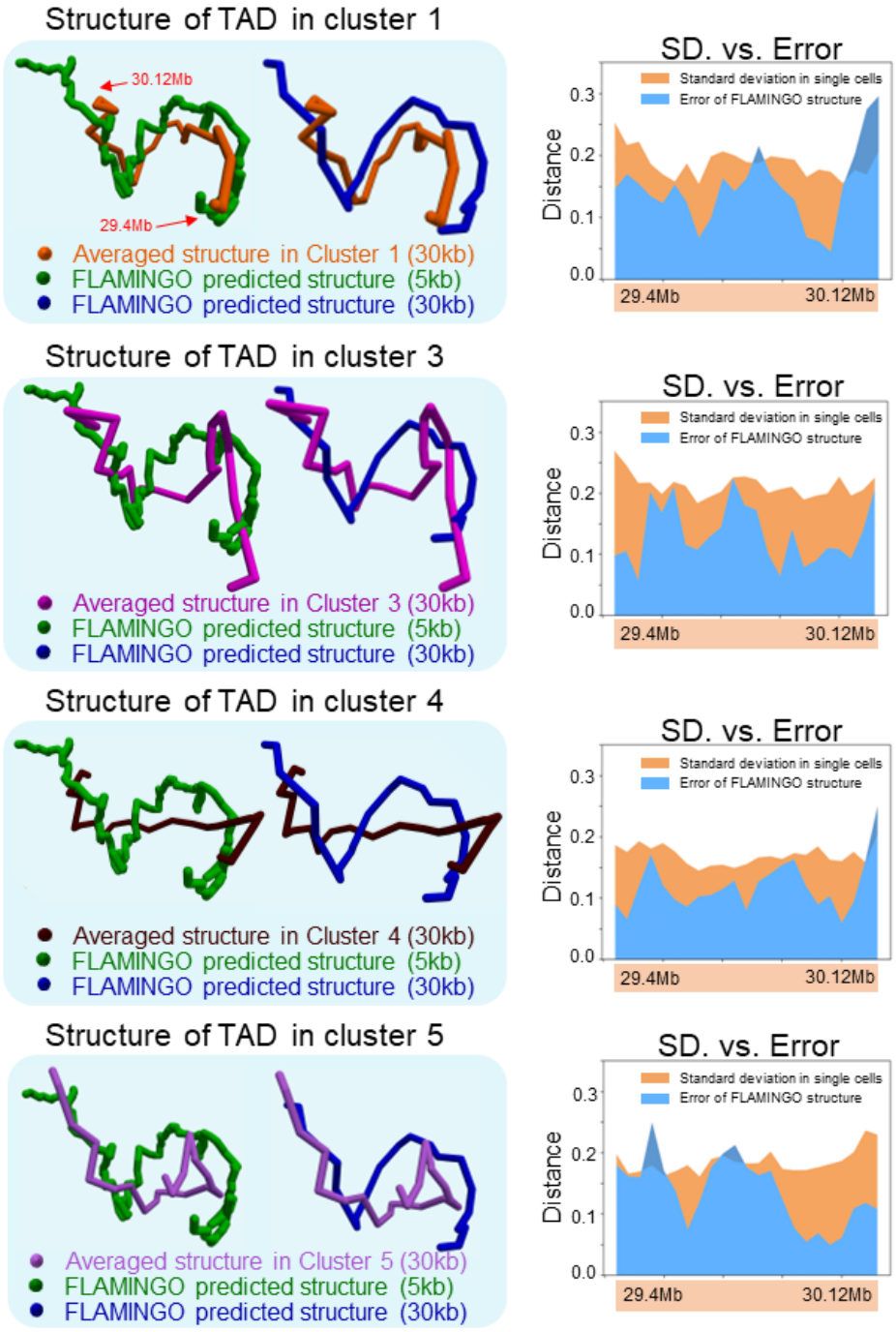



**Figure A.17 Comparison between the predicted structures with single-cell chromosome structures.**

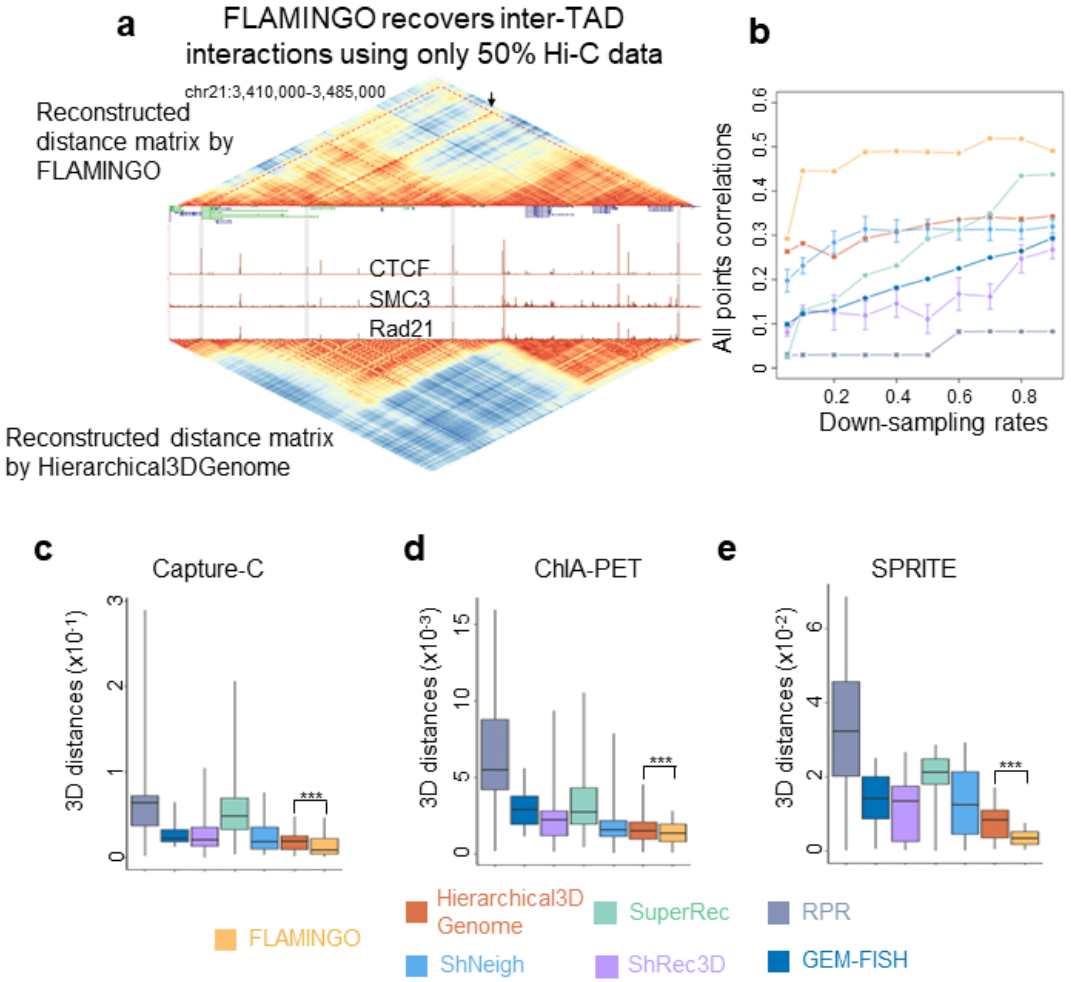



**Figure A.18 FLAMINGO robustly reconstructs the high-resolution 3D structures using a small fraction of observed Hi-C data.**

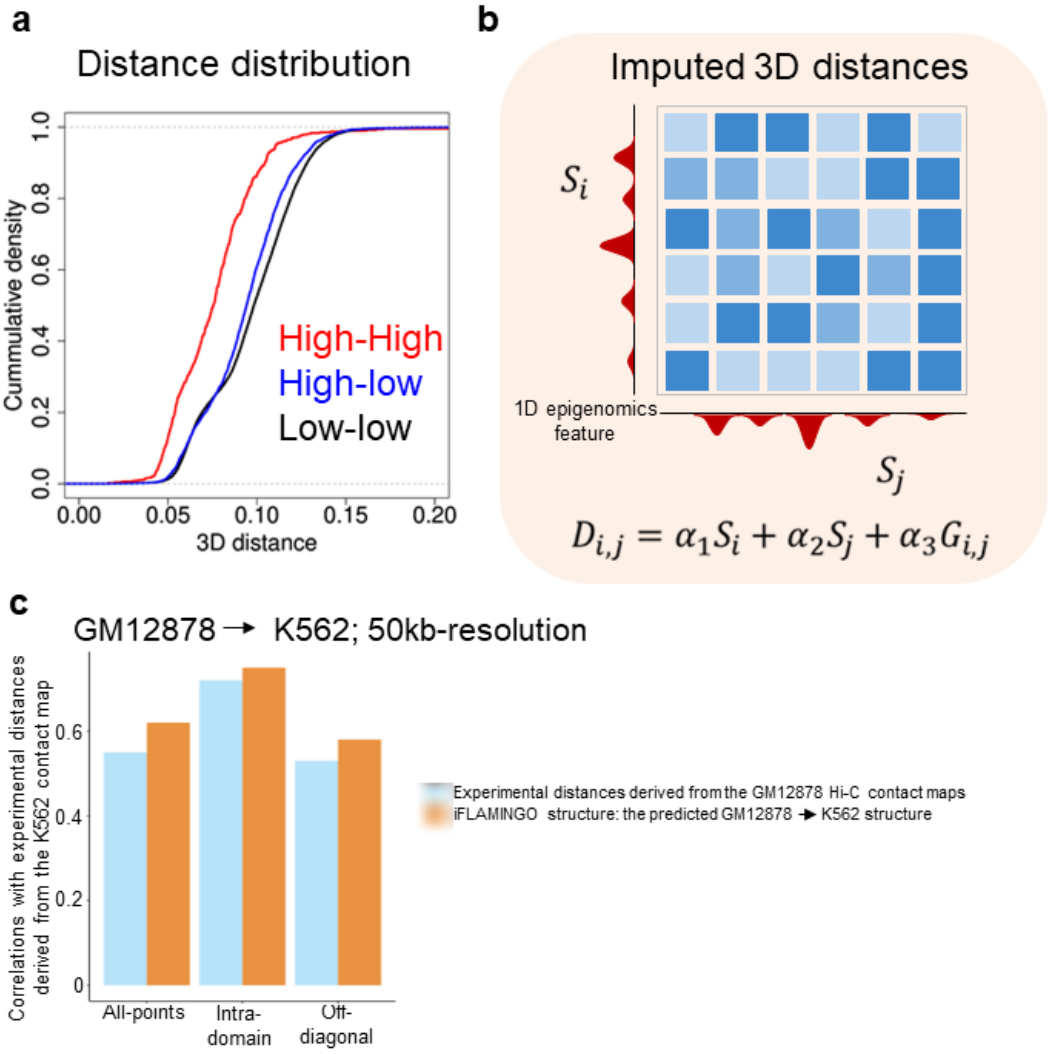



**Figure A.19 The imputation of 3D distances based on 1D epigenomics data in iFLAMINGO.**

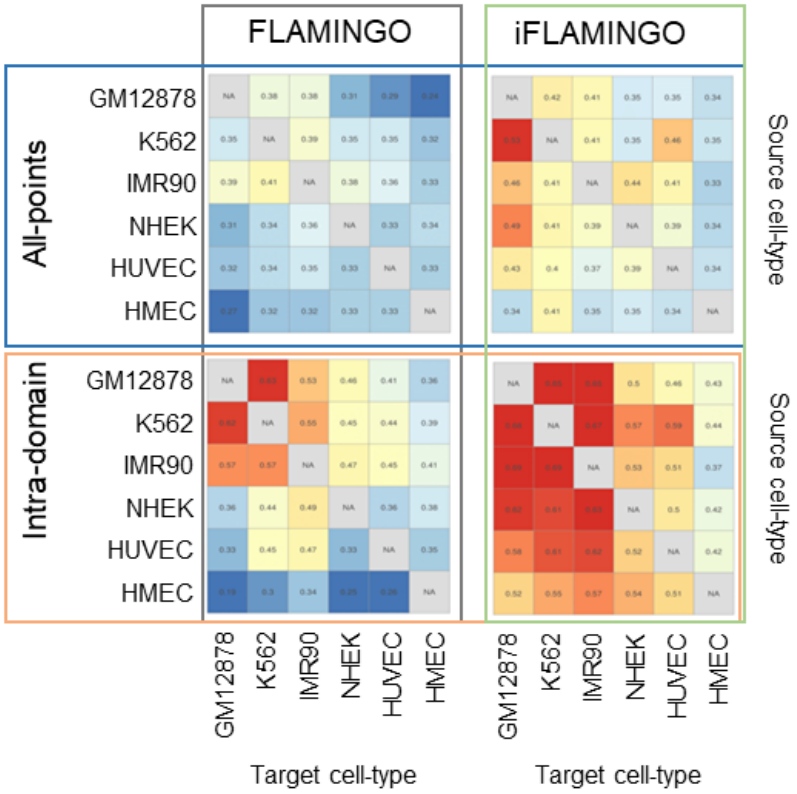



**Figure A.20 Performance of cross cell-type predictions using iFLAMINGO.**

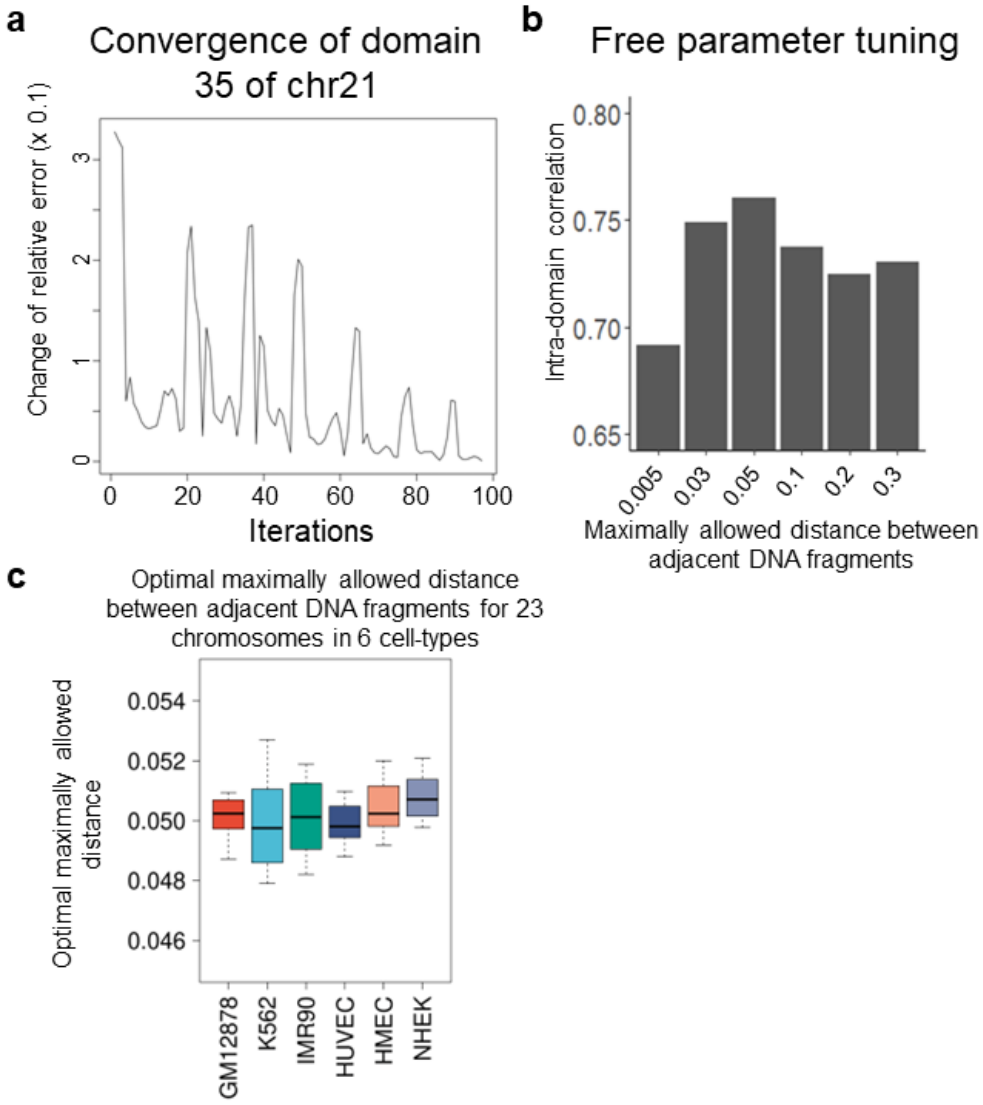



**Figure A.21 Convergence and parameter tuning of FLAMINGO.**

## APPENDIX B

## SUPPLEMENTARY FIGURES FOR CHAPTER 3

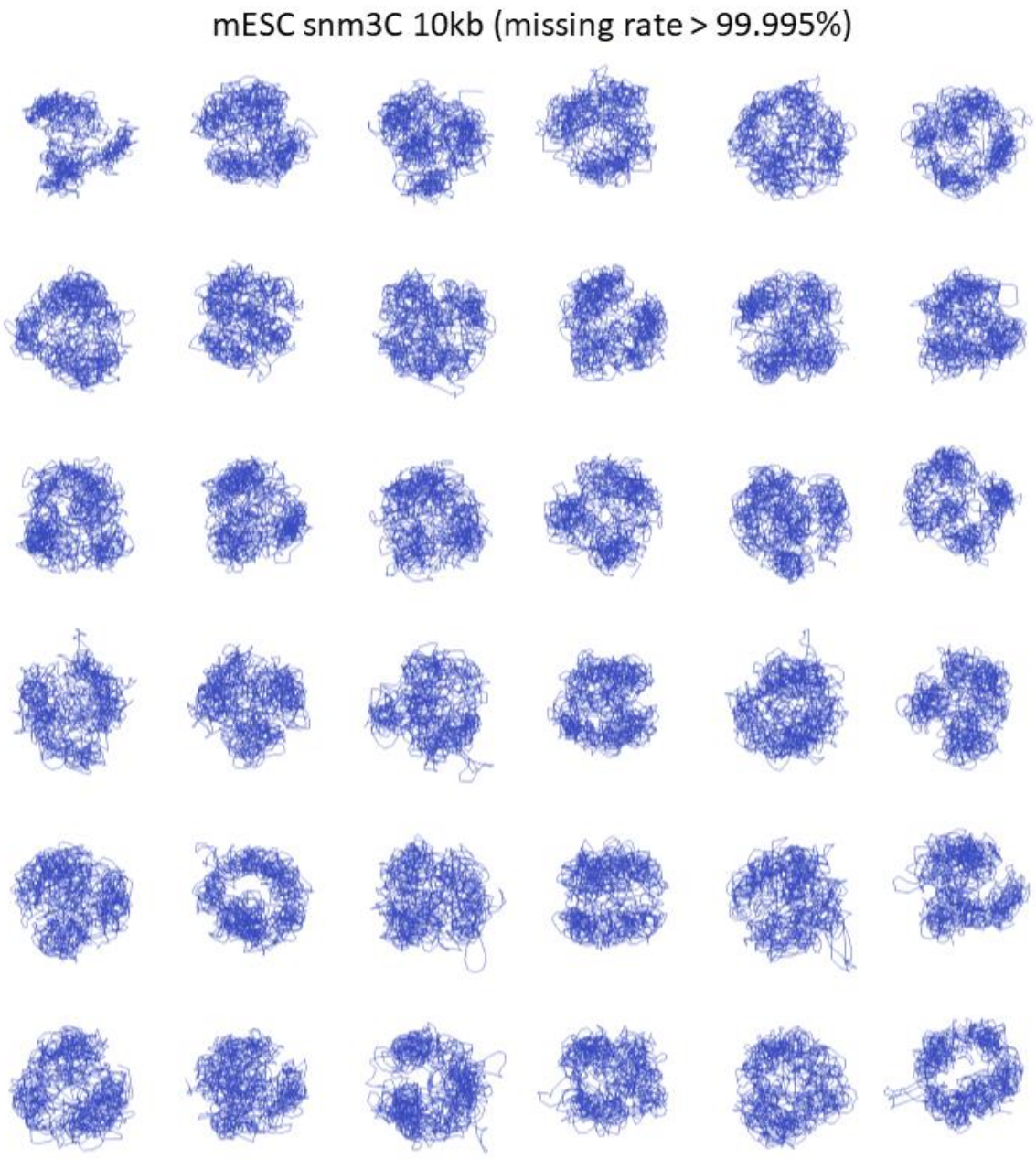



**Figure B.1 3D structures of chromosome 19 in 10kb-resolution for 351 mESC cells predicted by tFLAMINGO based on snm3C data.**

**Figure B.1 (cont'd)**

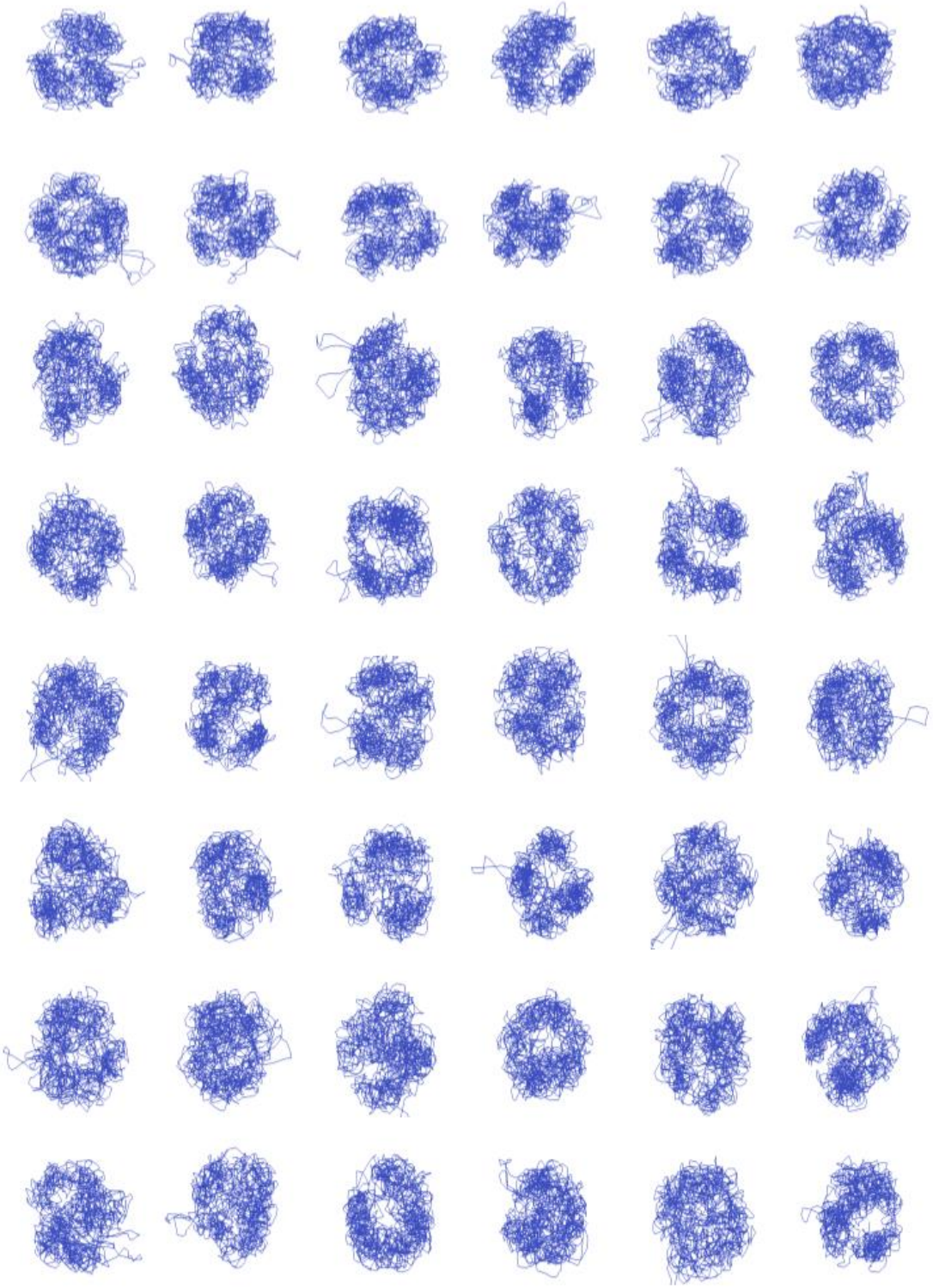


**Figure B.1 (cont'd)**

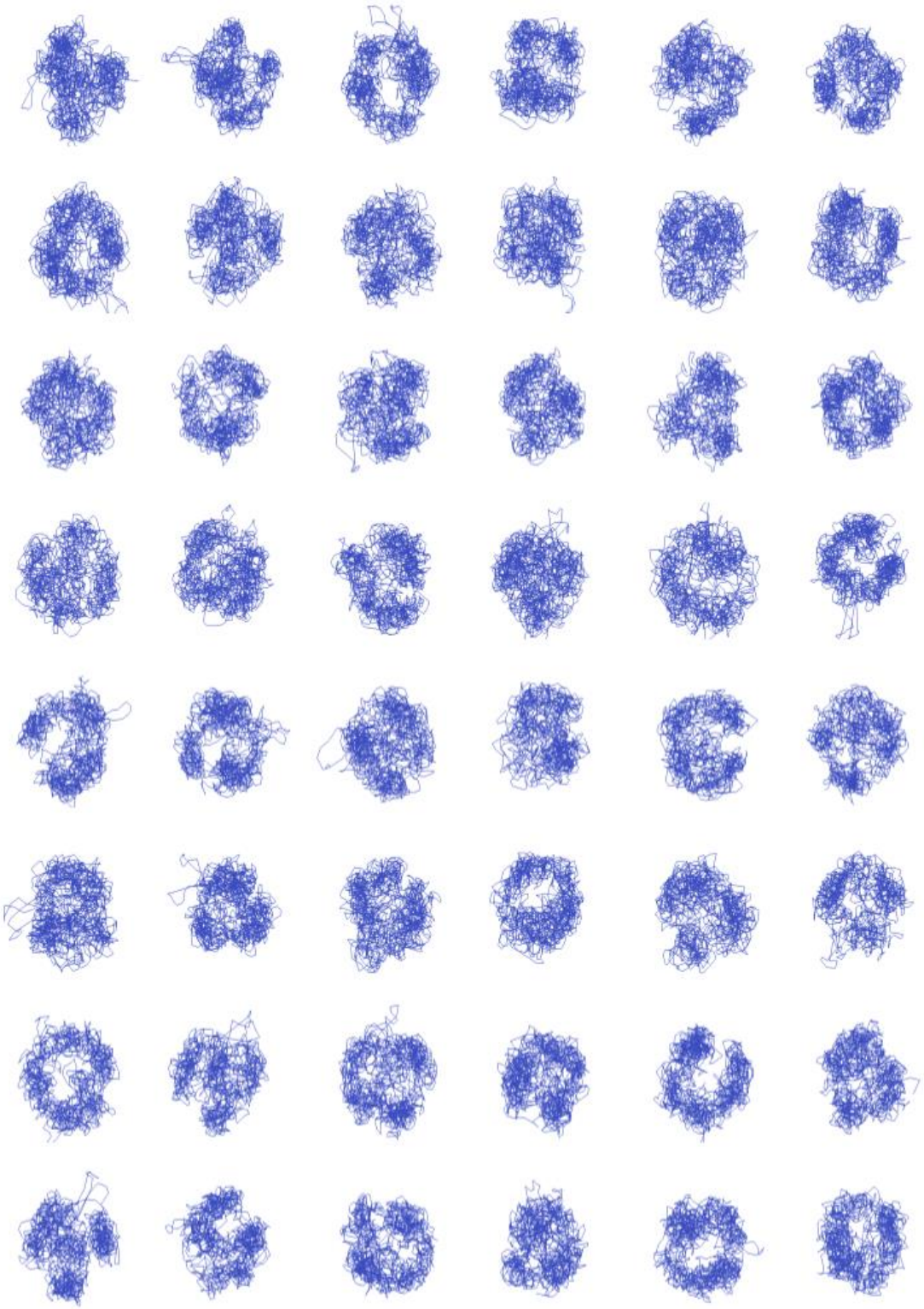


**Figure B.1 (cont'd)**

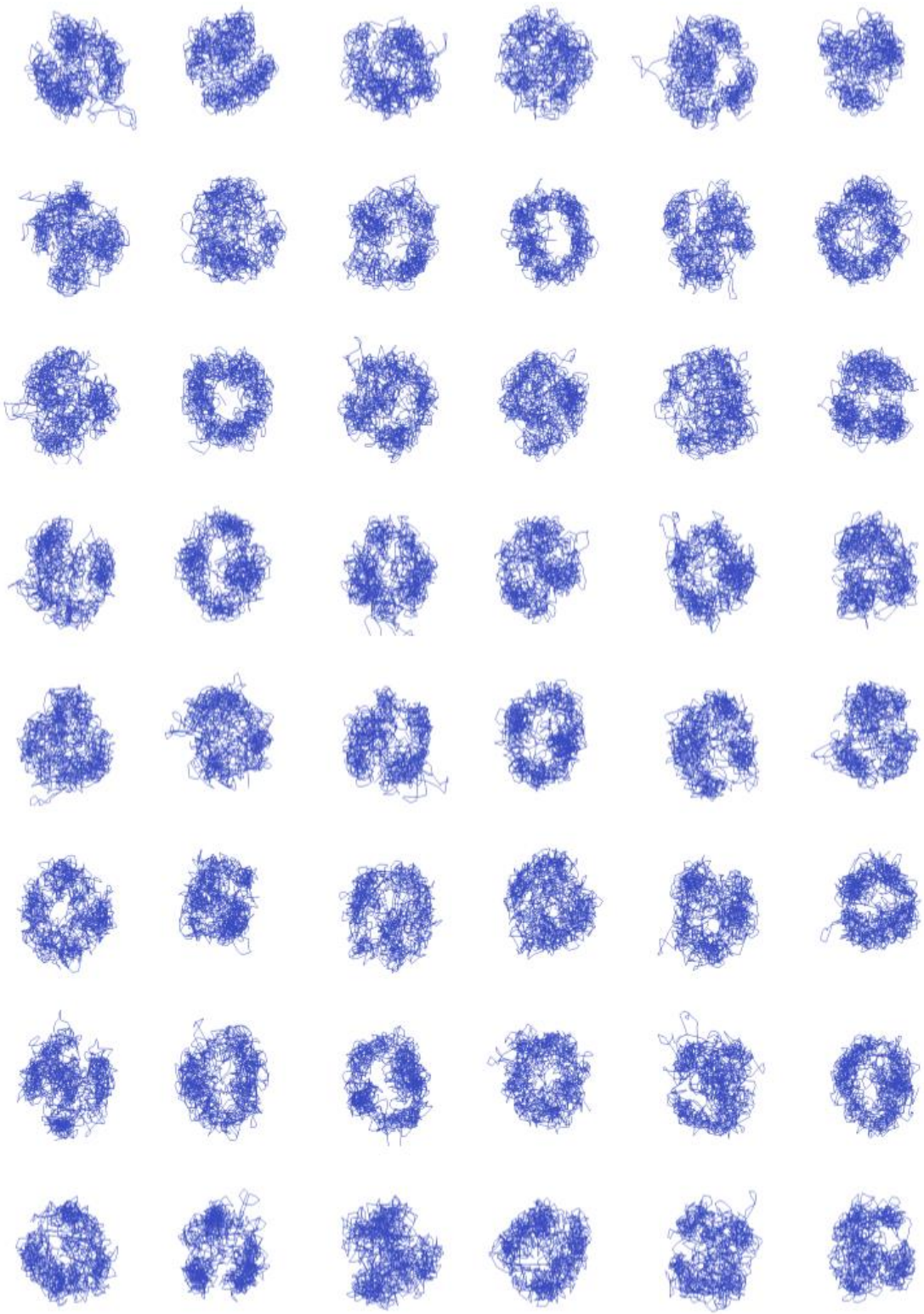


**Figure B.1 (cont'd)**

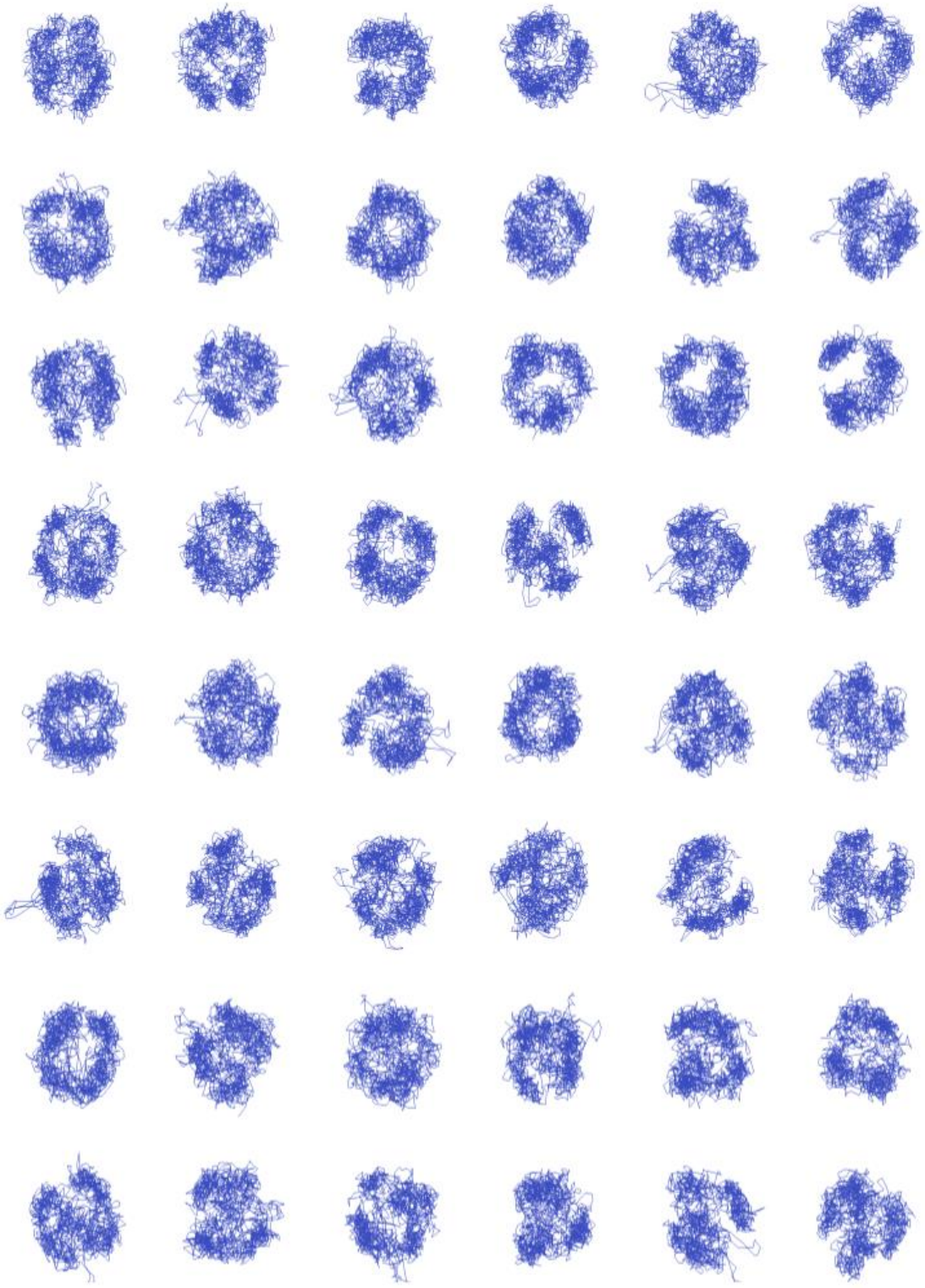


**Figure B.1 (cont'd)**

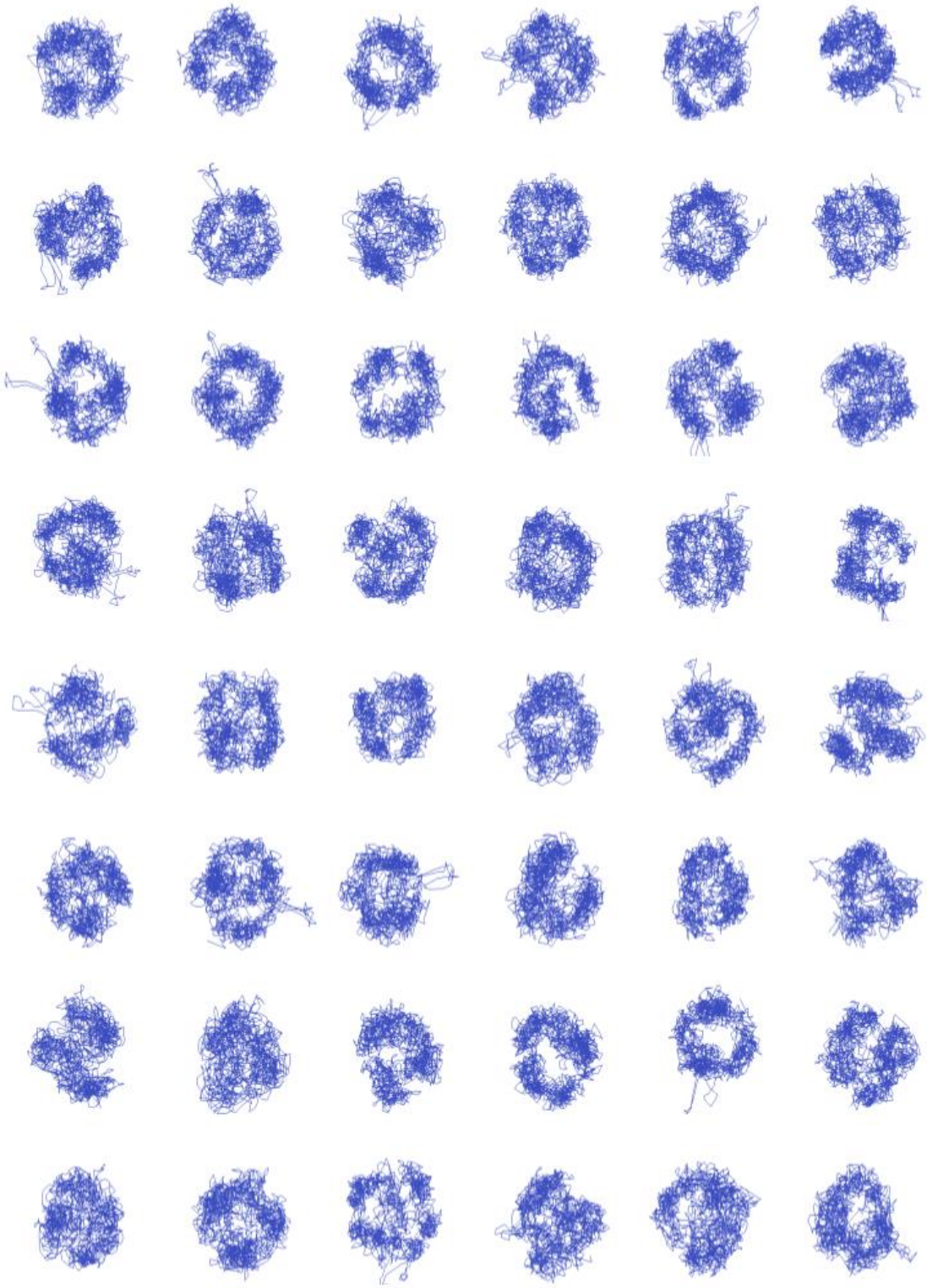


**Figure B.1 (cont'd)**

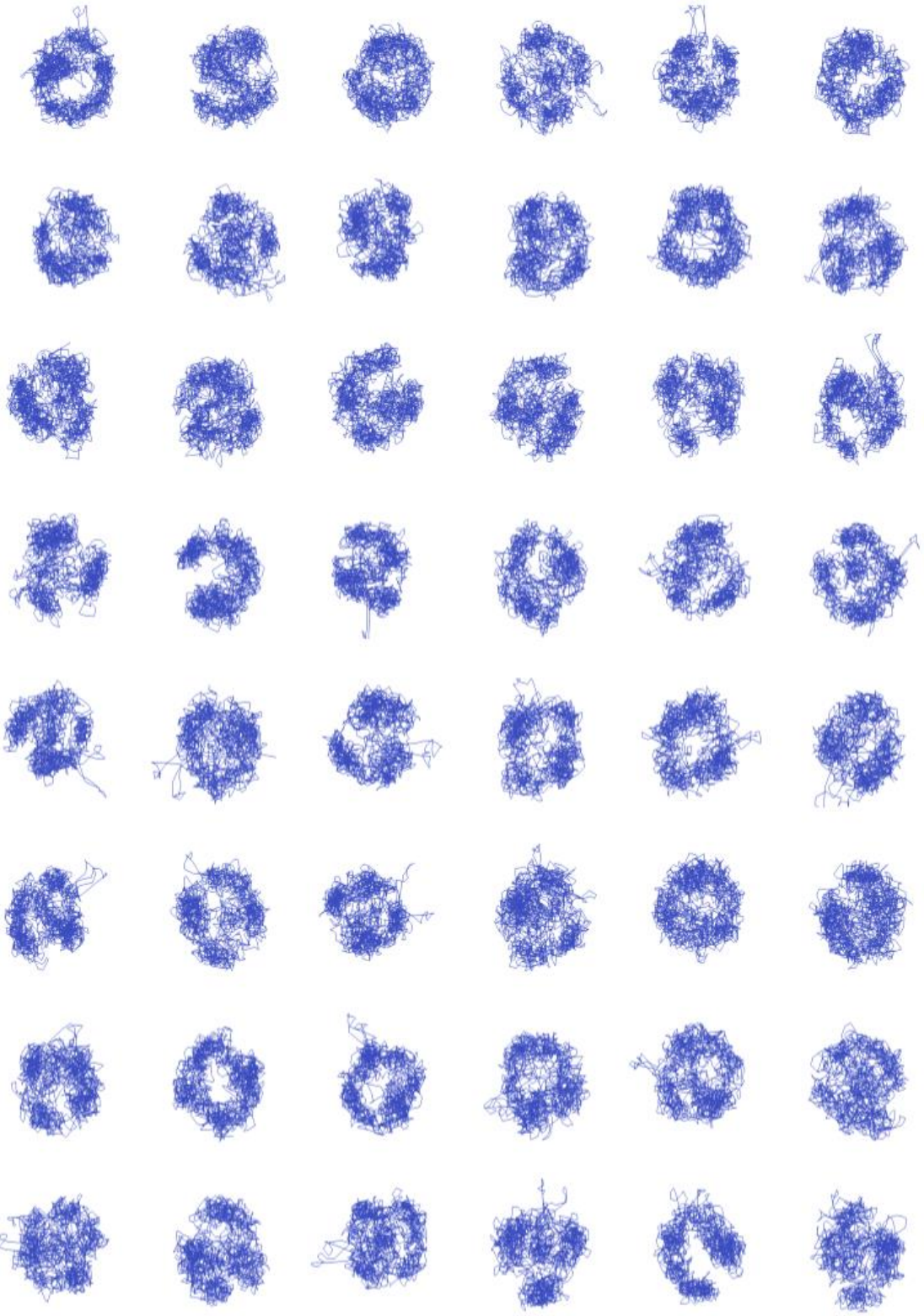
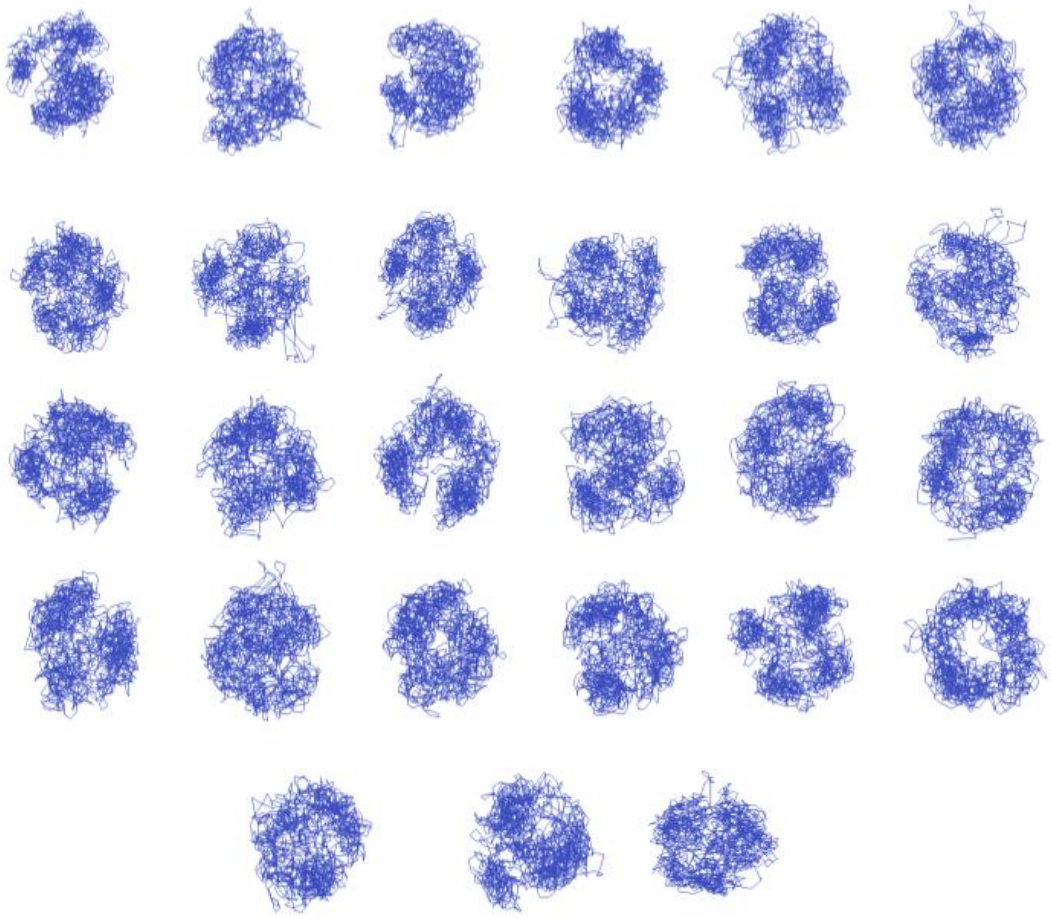


**Figure B.1 (cont'd)**



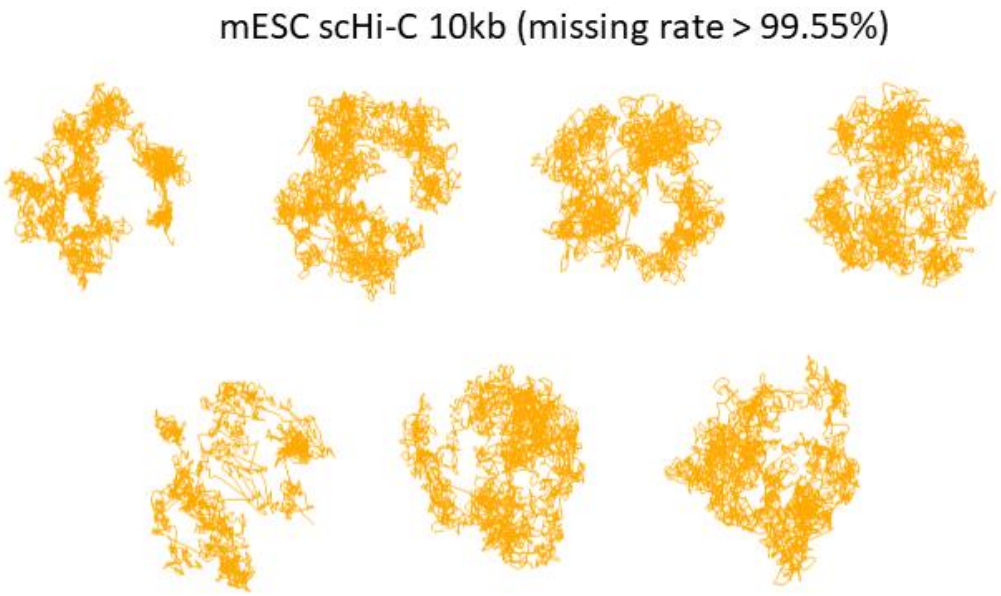



**Figure B.2 3D structures of chromosome 19 in 10kb-resolution for 7 mESC cells predicted by tFLAMINGO based on scHi-C data.**

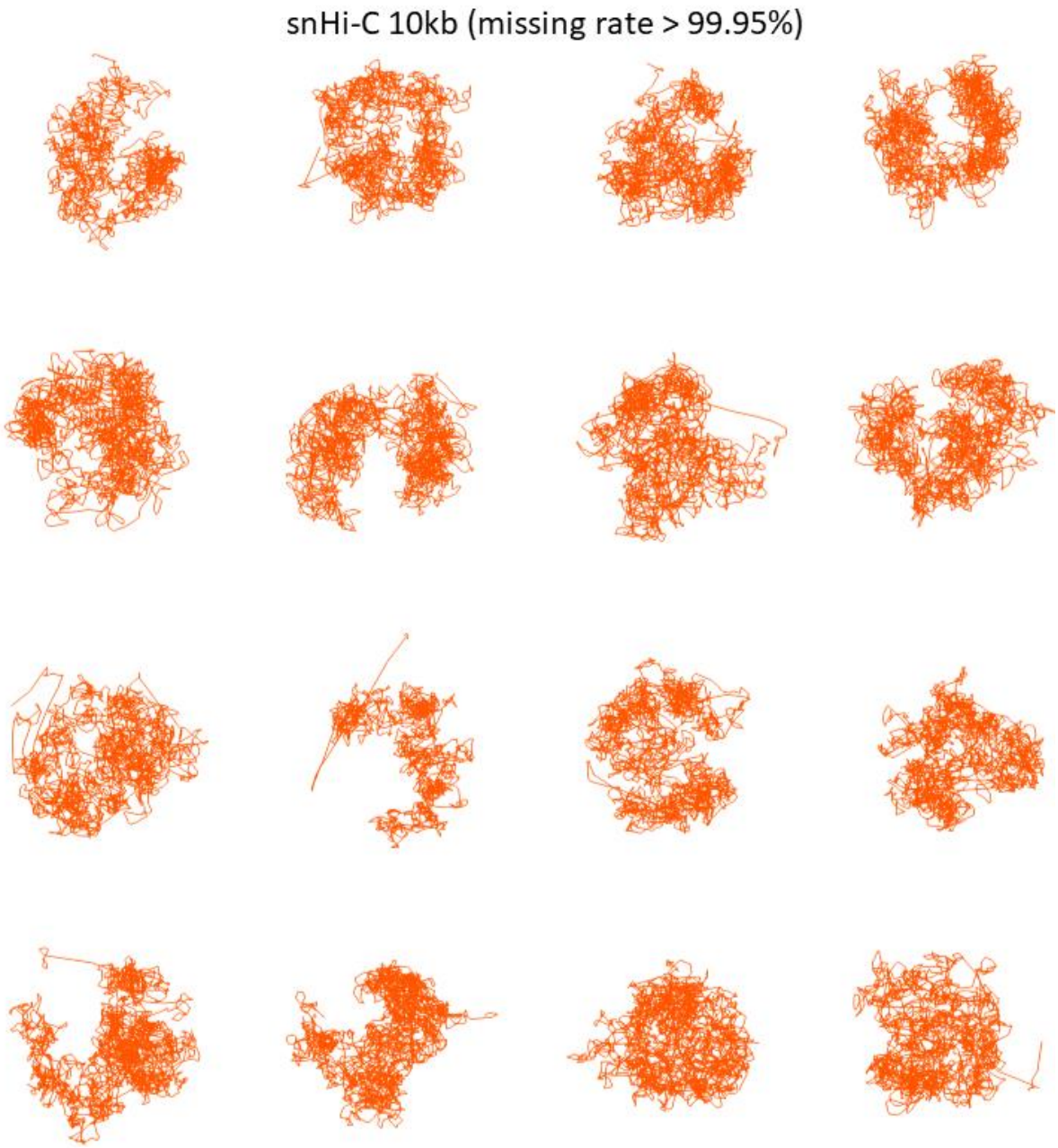



**Figure B.3 3D structures of chromosome 21 in 10kb-resolution for 16 K562 cells predicted by tFLAMINGO based on scHi-C data.**

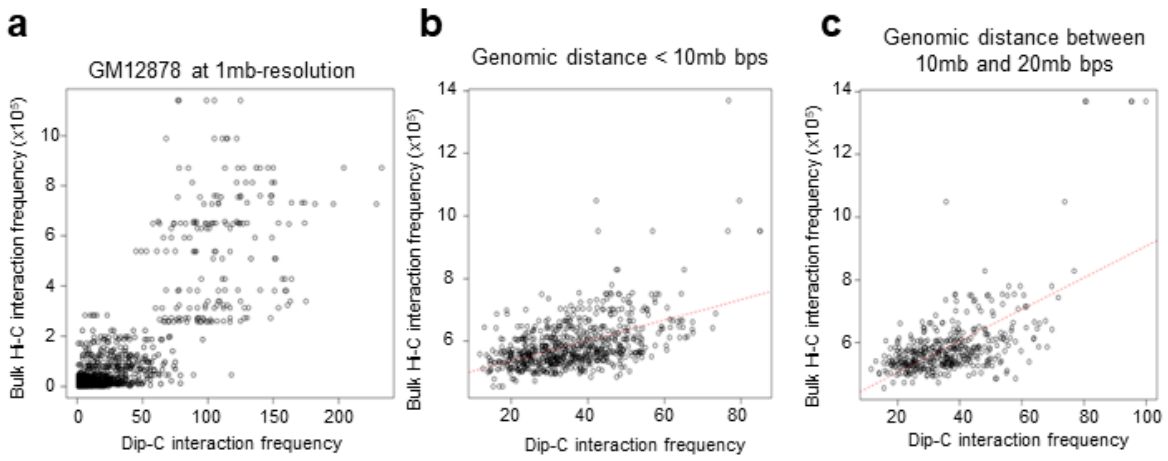



**Figure B.4 Differential linear relationships between single-cell 3C datasets and bulk Hi-C datasets.**

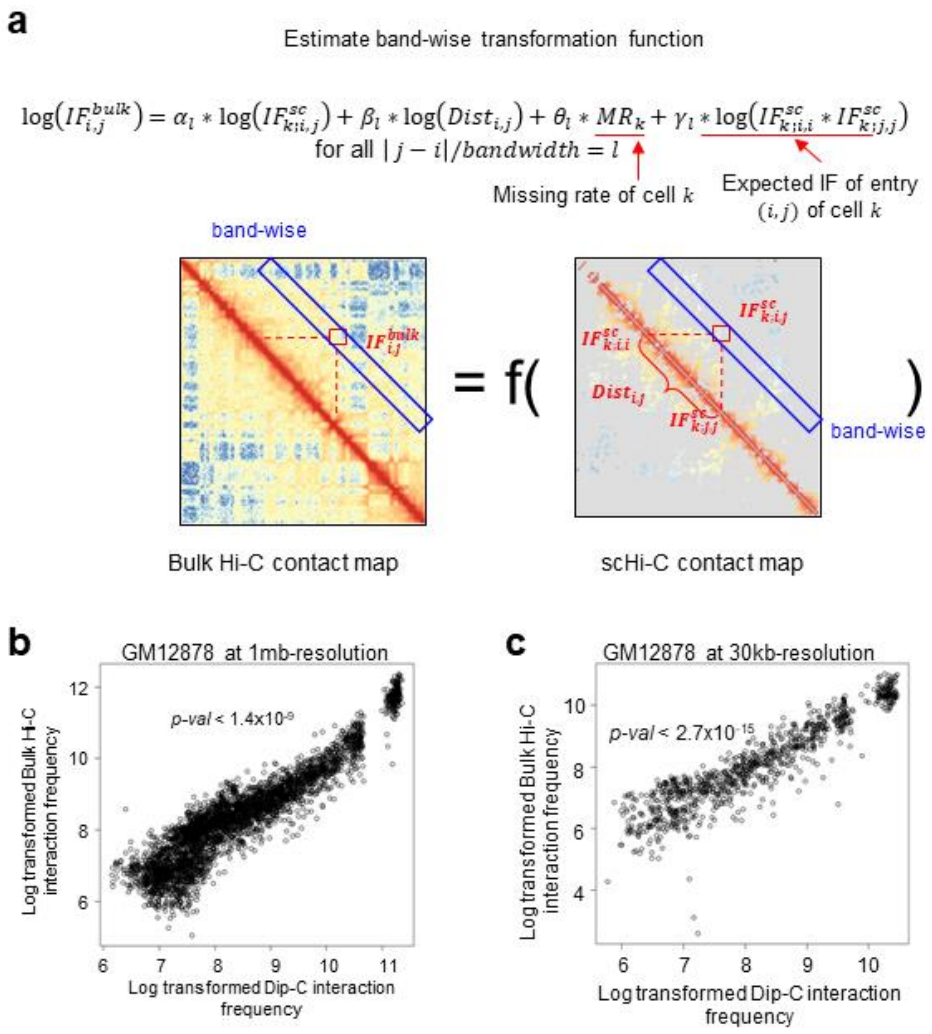



**Figure B.5 Schema of the band wise log-regression method to rescale the single-cell interaction frequencies.**

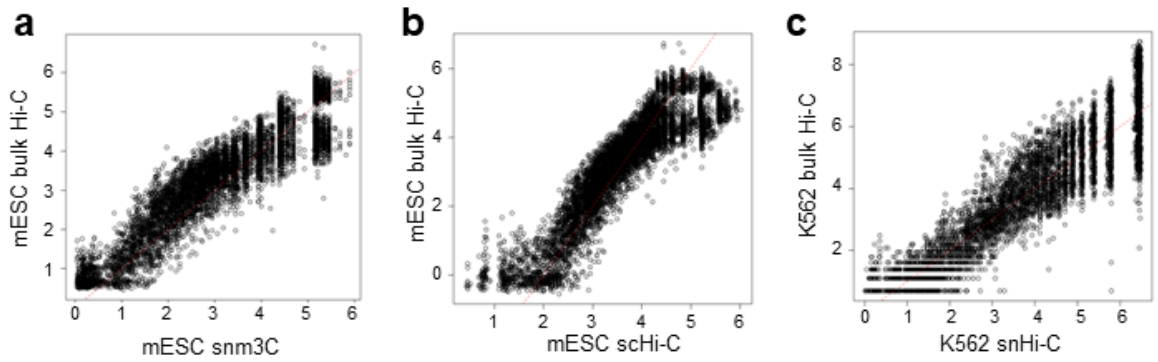



**Figure B.6 3D Validation of the transformed single-cell interaction frequencies based on three additional datasets.**

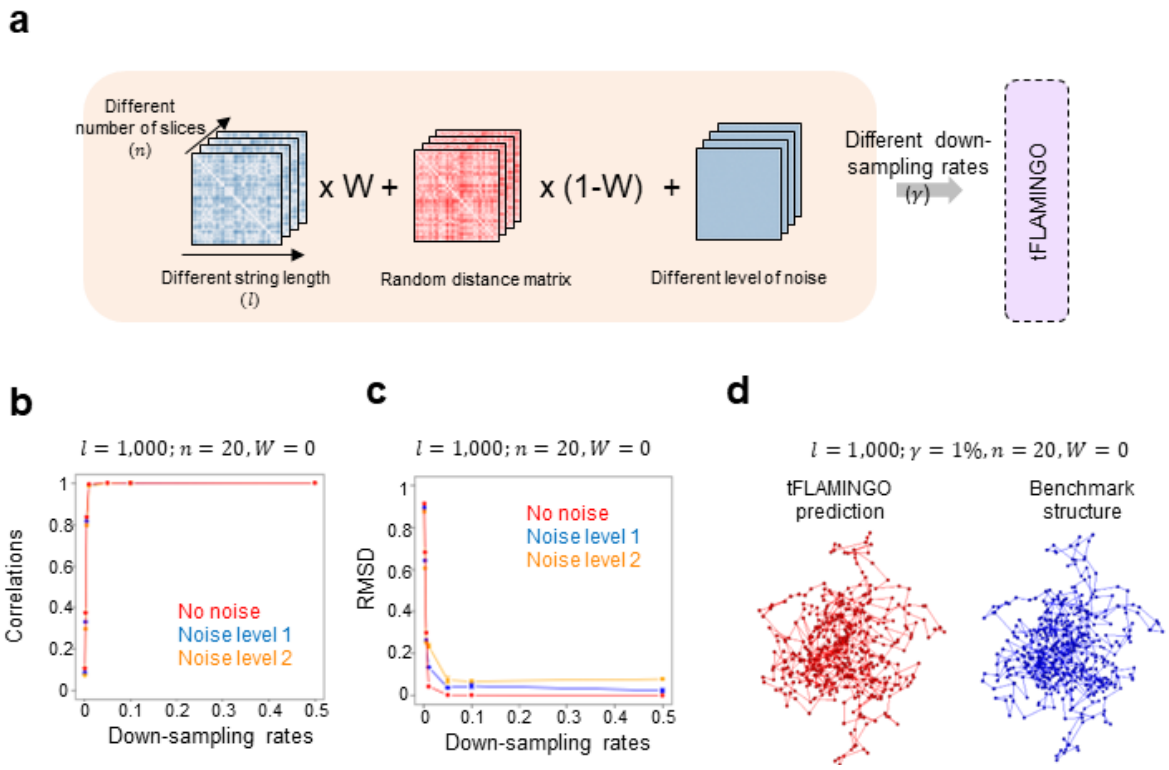



**Figure B.7 Robust performance of tFLAMINGO under different settings based on simulations.**

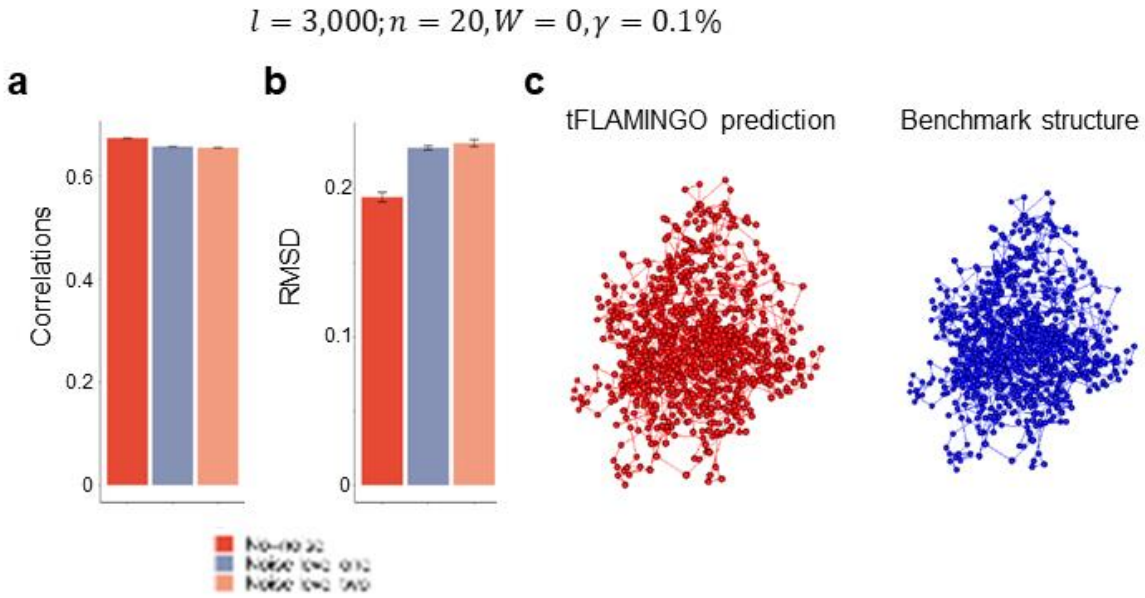



**Figure B.8 Accurate reconstruction of a simulated structure with 3000 loci under the 0.5% down sampling rate.**

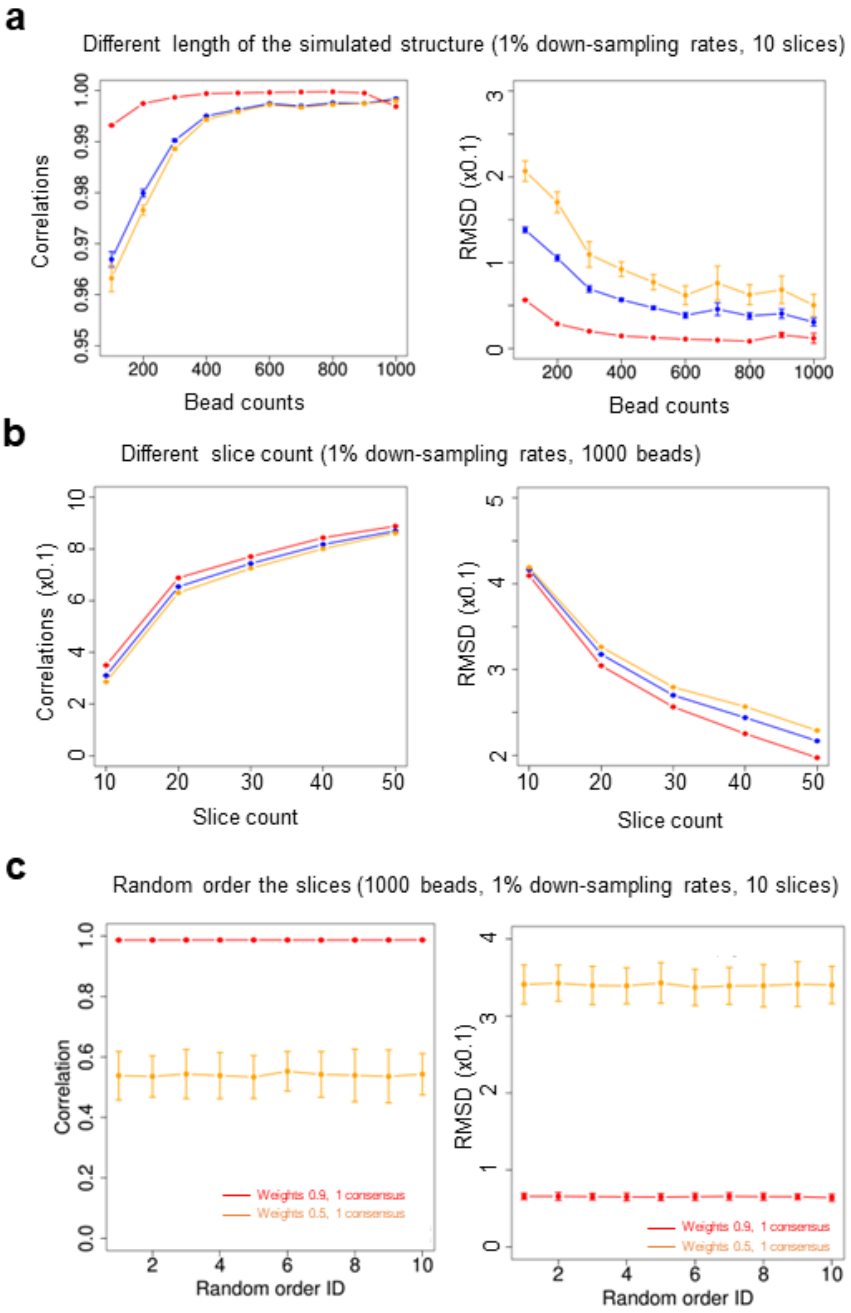



**Figure B.9 Systematic performance evaluation based on simulations.**

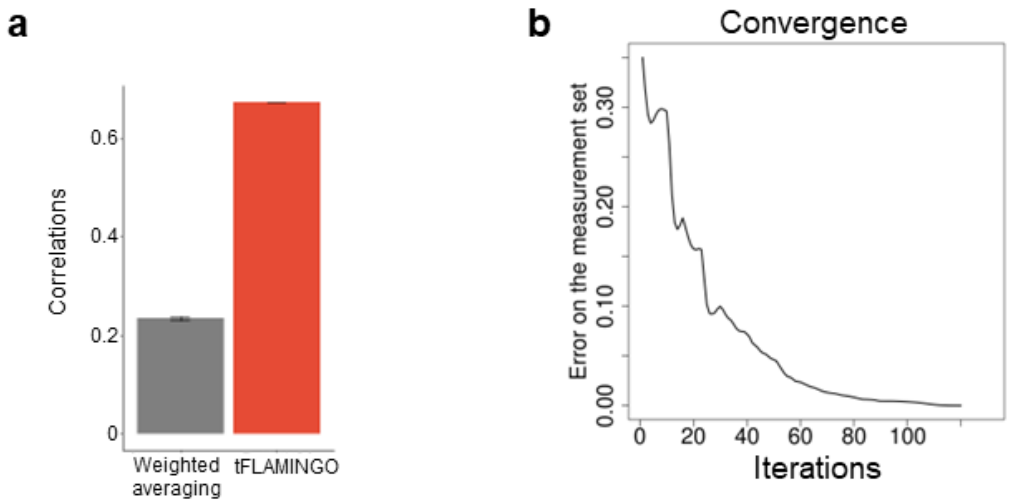



**Figure B.10 Convergence of tFLAMINGO.**

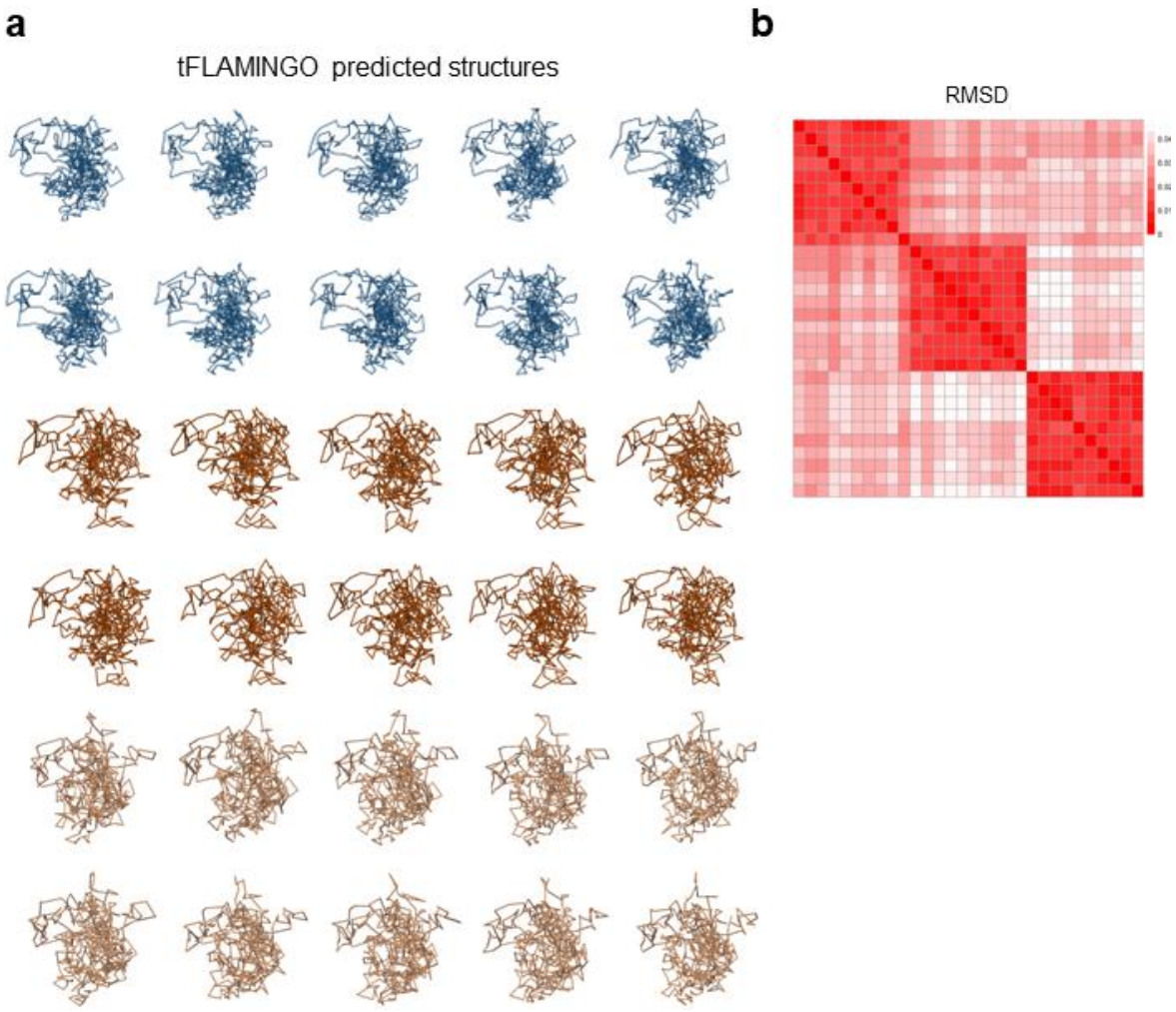



**Figure B.11 tFLAMINGO identifies underlying structural variations.**


**Figure B.12 Single-cell compartment and TAD analyses in GM12878.**

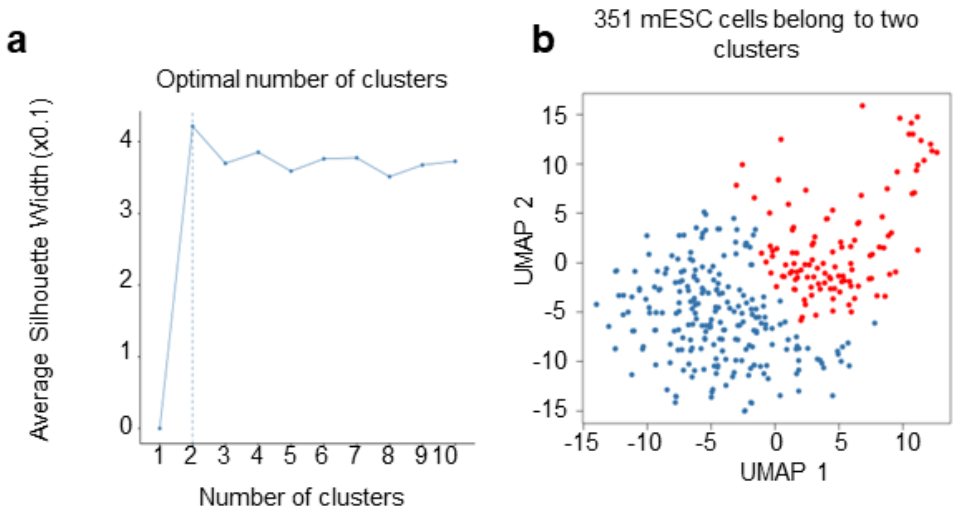



**Figure B.13 Justification of the optimal number of clusters.**

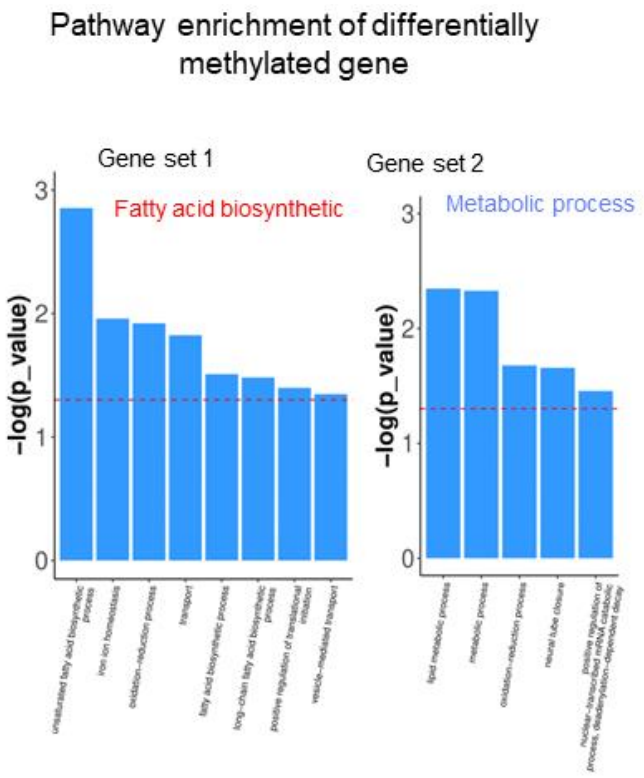



**Figure B.14 Pathway enrichments of differential methylated genes.**

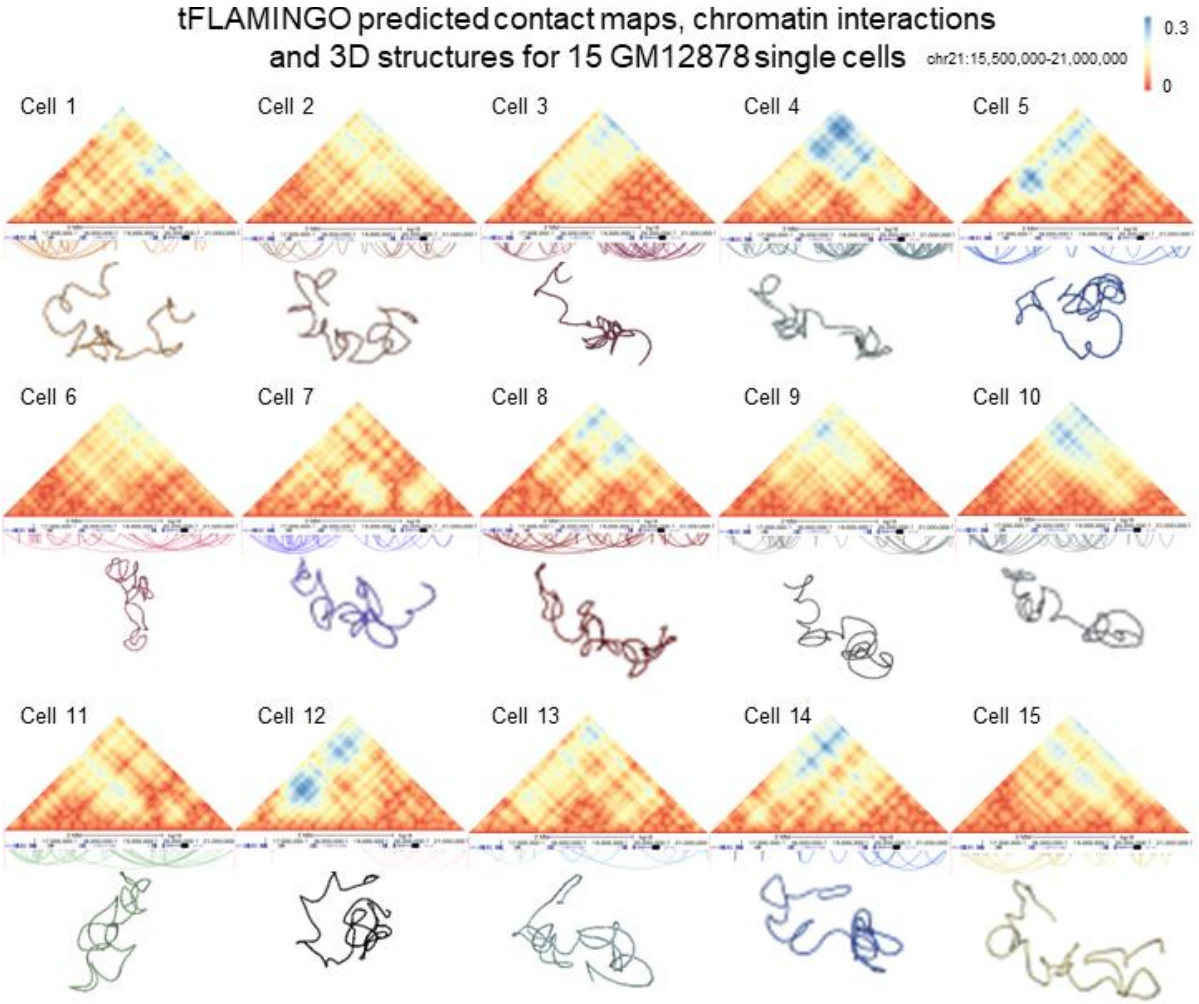



**Figure B.15 Dynamic 3D structures across 15 single cells.**

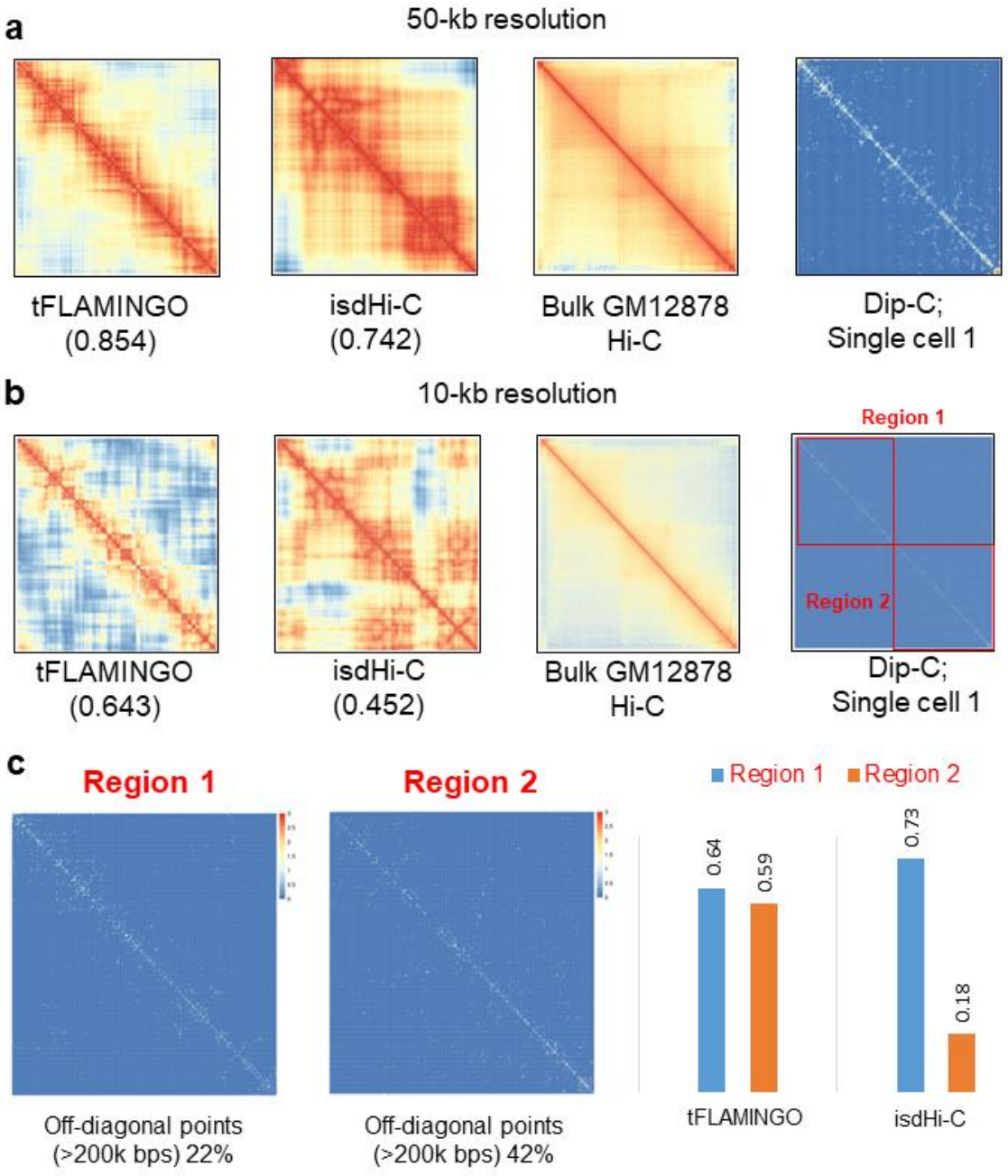



**Figure B.16 Simulation-based methods fail to handle long-range interactions.**

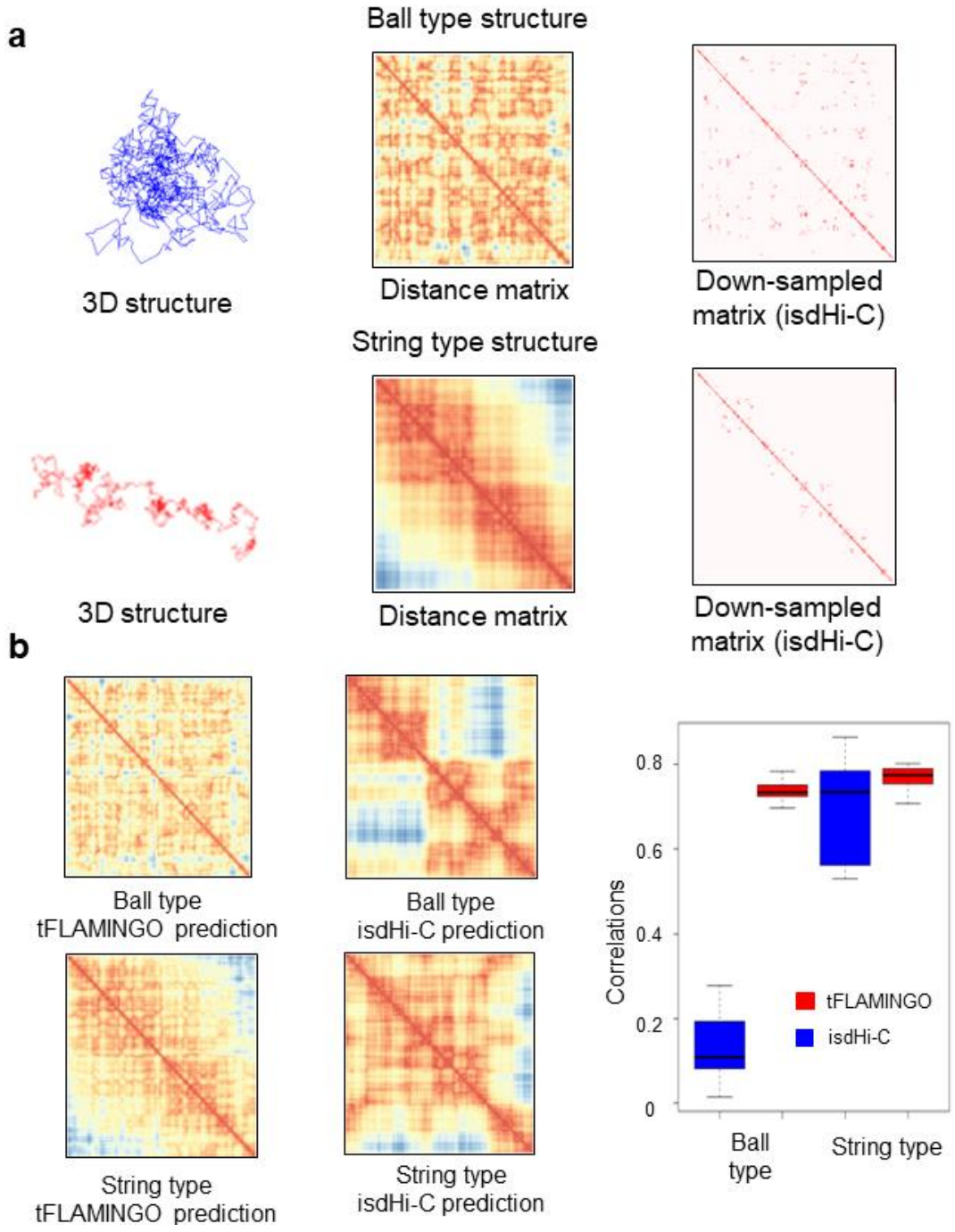



**Figure B.17 Simulation analyses confirms the limitation of simulation-based models.**

## APPENDIX C

## SUPPLEMENTARY FIGURES FOR CHAPTER 4

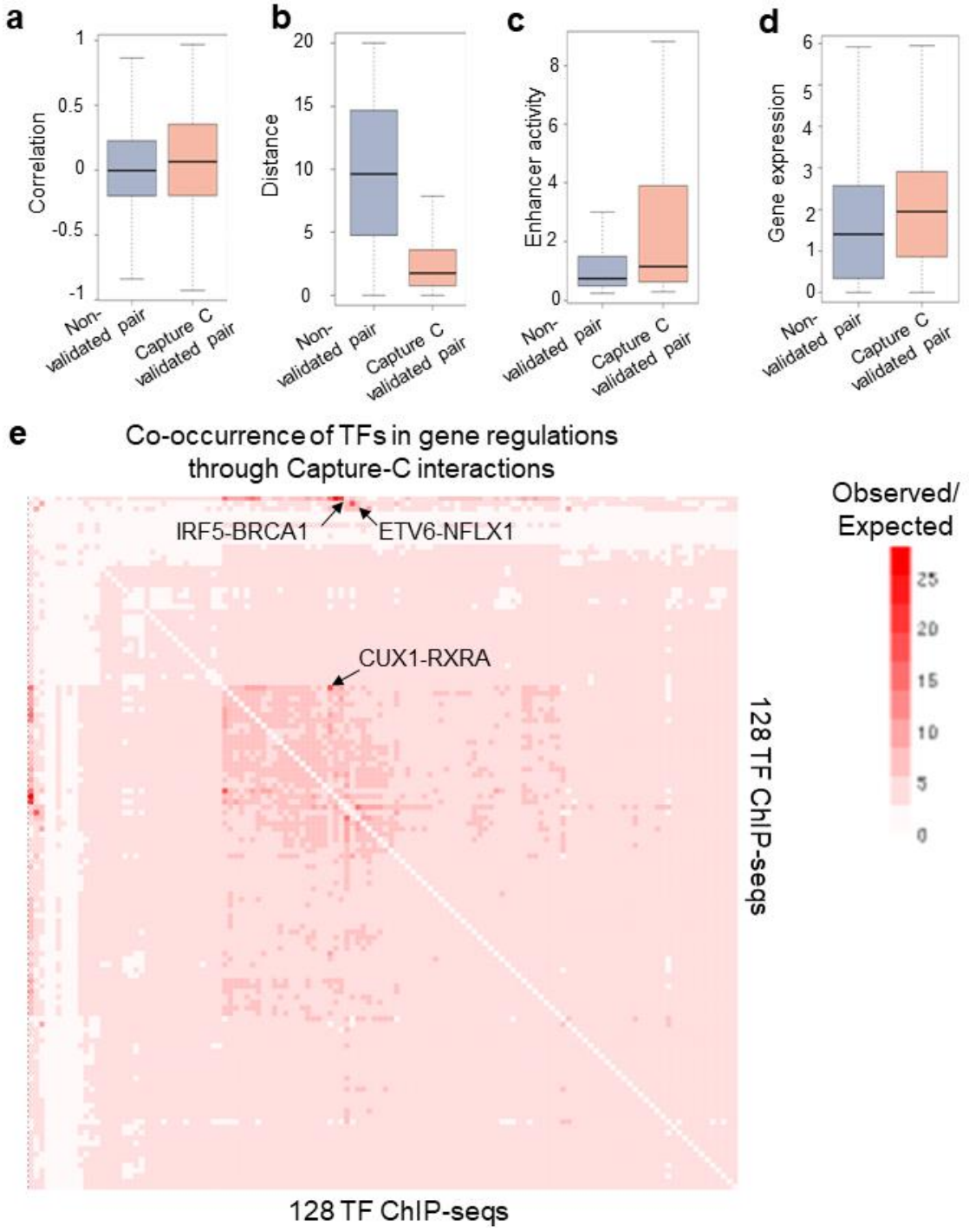



**Figure C.1 Predictive power of the features used in ComMUTE.**

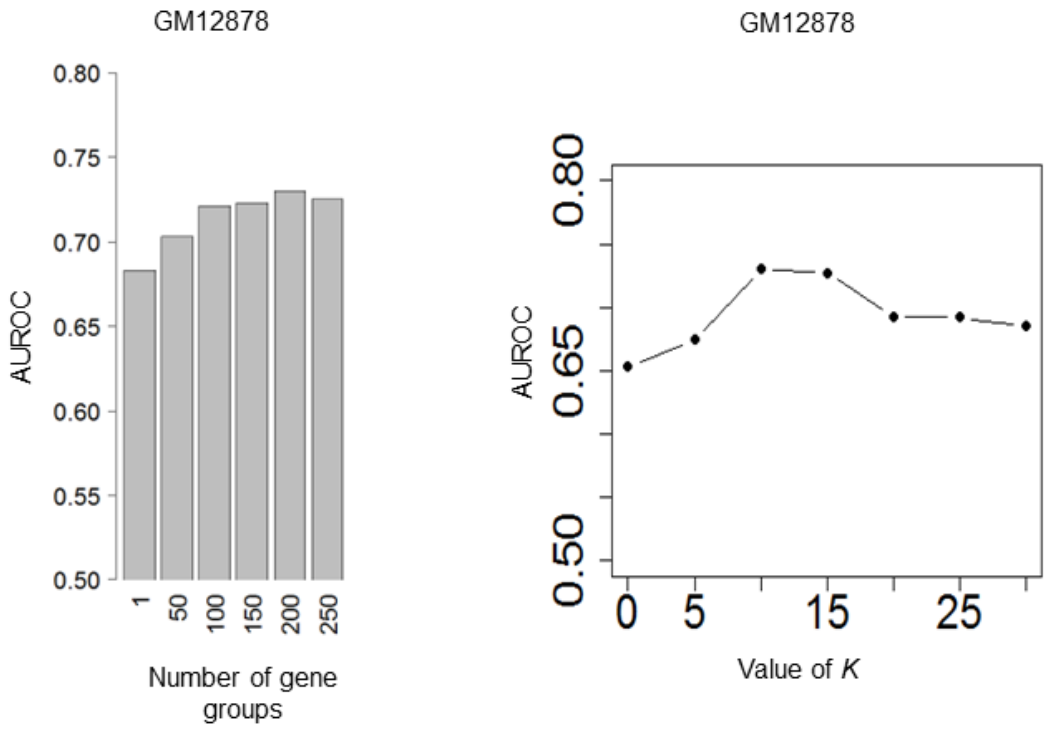



**Figure C.2 Parameter selection based on the optimal AUROC.**

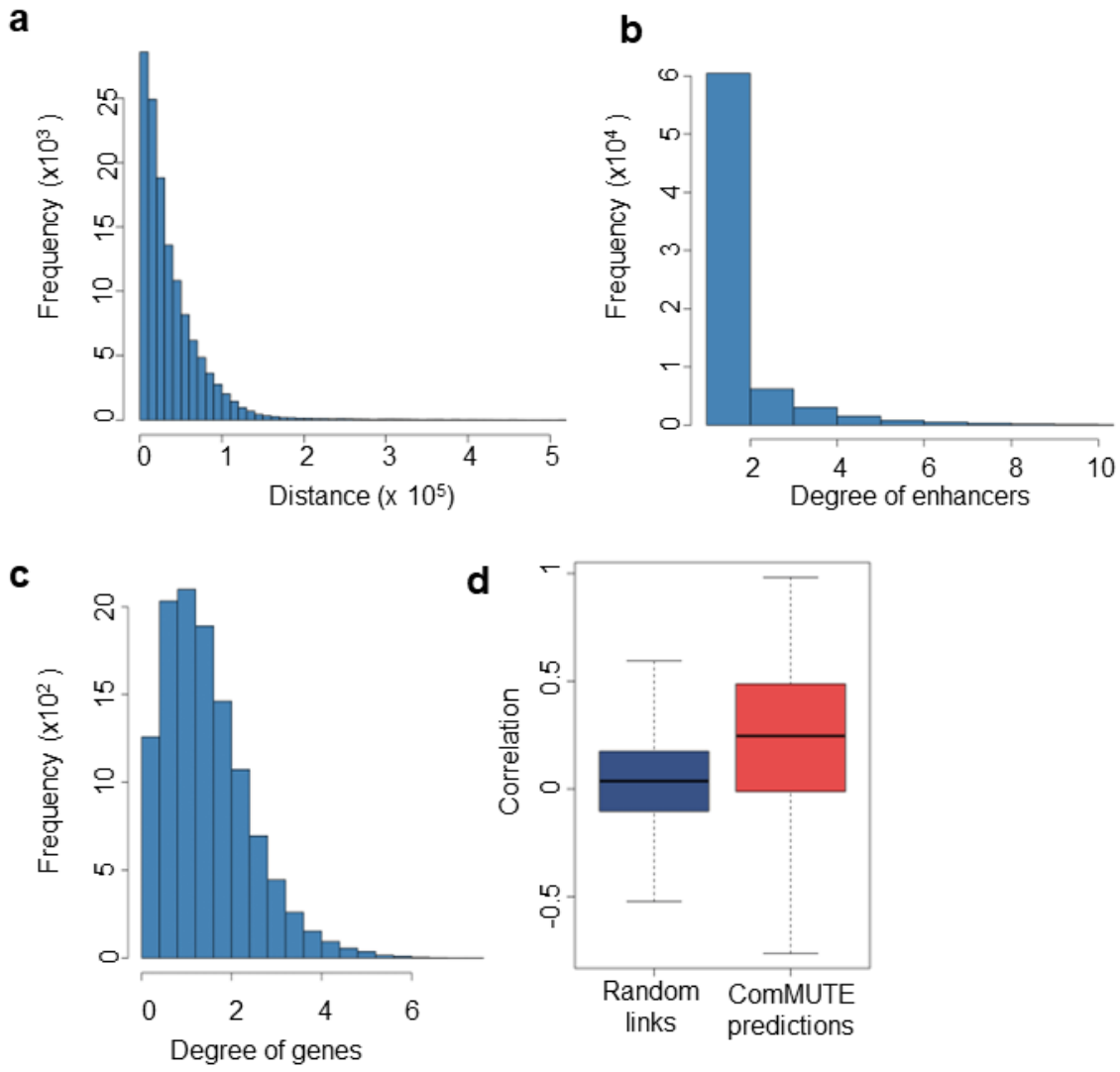



**Figure C.3 Summary statistics of the predicted enhancer-gene links.**

Summary of input epigenomics data datasets

| | Enhancer activity | Gene expression | Number of cell lines | |
|---|---|---|---|---|
| Version 1 | Imputed DNase-seq | Imputed RNA-seq | 127 | Imputed data |
| Version 2 | Imputed H3K27ac | Imputed RNA-seq | 127 | |
| Version 3 | Non-imputed DNase-seq | Non-imputed RNA-seq | 29 | Non-imputed data |
| Version 4 | Non-imputed H3K27ac | Non-imputed RNA-seq | 29 | |

**Figure C.4 Summary of the input epigenomic datasets.**

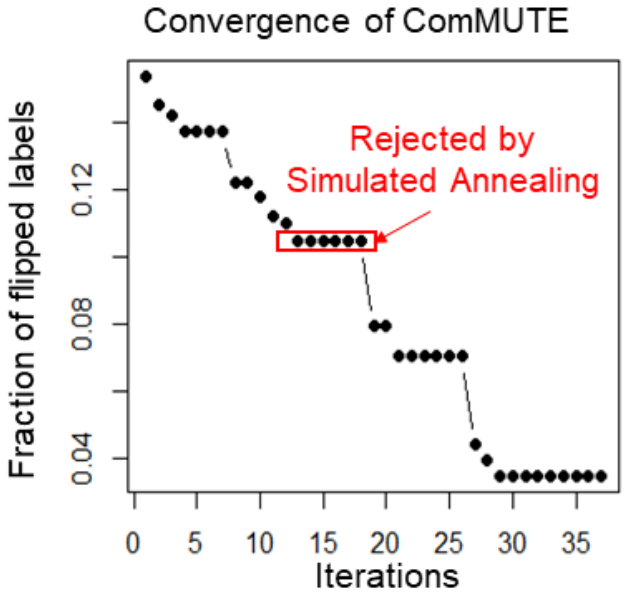



**Figure C.5 Convergence of ComMUTE.**

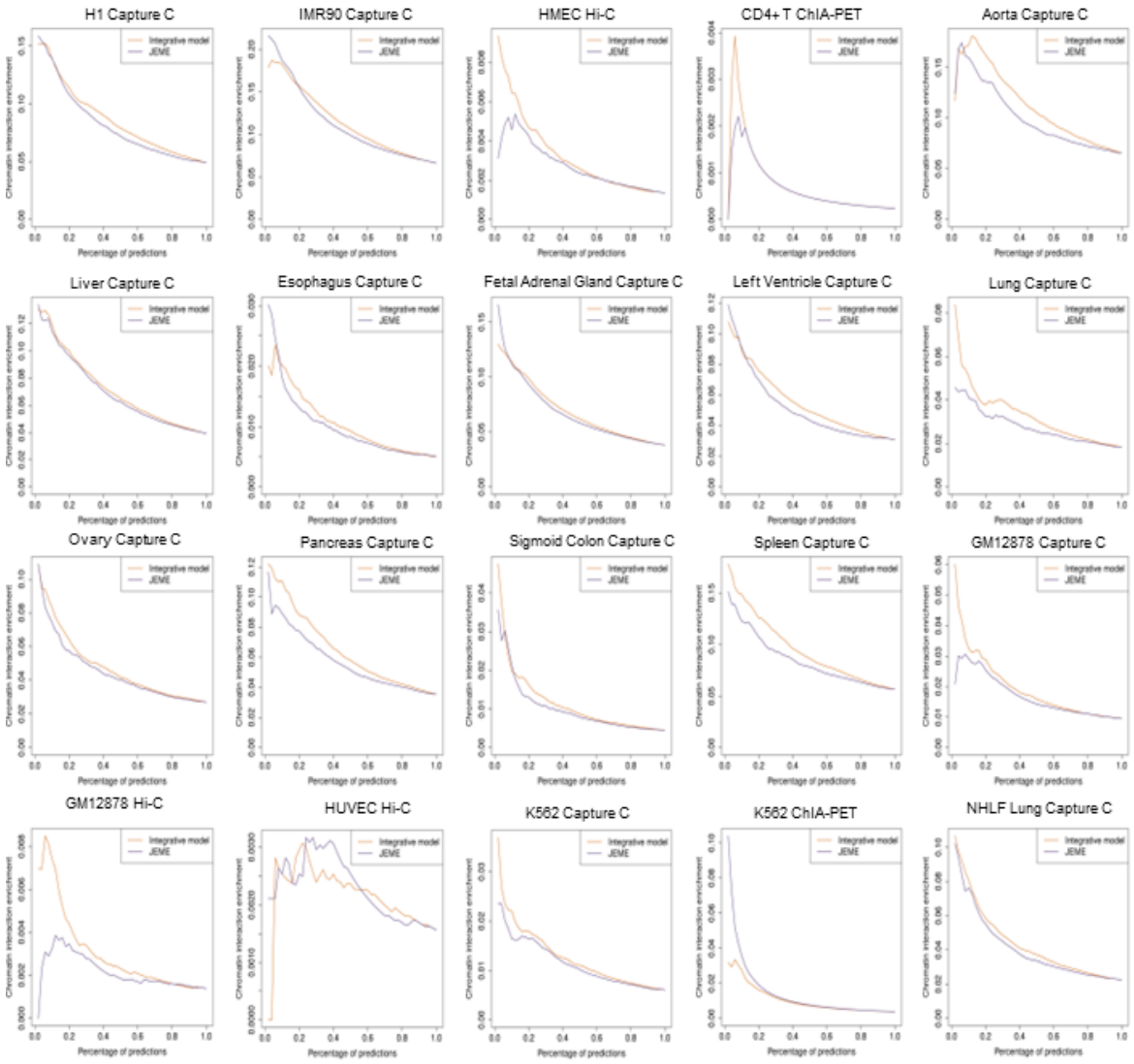



**Figure C.6 Performance comparison with JEME based on the enrichment analyses.**

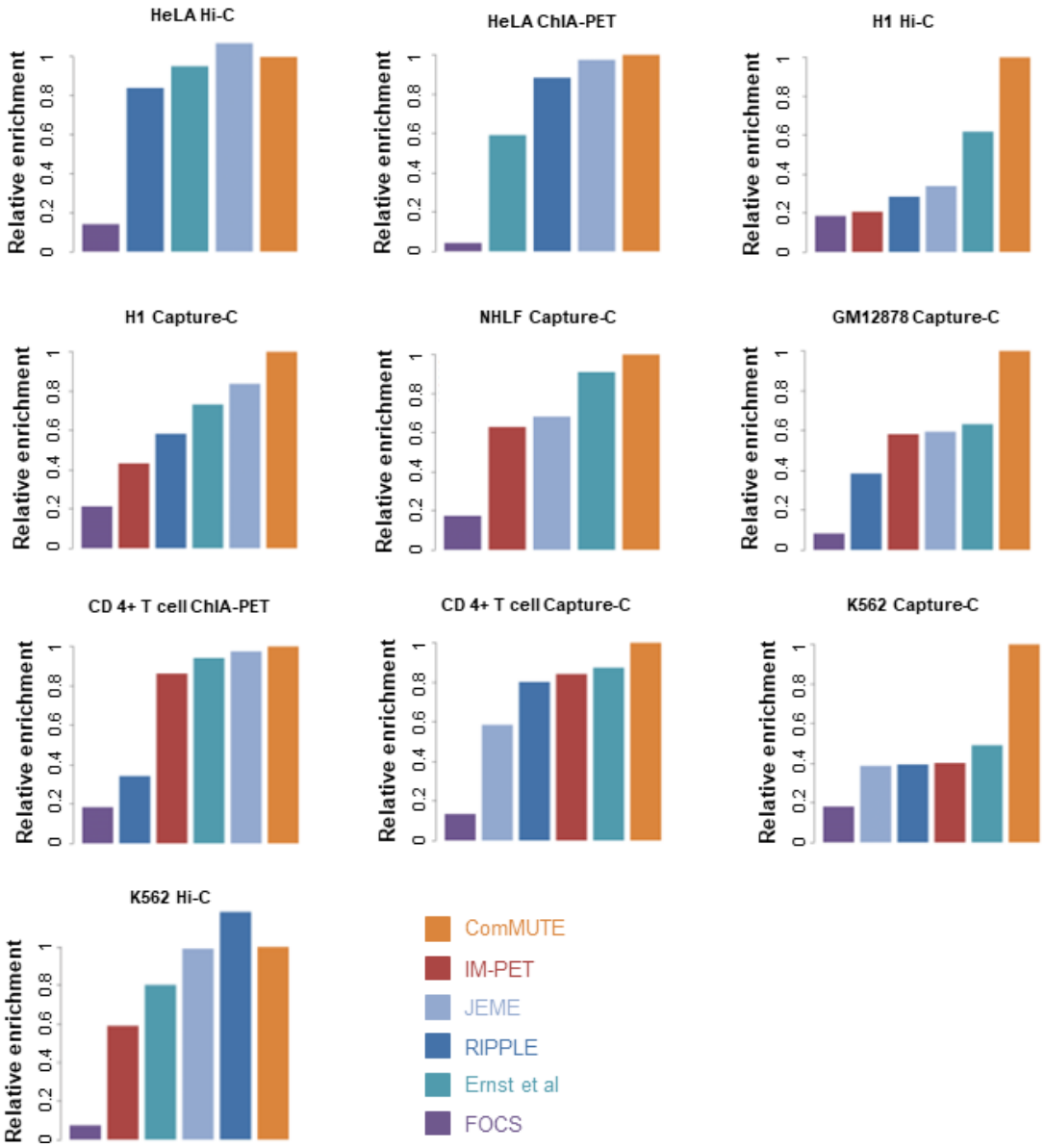



**Figure C.7 Performance comparison with existing methods based on the enrichment of experimental chromatin interactions.**

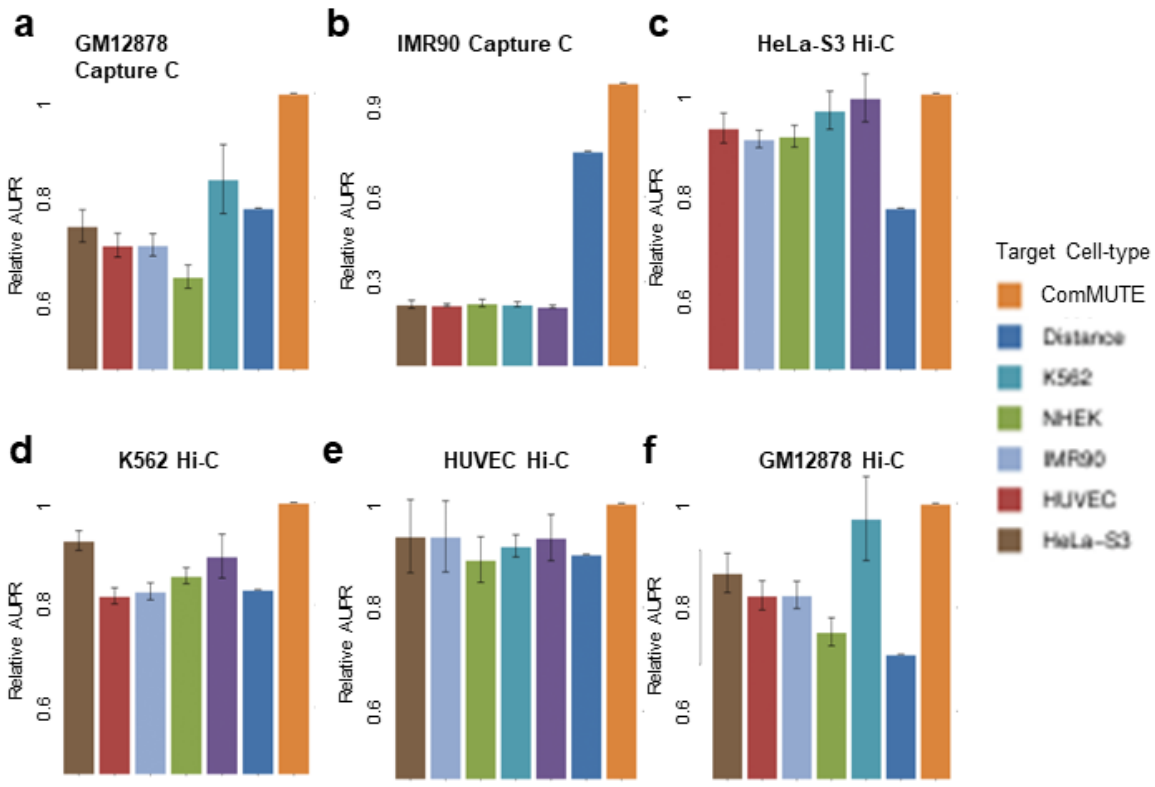



**Figure C.8 Cross-cell-type comparison with TargetFinder.**

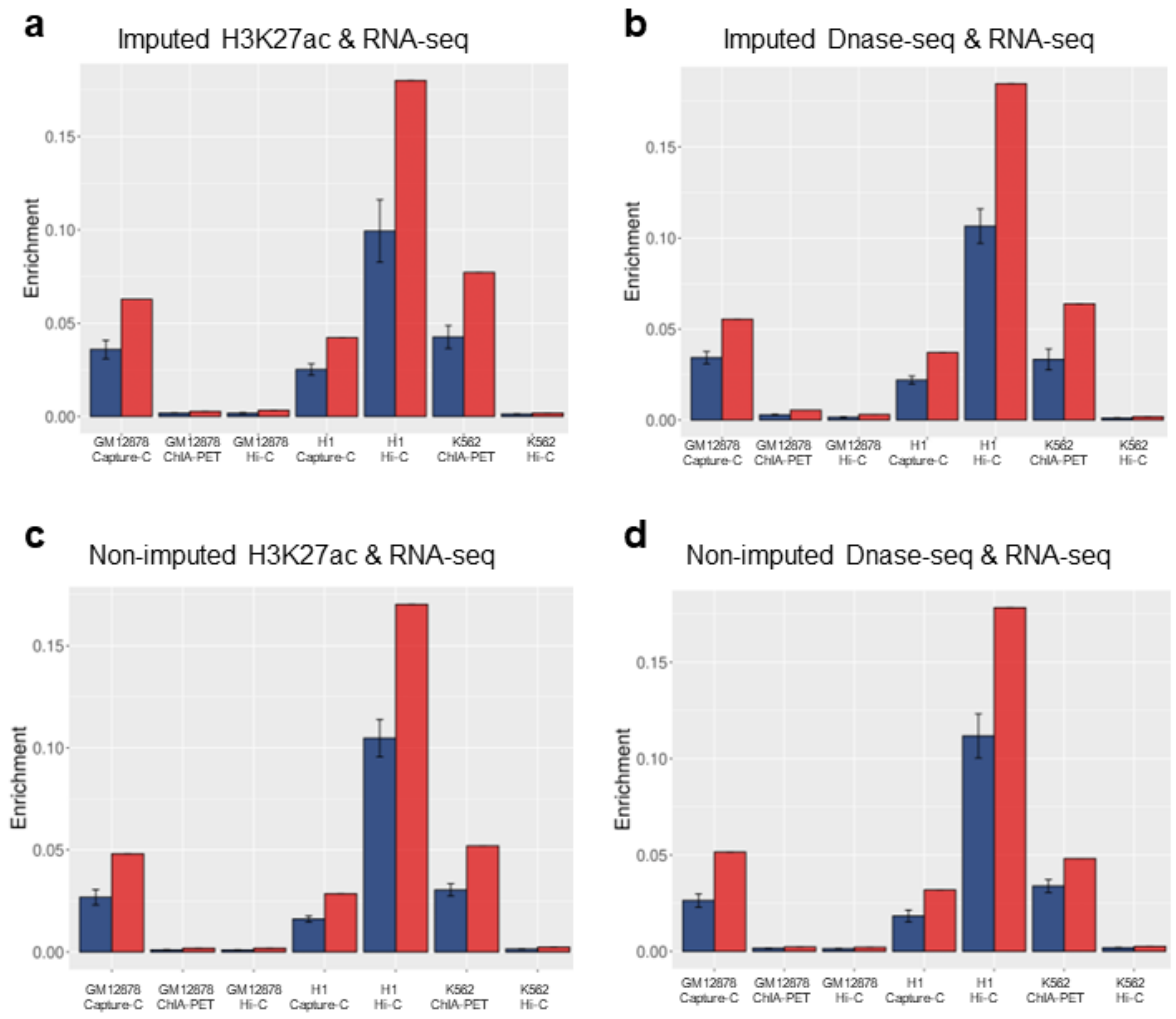



**Figure C.9 Evaluating the accuracy of predicted enhancer-gene links based on different epigenomic datasets.**

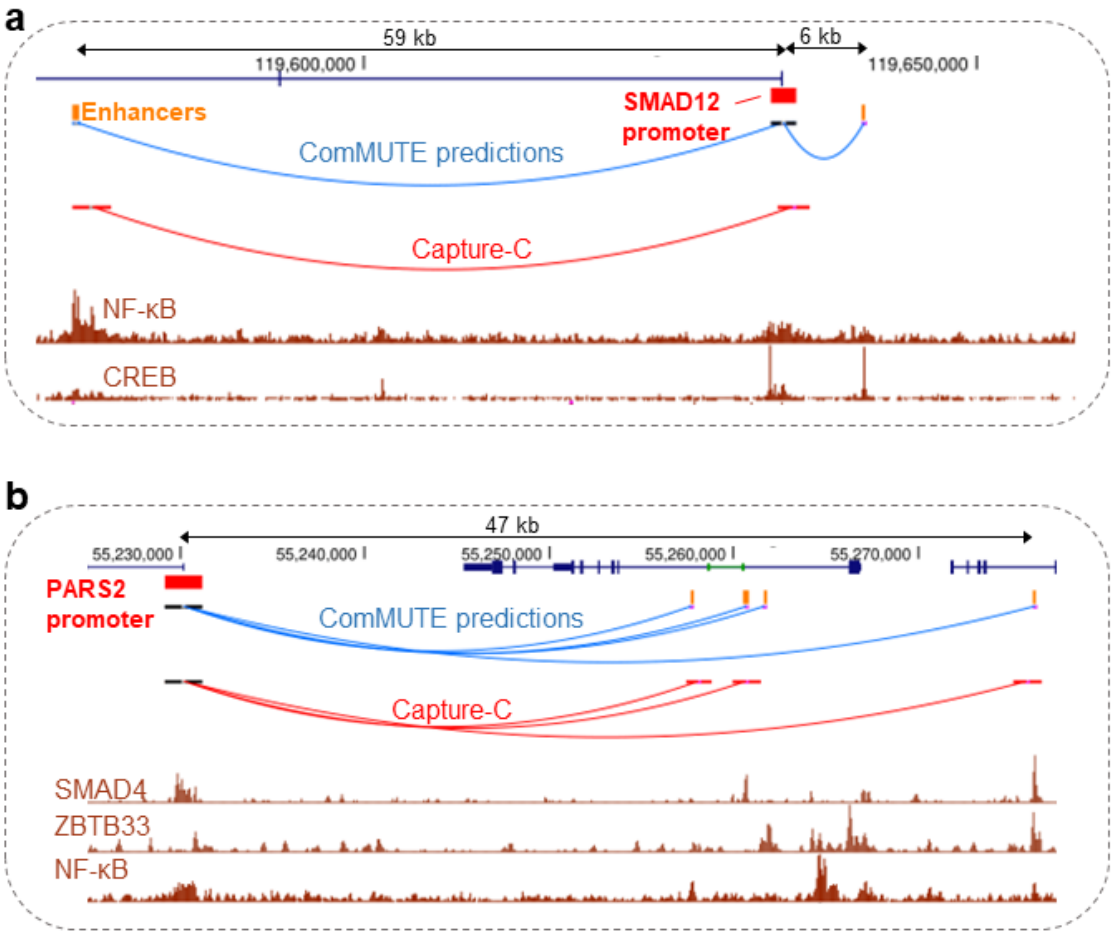



**Figure C.10 Example of predicted multi-enhancer regulations.**

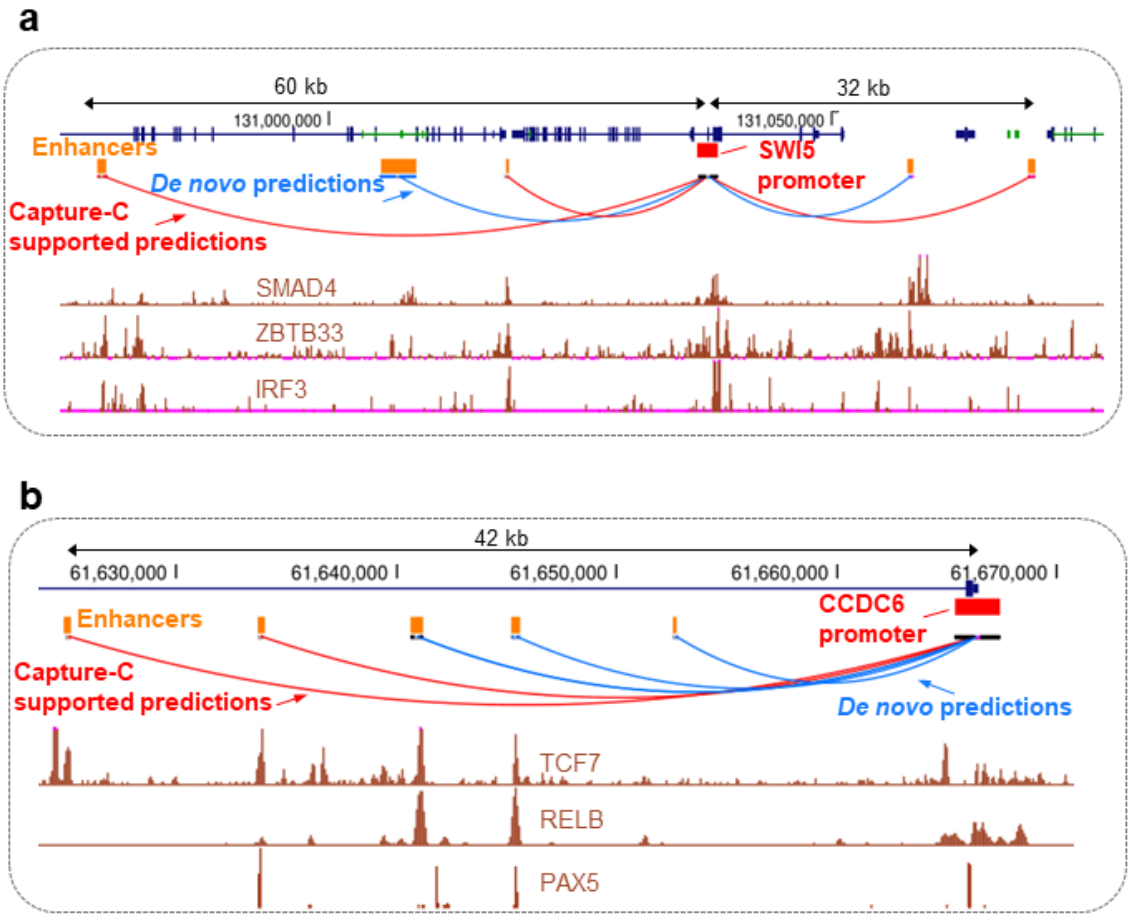



**Figure C.11 Example of predicted multi-enhancer regulations.**

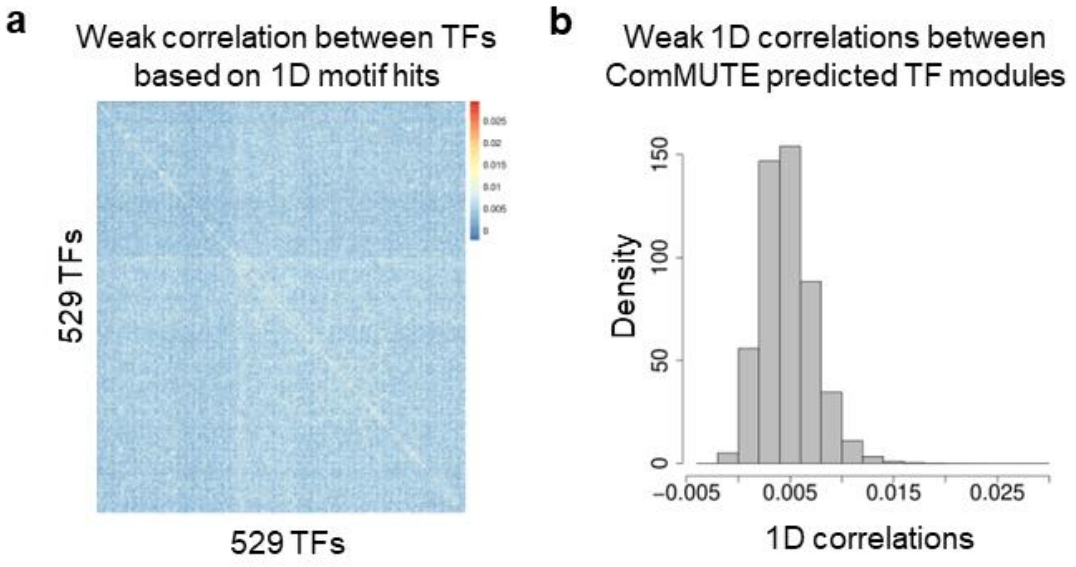



**Figure C.12 Co-binding analysis based on TF motif occurrence.**

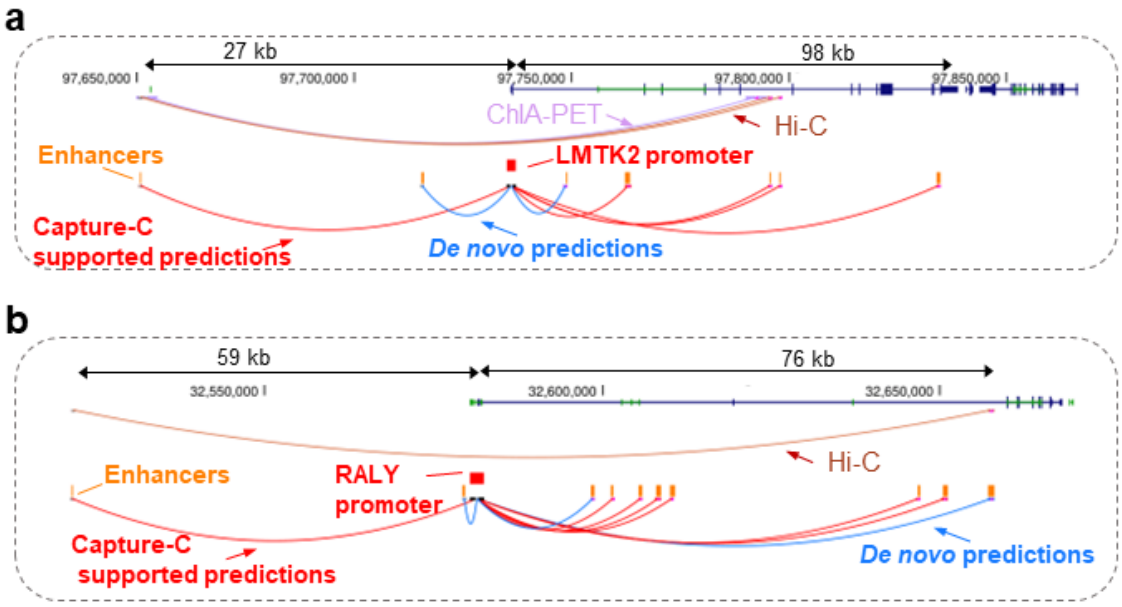



**Figure C.13 Example of direct chromatin interactions between co-regulating enhancers.**

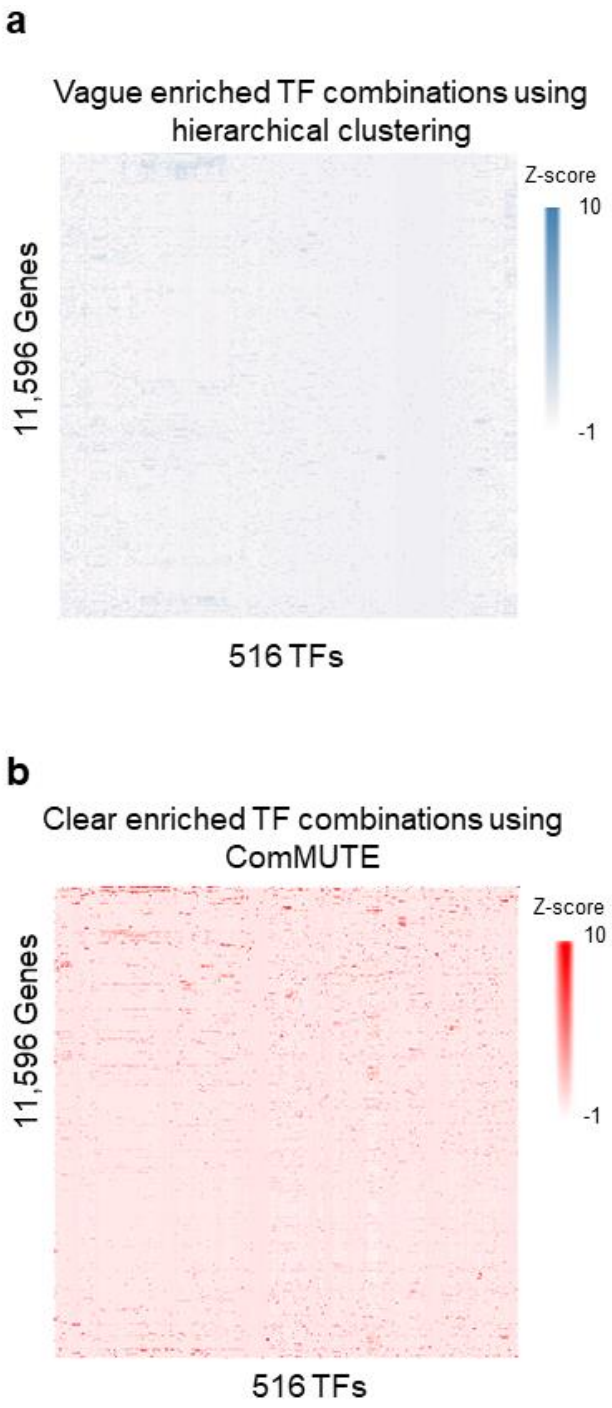



**Figure C.14 ComMUTE discovers clear TF grammars for gene regulations.**

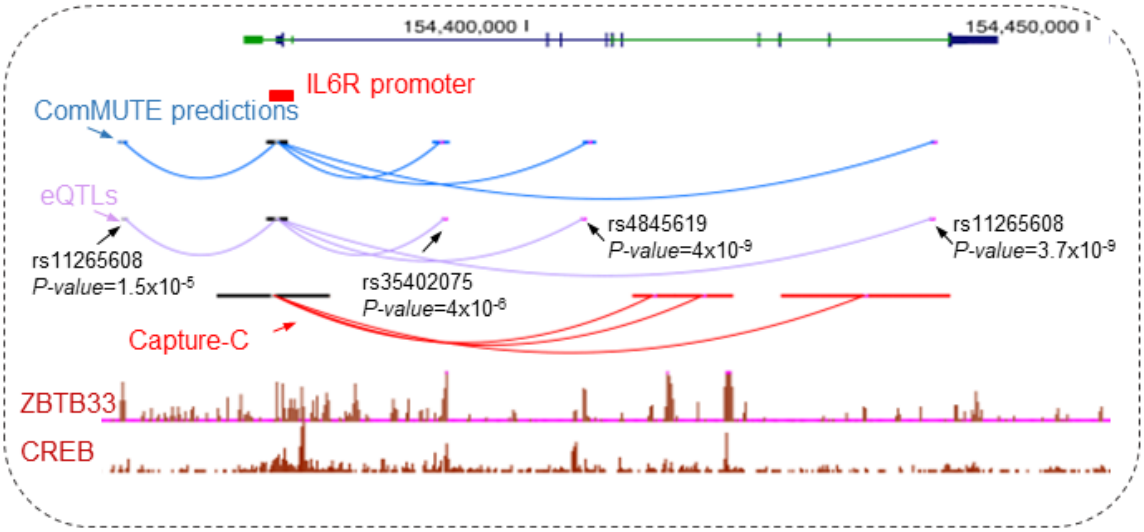



**Figure C.15 Convergence of ComMUTE.**

## APPENDIX D

## SUPPLEMENTARY FIGURES FOR CHAPTER 5

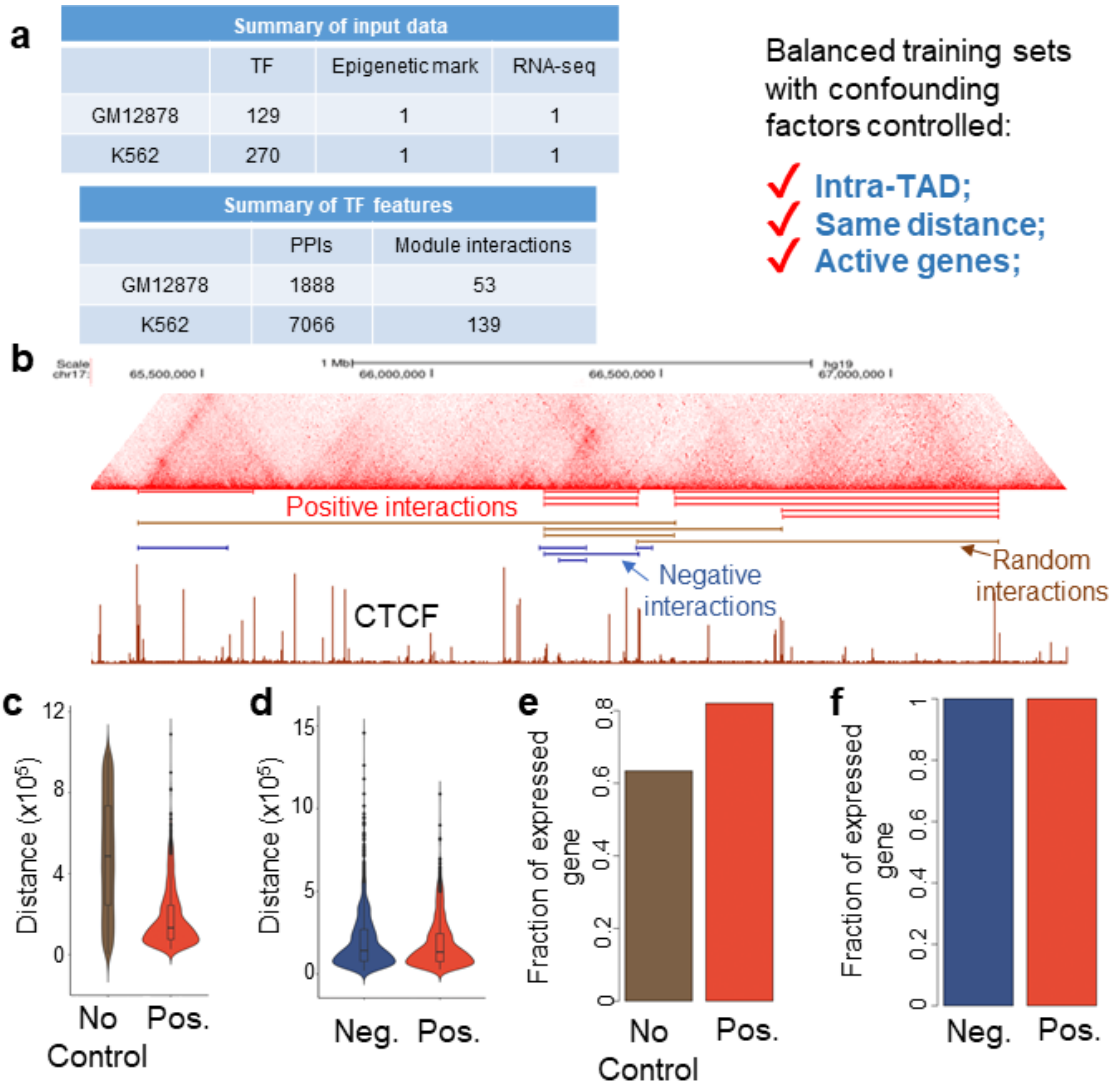



**Figure D.1 Summary of training dataset generation and confounding factor controls.**

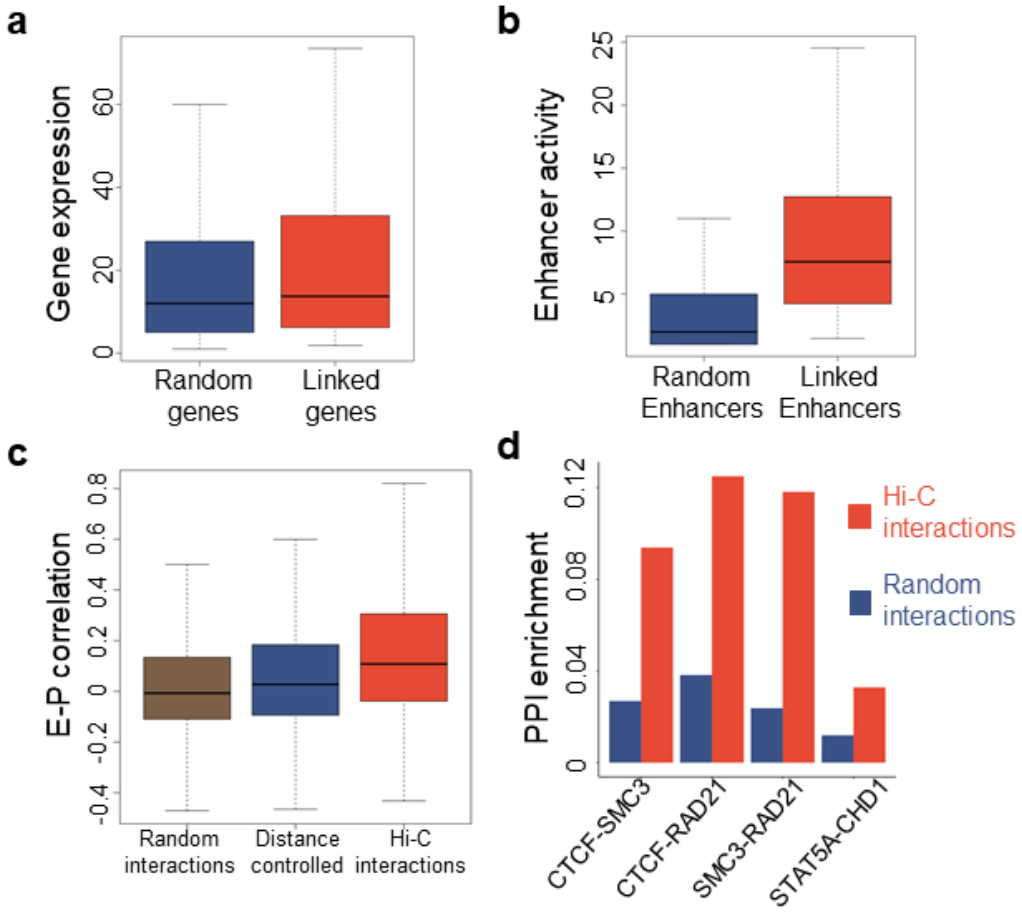



**Figure D.2 Predictive power of features are supported by the differential distributions of features.**

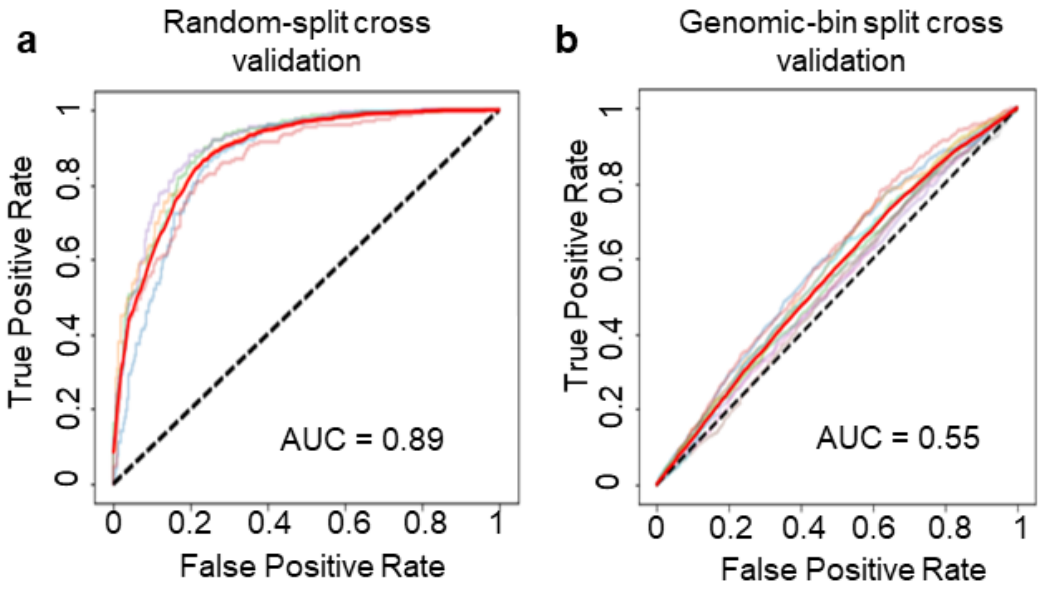



**Figure D.3 Advanced feature dimension reduction is needed due to the risk of overfitting.**

a

| p | Module count (GM12878) | Module count (K562) |
|---|---|---|
| 10 | 831 | 1083 |
| 20 | 401 | 526 |
| 30 | 134 | 419 |

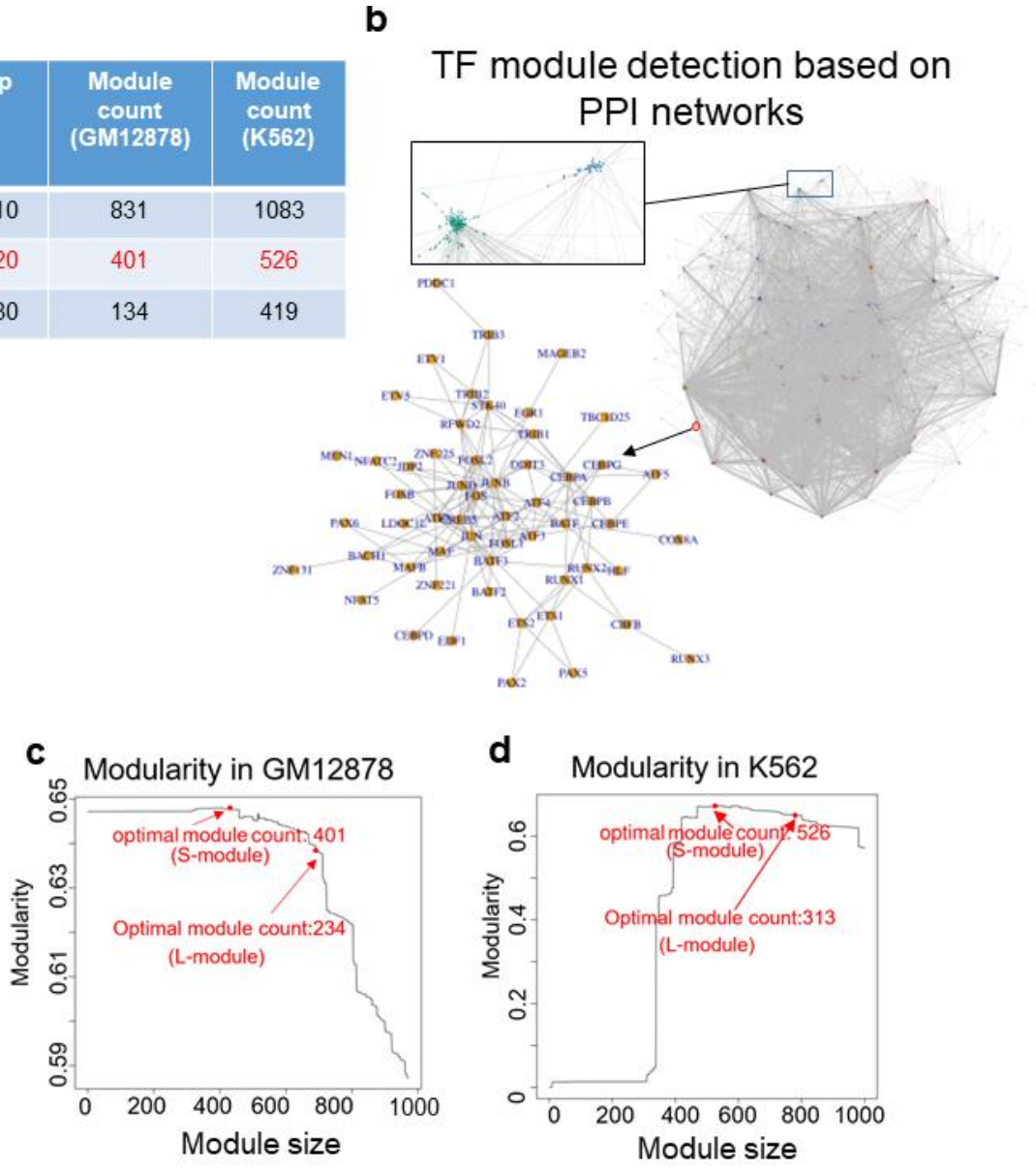



**Figure D.4 Hierarchical network-community detection based on the PPI network to construct model-level TF PPI features.**

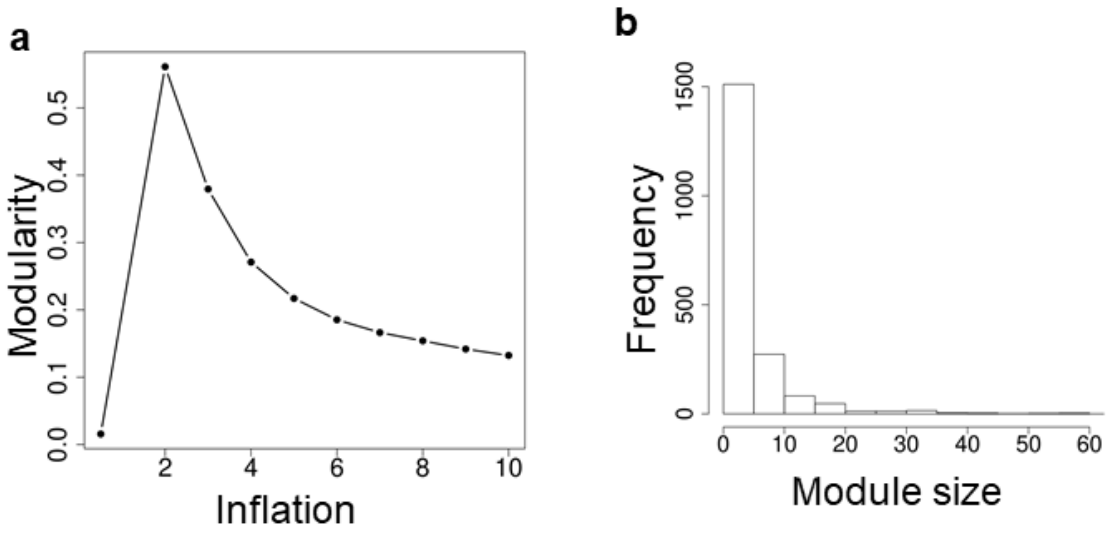



**Figure D.5 PPI community detection based on the MCL.**

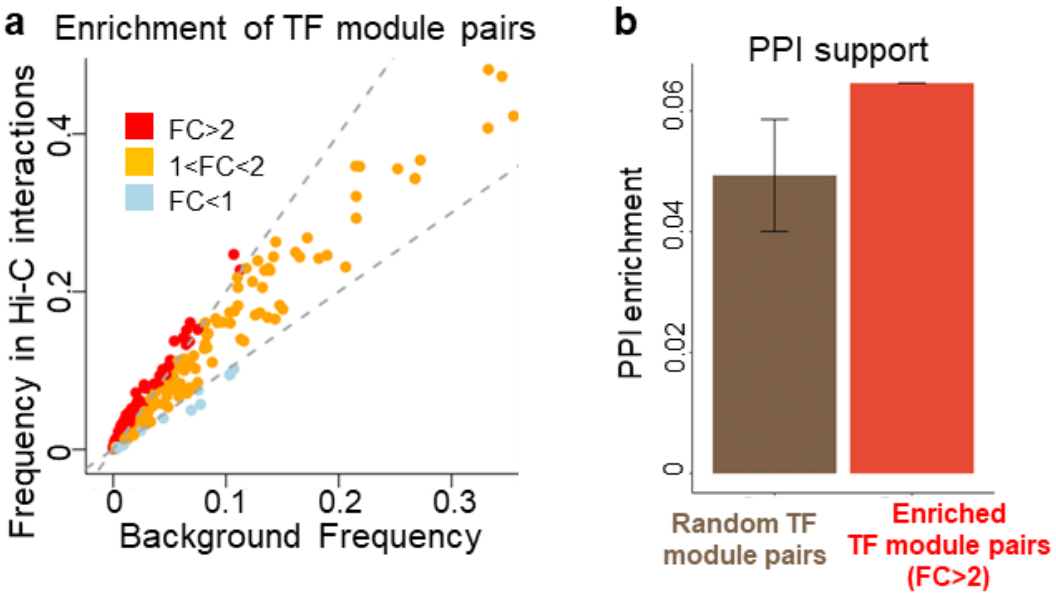



**Figure D.6 Enrichment analysis and PPI support analysis for TF module pairs.**

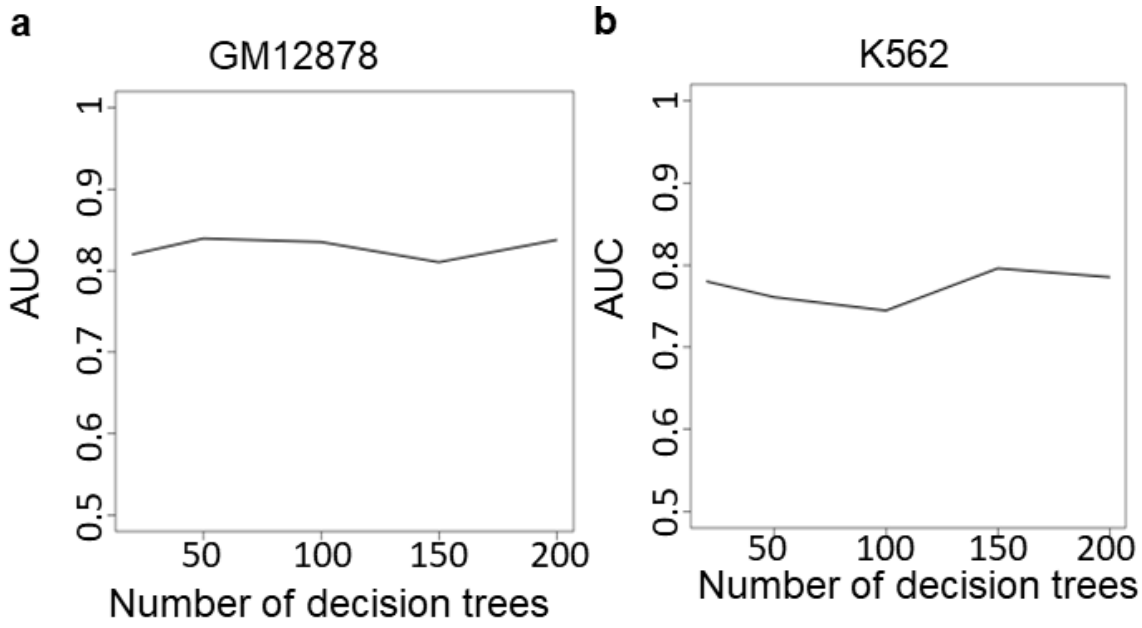



**Figure D.7 Model performance as a function of the number of decision trees.**

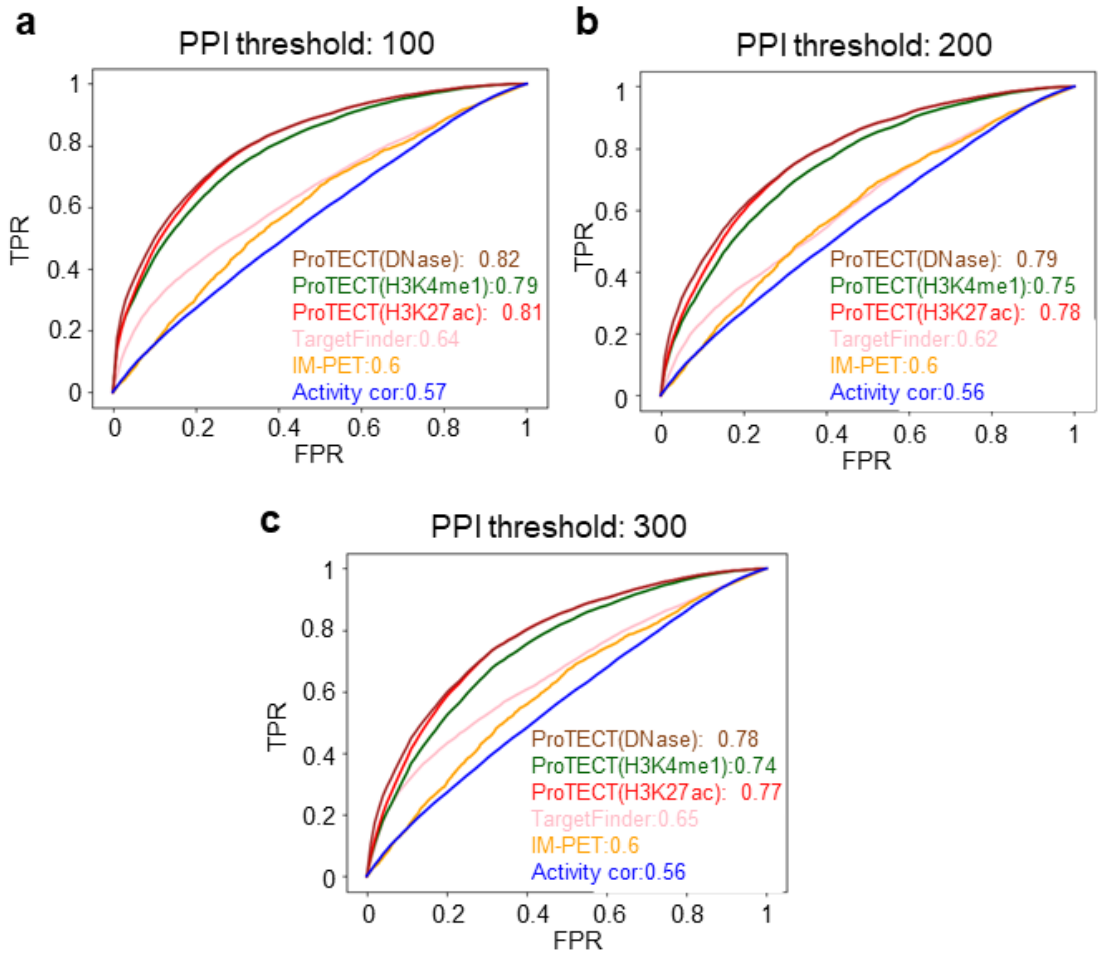



**Figure D.8 Performance of ProTECT using different epigenomic signals.**

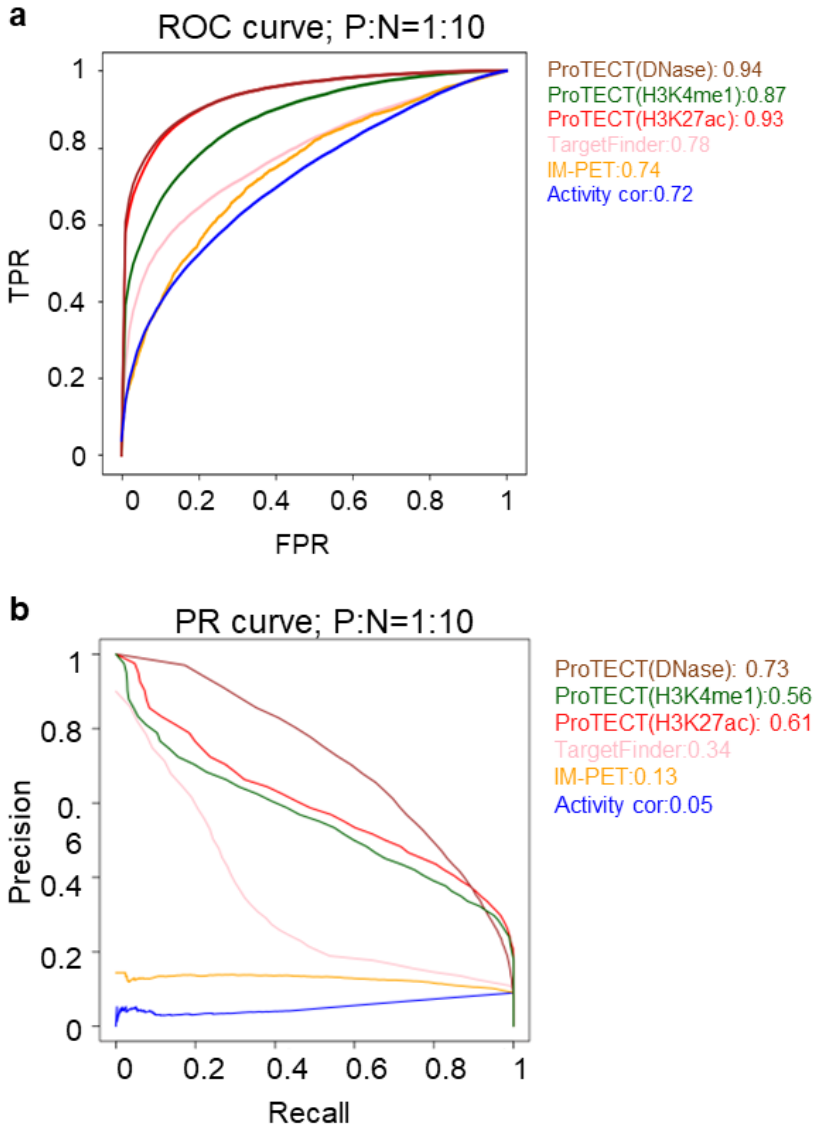



**Figure D.9 Performance comparison based on the imbalanced training data and the genomic bin-split cross-validation.**

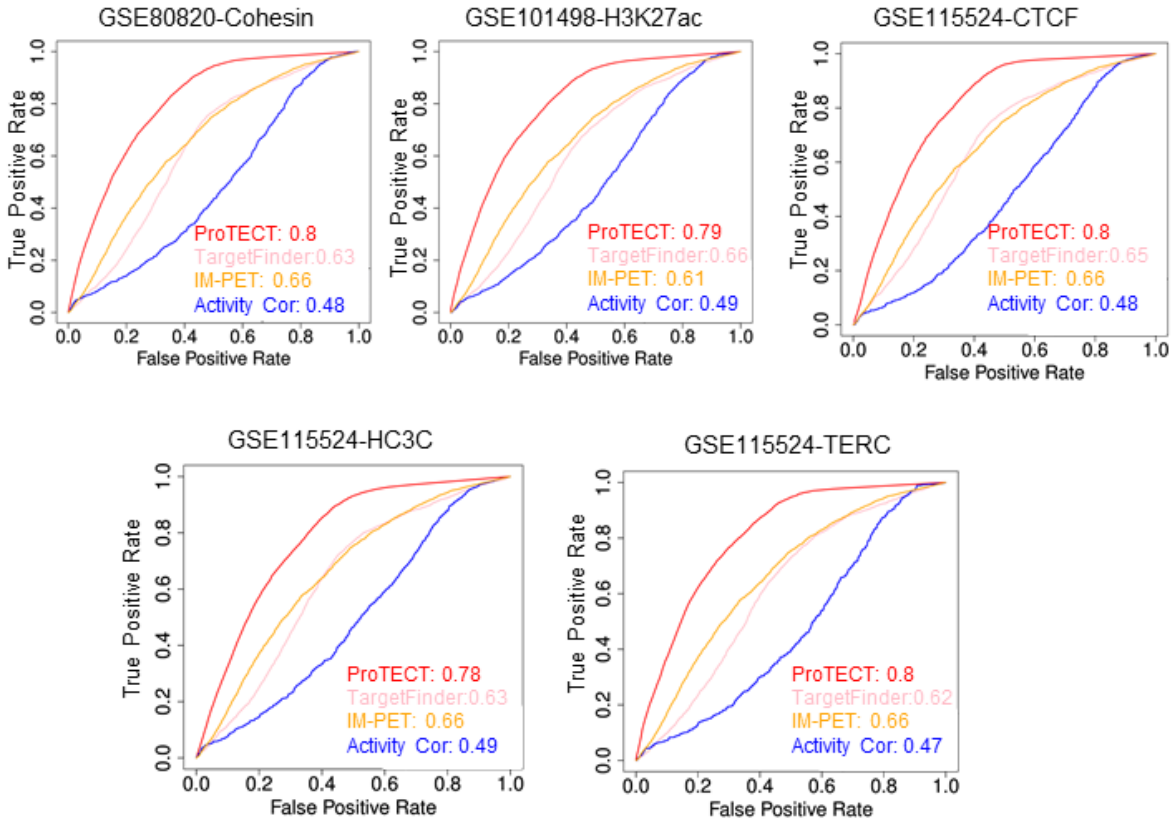



**Figure D.10 Performance comparison using five Hi-ChIP datasets.**

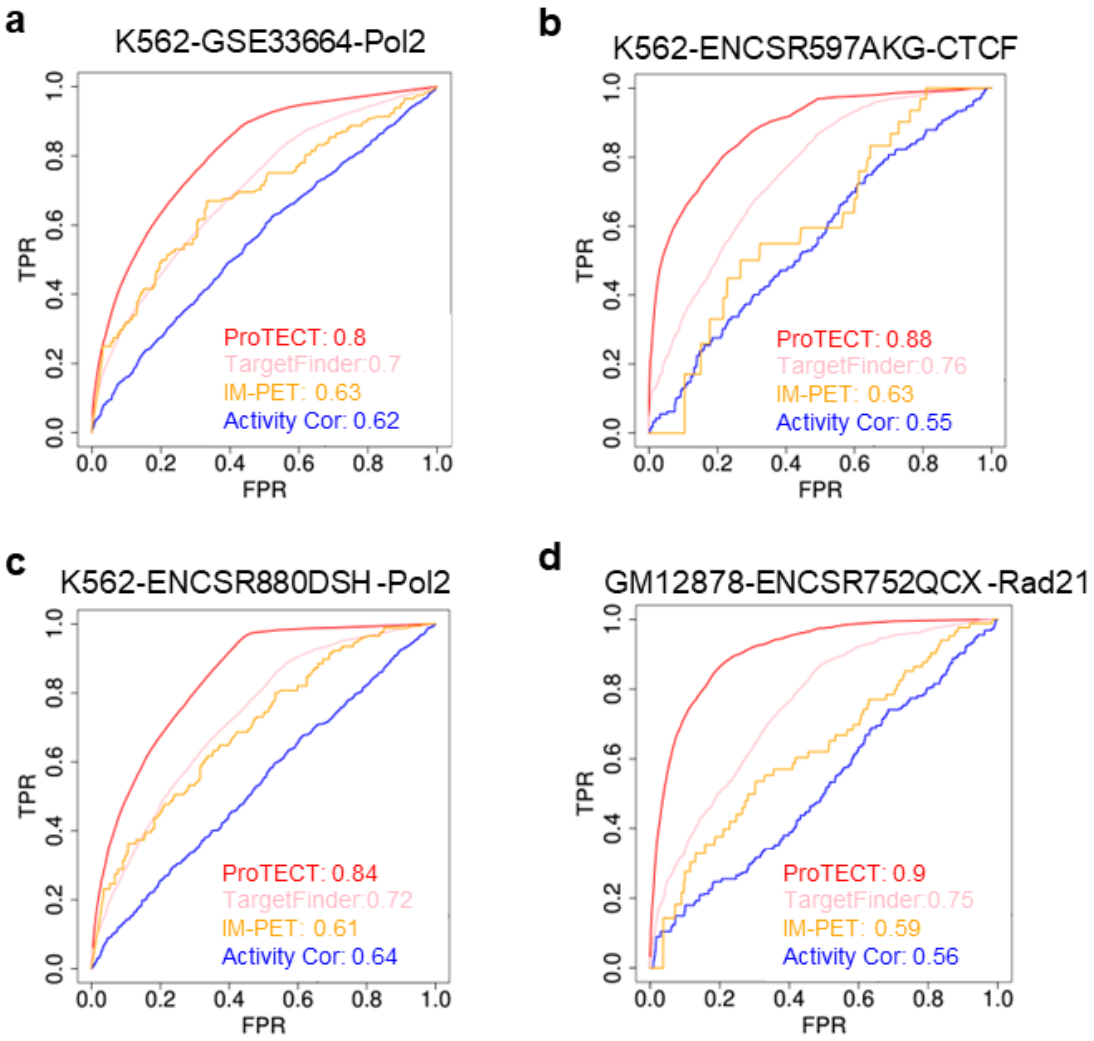



**Figure D.11 Performance comparison using four different ChIA-PET datasets.**

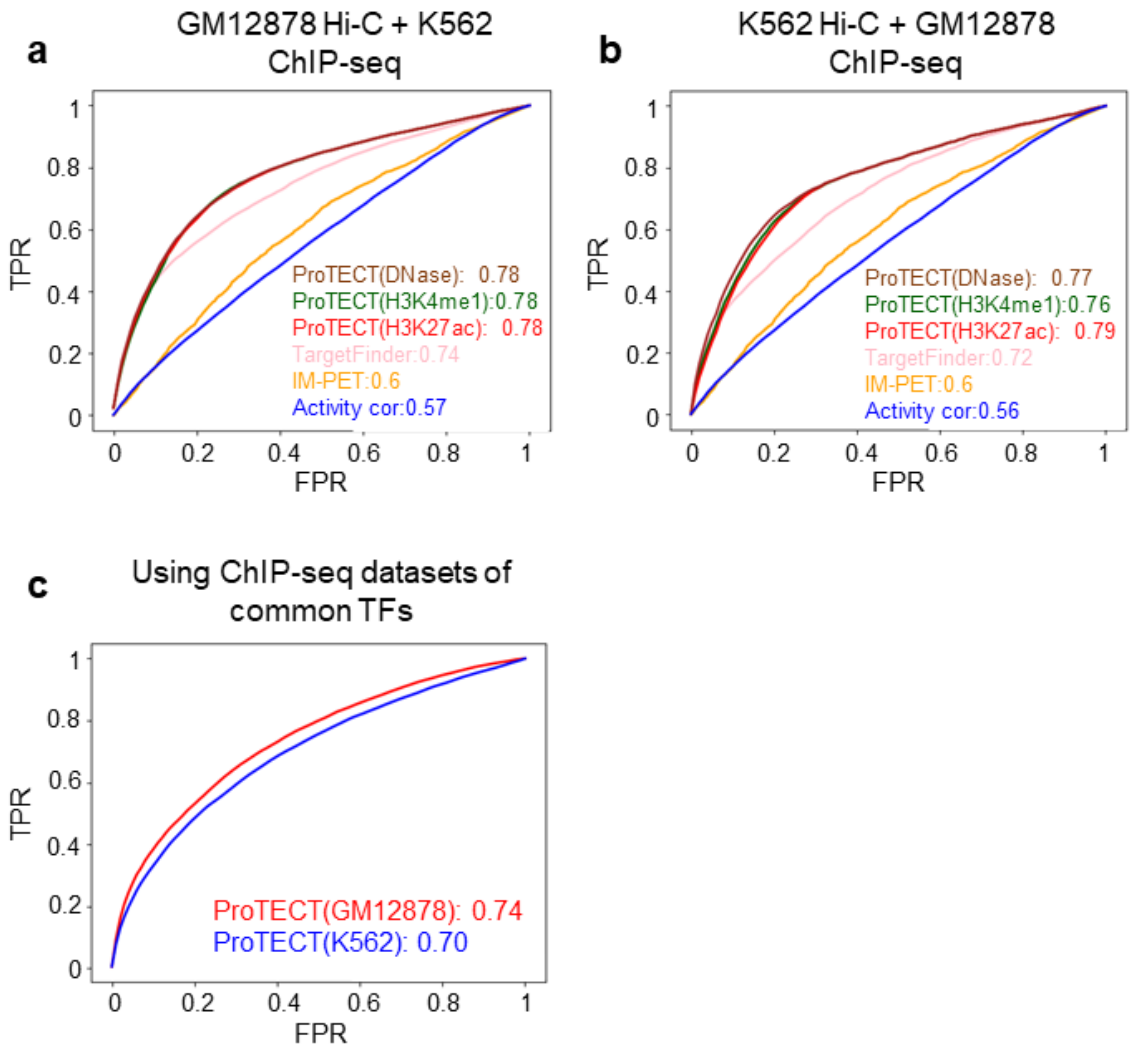



**Figure D.12 Performance comparison based on different combinations of Hi-C data and TF ChIP-seq data.**

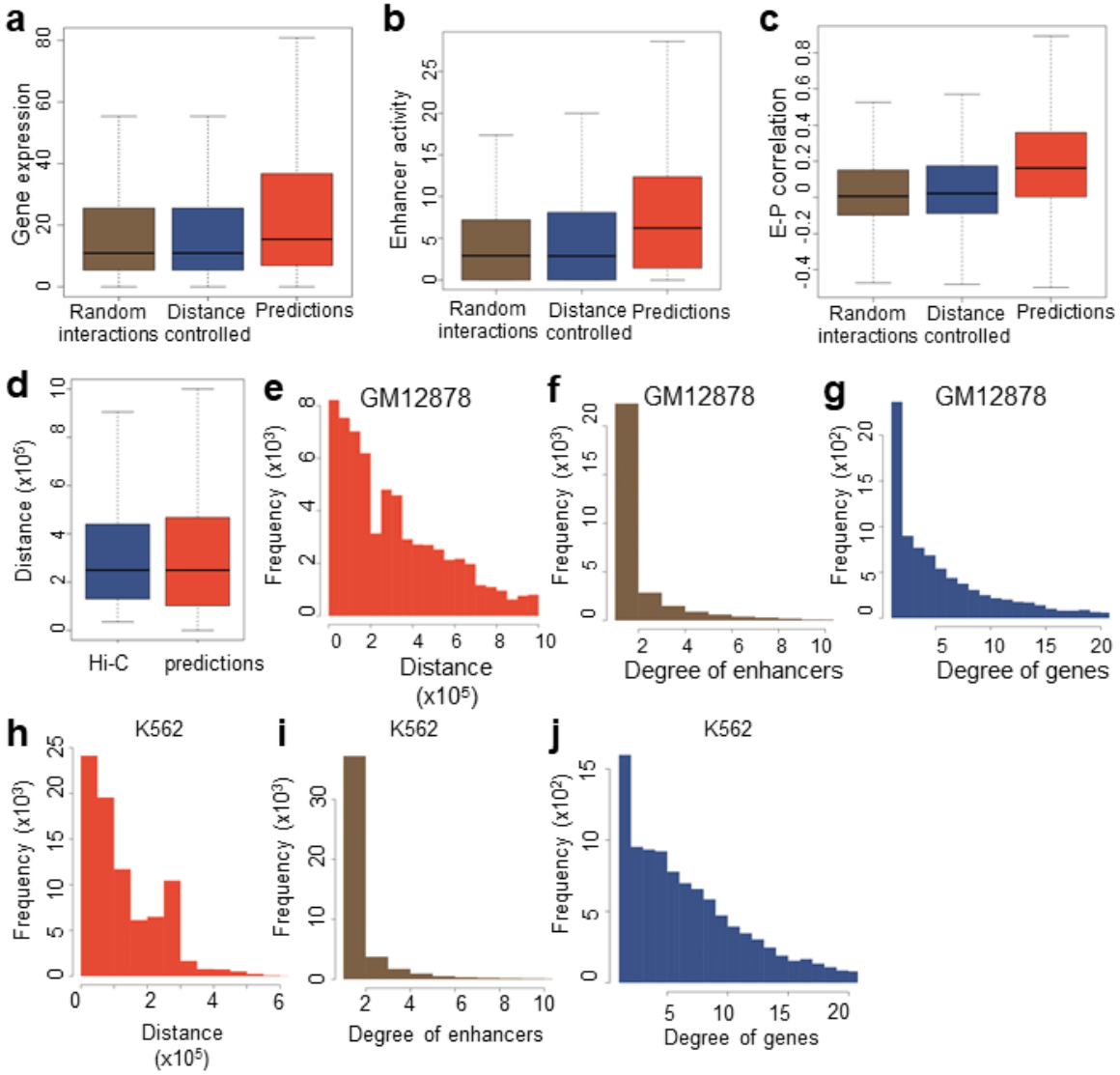



**Figure D.13 Summary of genome-wide predictions by ProTECT in GM12878 and K562.**

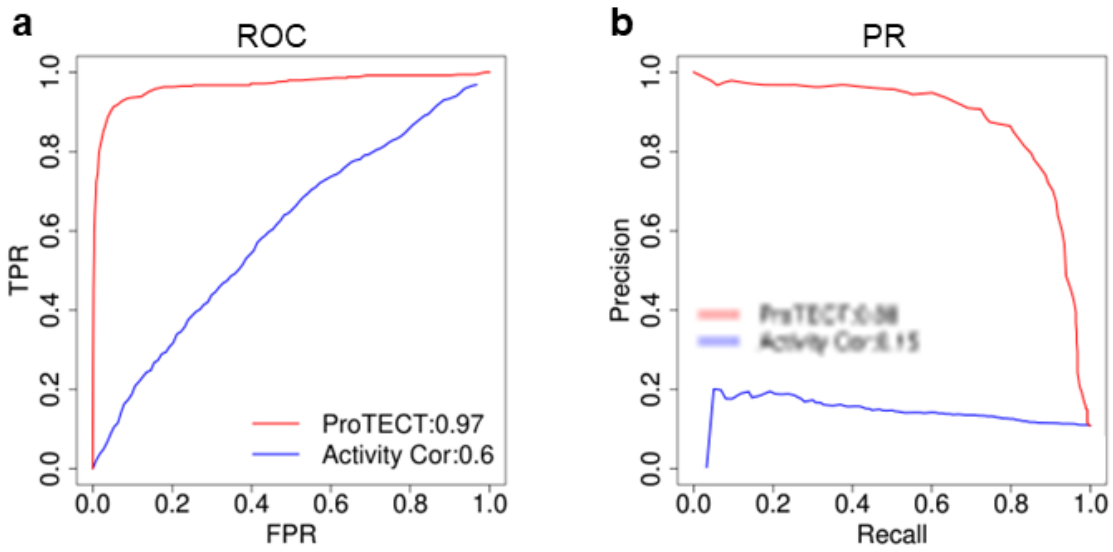



**Figure D.14 Validation of ProTECT predicted enhancer-gene links with enhancer degree greater than one.**

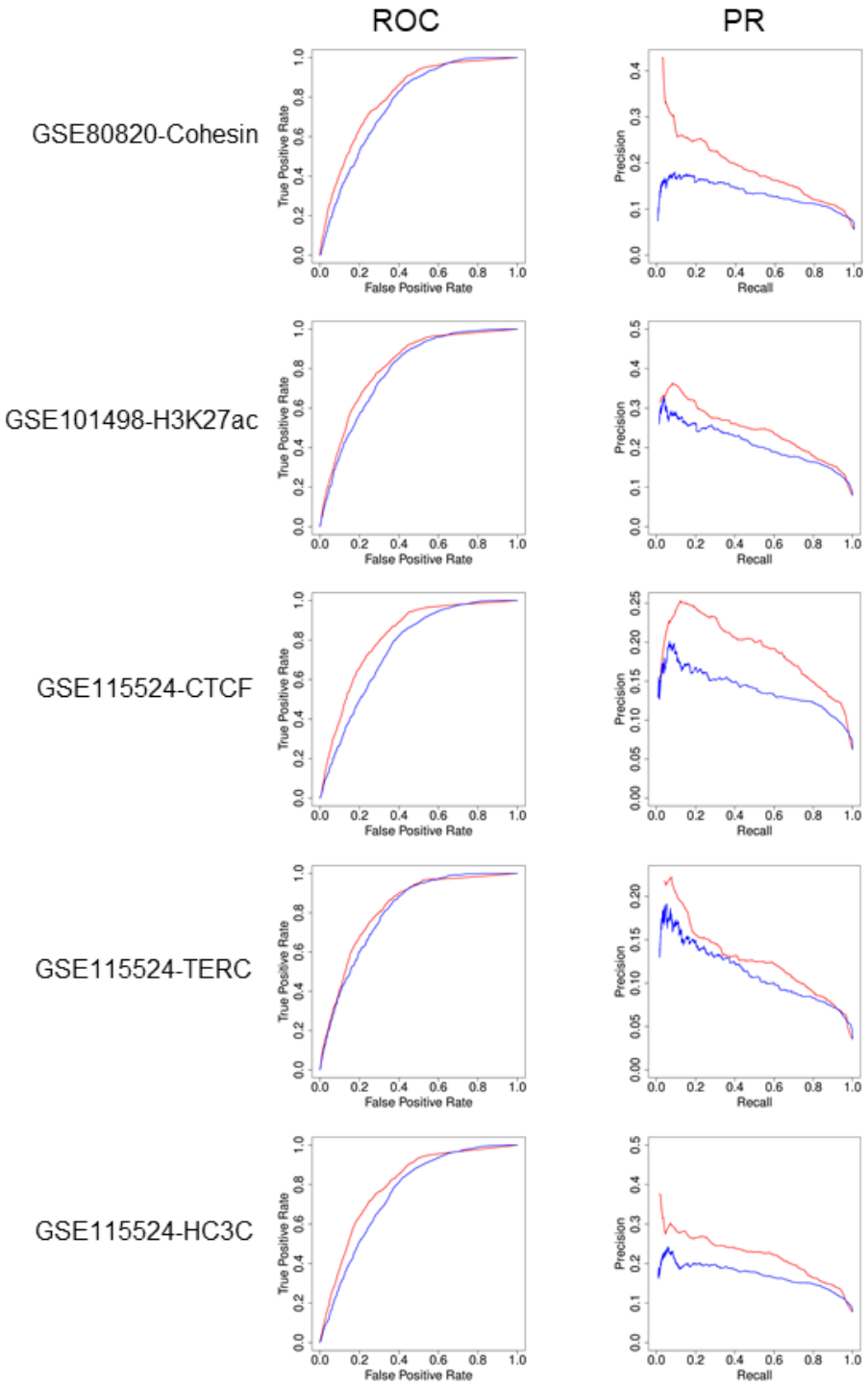



**Figure D.15 Performance comparison with the ABC model in the whole genome-wide.**

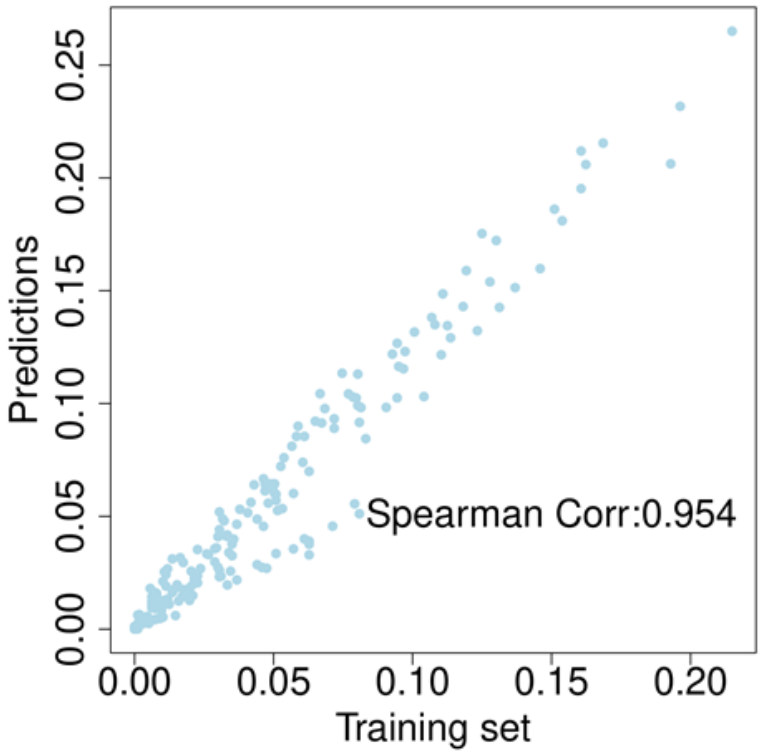



**Figure D.16 Comparing the TF PPI abundance score in the Hi-C supported enhancer-gene links and the ProTECT predictions.**

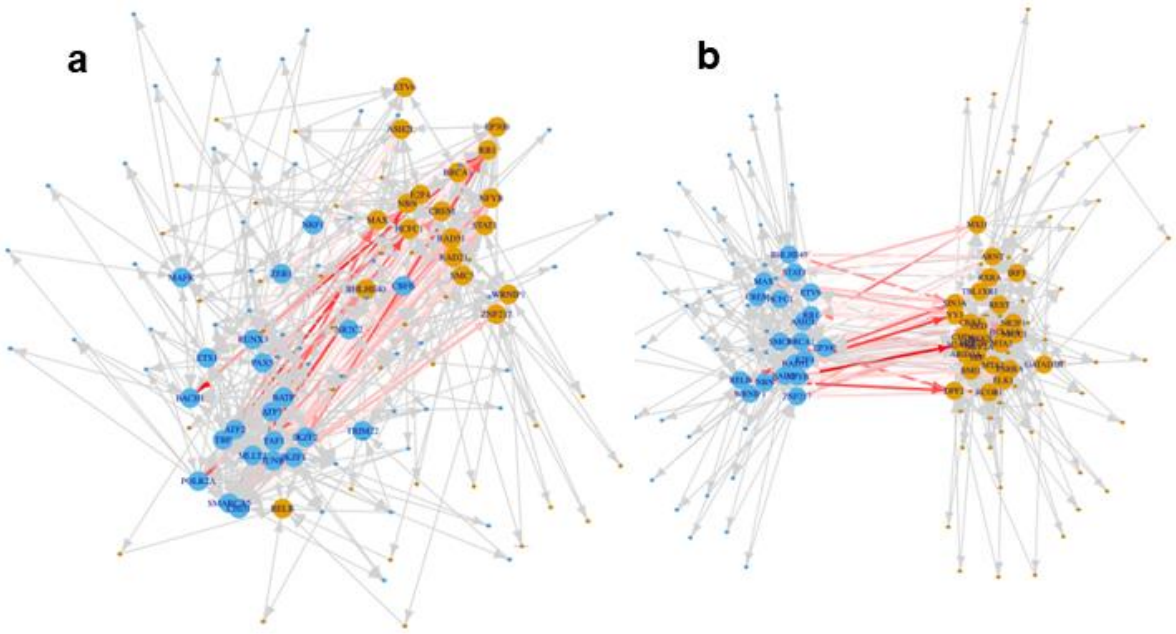



**Figure D.17 Examples of prioritized module-level TF PPIs features.**

**a** Un-directional TF PPI features


| | TF-TF abundance score | Fold |
|---|---|---|
| RELB-YY1 | 0.238523 | 1.23 |
| YY1-RELB | 0.194012 | |
| IKZF-CREM | 0.110332 | 1.17 |
| CREM-IKZF | 0.0941 | |

**b** Directional TF PPI features


| | TF-TF abundance score | Fold |
|---|---|---|
| EP300-POL2R2A | 0.193 | 9.19 |
| POL2R2A-EP300 | 0.021 | |
| SMC3-MXI1 | 0.234 | 7.8 |
| MXI1-SMC3 | 0.030 | |

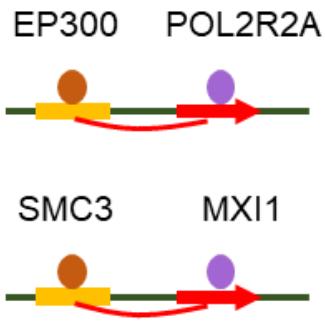



**Figure D.18 Identification of the directions of TF PPI features.**

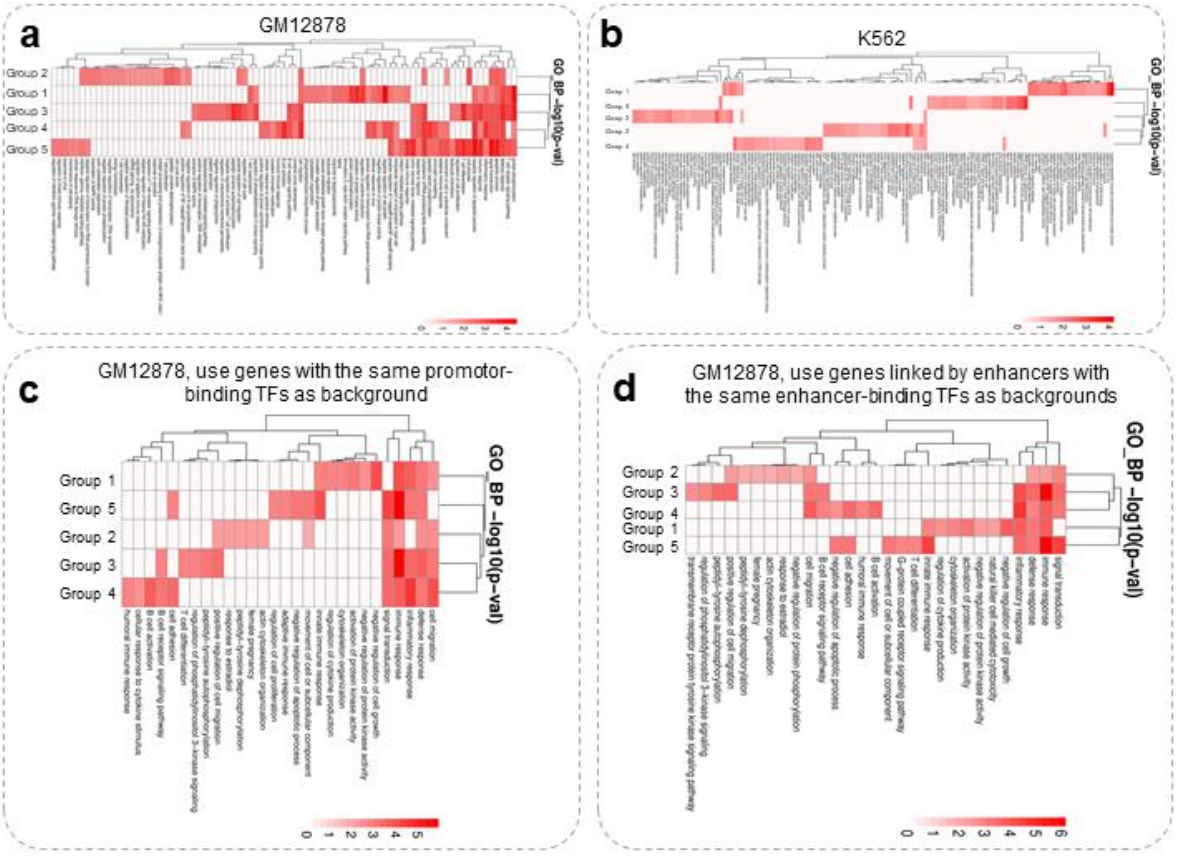



**Figure D.19 Differential pathway enrichments of genes regulated by different module-level TF PPIs based on the ProTECT predictions.**

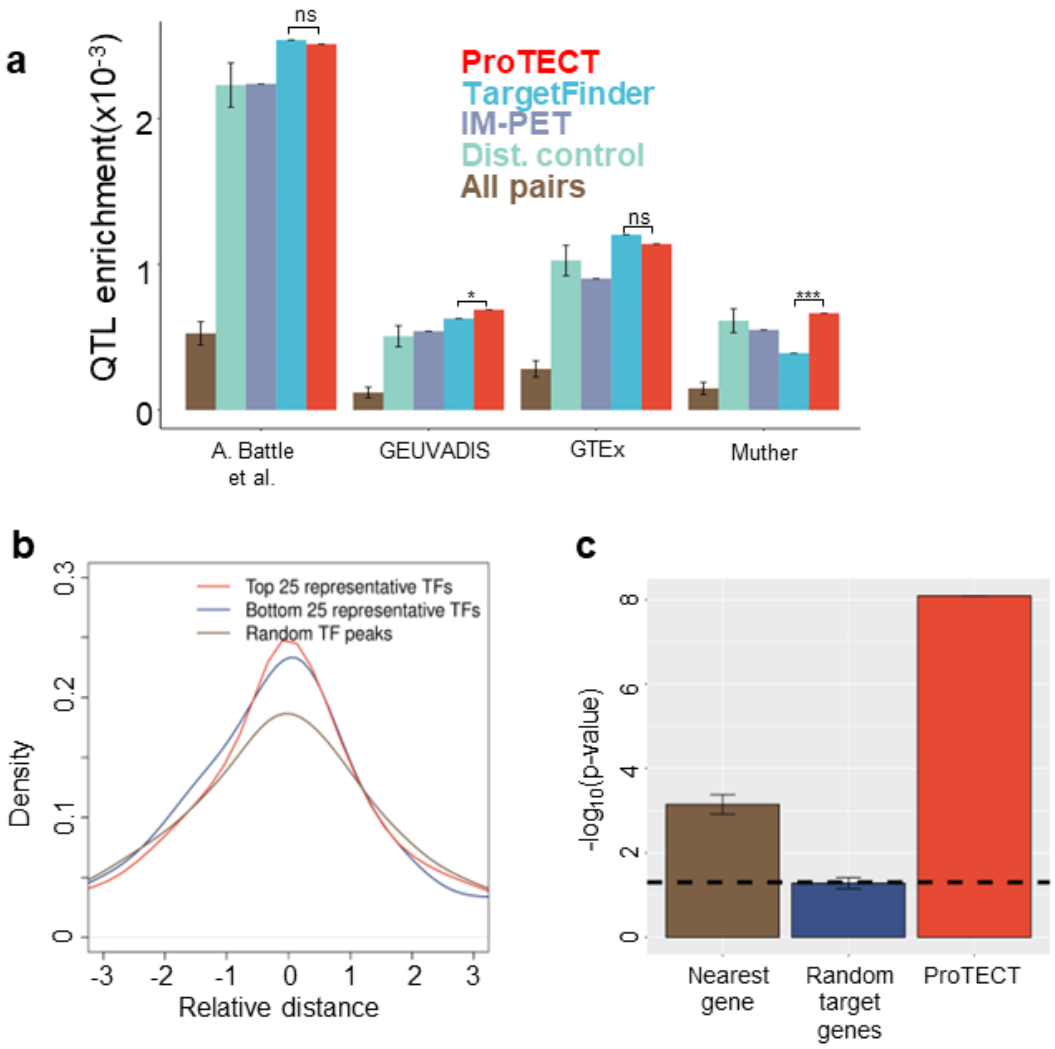



**Figure D.20 QTL enrichment analysis in K562.**

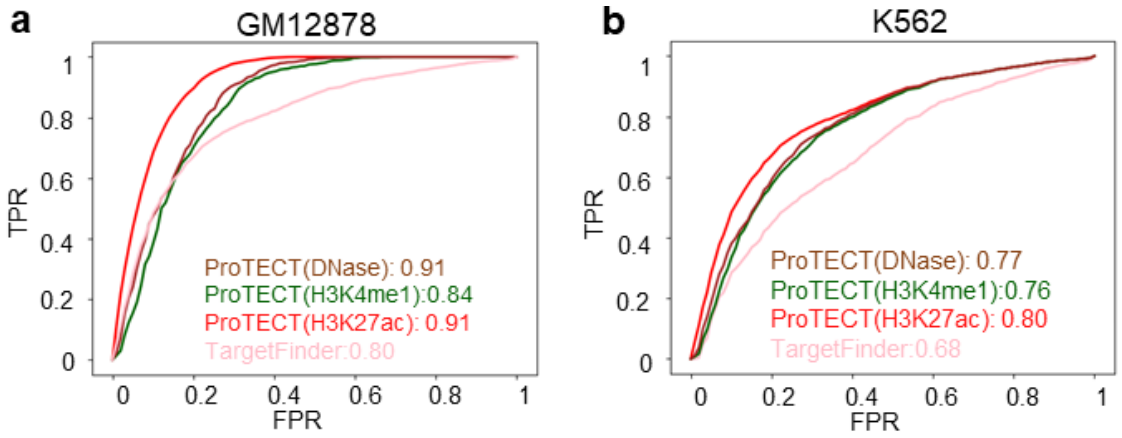



**Figure D.21 ProTECT predicts enhancer-gene links based on the imputed TF binding sites.**

# BIBLIOGRAPHY

## BIBLIOGRAPHY


1. Bickmore, W.A. The spatial organization of the human genome. Annu Rev Genomics Hum Genet **14**, 67-84 (2013).

2. Cremer, T. & Cremer, C. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. Nat Rev Genet **2**, 292-301 (2001).

3. Sexton, T., Schober, H., Fraser, P. & Gasser, S.M. Gene regulation through nuclear organization. Nat Struct Mol Biol **14**, 1049-1055 (2007).

4. Liu, M. et al. Multiplexed imaging of nucleome architectures in single cells of mammalian tissue. Nat Commun **11**, 2907 (2020).

5. Gorkin, D.U., Leung, D. & Ren, B. The 3D genome in transcriptional regulation and pluripotency. Cell Stem Cell **14**, 762-775 (2014).

6. Zhou, X. et al. The Human Epigenome Browser at Washington University. Nat Methods **8**, 989-990 (2011).

7. Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science **326**, 289-293 (2009).

8. Zheng, Y. & Keles, S. FreeHi-C simulates high-fidelity Hi-C data for benchmarking and data augmentation. Nat Methods **17**, 37-40 (2020).

9. Dixon, J.R. et al. Chromatin architecture reorganization during stem cell differentiation. Nature **518**, 331-336 (2015).

10. Rao, S.S. et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell **159**, 1665-1680 (2014).

11. Wang, Y. et al. The 3D Genome Browser: a web-based browser for visualizing 3D genome organization and long-range chromatin interactions. Genome Biol **19**, 151 (2018).

12. Schmitt, A.D. et al. A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome. Cell Rep **17**, 2042-2059 (2016).

13. Duan, Z. et al. A three-dimensional model of the yeast genome. Nature **465**, 363-367 (2010).

14. Wang, S. et al. Spatial organization of chromatin domains and compartments in single chromosomes. Science **353**, 598-602 (2016).

15. Fudenberg, G. et al. Formation of Chromosomal Domains by Loop Extrusion. Cell Rep **15**, 2038-2049 (2016).

16. Sanborn, A.L. et al. Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. Proc Natl Acad Sci U S A **112**, E6456-6465 (2015).

17. Oluwadare, O., Highsmith, M. & Cheng, J. An Overview of Methods for Reconstructing 3-D Chromosome and Genome Structures from Hi-C Data. Biol Proced Online **21**, 7 (2019).

18. Rieber, L. & Mahony, S. miniMDS: 3D structural inference from high-resolution Hi-C data. Bioinformatics **33**, i261-i266 (2017).

19. Szalaj, P. et al. An integrated 3-Dimensional Genome Modeling Engine for data-driven simulation of spatial genome organization. Genome Res **26**, 1697-1709 (2016).

20. Paulsen, J. et al. Chrom3D: three-dimensional genome modeling from Hi-C and nuclear lamin-genome contacts. Genome Biol **18**, 21 (2017).

21. Mishra, B., Meyer, G. & Sepulchre, R. in 2011 50th IEEE Conference on Decision and Control and European Control Conference 4455-4460 (2011).

22. Zhang, Z., Li, G., Toh, K.C. & Sung, W.K. 3D chromosome modeling with semi-definite programming and Hi-C data. J Comput Biol **20**, 831-846 (2013).

23. Adhikari, B., Trieu, T. & Cheng, J.L. Chromosome3D: reconstructing three-dimensional chromosomal structures from Hi-C interaction frequency data using distance geometry simulated annealing. Bmc Genomics **17** (2016).

24. Trieu, T. & Cheng, J.L. MOGEN: a tool for reconstructing 3D models of genomes from chromosomal conformation capturing data. Bioinformatics **32**, 1286-1292 (2016).

25. Wang, S., Xu, J. & Zeng, J. Inferential modeling of 3D chromatin structure. Nucleic Acids Res **43**, e54 (2015).

26. Peng, C. et al. The sequencing bias relaxed characteristics of Hi-C derived data and implications for chromatin 3D modeling. Nucleic Acids Res **41**, e183 (2013).

27. Kapilevich, V., Seno, S., Matsuda, H. & Takenaka, Y. Chromatin 3D Reconstruction from Chromosomal Contacts Using a Genetic Algorithm. IEEE/ACM Trans Comput Biol Bioinform **16**, 1620-1626 (2019).

28. Varoquaux, N., Ay, F., Noble, W.S. & Vert, J.P. A statistical approach for inferring the 3D structure of the genome. Bioinformatics **30**, i26-33 (2014).

29. Hu, M. et al. Bayesian inference of spatial organizations of chromosomes. PLoS Comput Biol **9**, e1002893 (2013).

30. Zou, C., Zhang, Y. & Ouyang, Z. HSA: integrating multi-track Hi-C data for genome-scale reconstruction of 3D chromatin structure. Genome Biol **17**, 40 (2016).

31. Rousseau, M., Fraser, J., Ferraiuolo, M.A., Dostie, J. & Blanchette, M. Three-dimensional modeling of chromatin structure from interaction frequency data using Markov chain Monte Carlo sampling. Bmc Bioinformatics **12** (2011).

32. Carstens, S., Nilges, M. & Habeck, M. Inferential Structure Determination of Chromosomes from Single-Cell Hi-C Data. PLoS Comput Biol **12**, e1005292 (2016).

33. Lesne, A., Riposo, J., Roger, P., Cournac, A. & Mozziconacci, J. 3D genome reconstruction from chromosomal contacts. Nat Methods **11**, 1141-1143 (2014).

34. Abbas, A. et al. Integrating Hi-C and FISH data for modeling of the 3D organization of chromosomes. Nat Commun **10**, 2049 (2019).

35. Trieu, T., Oluwadare, O. & Cheng, J. Hierarchical Reconstruction of High-Resolution 3D Models of Large Chromosomes. Sci Rep **9**, 4971 (2019).

36. Hirata, Y., Oda, A., Ohta, K. & Aihara, K. Three-dimensional reconstruction of single-cell chromosome structure using recurrence plots. Sci Rep-Uk **6** (2016).

37. Zhang, Y.L., Liu, W.W., Lin, Y., Ng, Y.K. & Li, S.C. Large-scale 3D chromatin reconstruction from chromosomal contacts. Bmc Genomics **20** (2019).

38. Li, F.Z. et al. Chromatin 3D structure reconstruction with consideration of adjacency relationship among genomic loci. Bmc Bioinformatics **21** (2020).

39. DeVience, S.J. & Mayer, D. Speeding up dynamic spiral chemical shift imaging with incoherent sampling and low-rank matrix completion. Magn Reson Med **77**, 951-960 (2017).

40. Shin, P.J. et al. Calibrationless parallel imaging reconstruction based on structured low-rank matrix completion. Magn Reson Med **72**, 959-970 (2014).

41. Kim, J.H., Sim, J.Y. & Kim, C.S. Video deraining and desnowing using temporal correlation and low-rank matrix completion. IEEE Trans Image Process **24**, 2658-2670 (2015).

42. Gower, J.C. Properties of Euclidean and non-Euclidean distance matrices. Linear Algebra and its Applications **67**, 81-97 (1985).

43. Fullwood, M.J. et al. An oestrogen-receptor-alpha-bound human chromatin interactome. Nature **462**, 58-64 (2009).

44. Hughes, J.R. et al. Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. Nat Genet **46**, 205-212 (2014).

45. Jung, I. et al. A compendium of promoter-centered long-range chromatin interactions in the human genome. Nat Genet **51**, 1442-1449 (2019).

46. Quinodoz, S.A. et al. Higher-Order Inter-chromosomal Hubs Shape 3D Genome Organization in the Nucleus. Cell **174**, 744-757 e724 (2018).

47. Baldi, S., Korber, P. & Becker, P.B. Beads on a string-nucleosome array arrangements and folding of the chromatin fiber. Nat Struct Mol Biol **27**, 109-118 (2020).

48. Eckstein, J. & Bertsekas, D.P. On the Douglas—Rachford splitting method and the proximal point algorithm for maximal monotone operators. Mathematical Programming **55**, 293-318 (1992).

49. Tasissa, A. & Lai, R. Exact Reconstruction of Euclidean Distance Geometry Problem Using Low-Rank Matrix Completion. IEEE Transactions on Information Theory **65**, 3124-3144 (2019).

50. Consortium, E.P. An integrated encyclopedia of DNA elements in the human genome. Nature **489**, 57-74 (2012).

51. Dekker, J. et al. The 4D nucleome project. Nature **549**, 219-226 (2017).

52. Zheng, H. & Xie, W. The role of 3D genome organization in development and cell differentiation. Nat Rev Mol Cell Biol **20**, 535-550 (2019).

53. Zheng, M. et al. Multiplex chromatin interactions with single-molecule precision. Nature **566**, 558-562 (2019).

54. Beagrie, R.A. et al. Complex multi-enhancer contacts captured by genome architecture mapping. Nature **543**, 519-524 (2017).

55. Hsieh, T.S. et al. Resolving the 3D Landscape of Transcription-Linked Mammalian Chromatin Folding. Mol Cell **78**, 539-553 e538 (2020).

56. Delaneau, O. et al. Chromatin three-dimensional interactions mediate genetic effects on gene expression. Science **364** (2019).

57. Koch, L. Adding another dimension to gene regulation. Nature Reviews Genetics **16**, 563-563 (2015).

58. Battle, A. et al. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. Genome Res **24**, 14-24 (2014).

59. Grundberg, E. et al. Mapping cis- and trans-regulatory effects across multiple tissues in twins. Nat Genet **44**, 1084-1089 (2012).

60. Lappalainen, T. et al. Transcriptome and genome sequencing uncovers functional variation in humans. Nature **501**, 506-511 (2013).

61. Consortium, G.T. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science **348**, 648-660 (2015).

62. Grubert, F. et al. Genetic Control of Chromatin States in Humans Involves Local and Distal Chromosomal Interactions. Cell **162**, 1051-1065 (2015).

63. Dekker, J. GC- and AT-rich chromatin domains differ in conformation and histone modification status and are differentially modulated by Rpd3p. Genome Biol **8**, R116 (2007).

64. Jabbari, K., Chakraborty, M. & Wiehe, T. DNA sequence-dependent chromatin architecture and nuclear hubs formation. Sci Rep **9**, 14646 (2019).

65. Sekelja, M., Paulsen, J. & Collas, P. 4D nucleomes in single cells: what can computational modeling reveal about spatial chromatin conformation? Genome Biol **17**, 54 (2016).

66. Nagano, T. et al. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. Nature **502**, 59-64 (2013).

67. Su, J.H., Zheng, P., Kinrot, S.S., Bintu, B. & Zhuang, X. Genome-Scale Imaging of the 3D Organization and Transcriptional Activity of Chromatin. Cell **182**, 1641-1659 e1626 (2020).

68. Bintu, B. et al. Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells. Science **362** (2018).

69. Tjong, H. et al. Population-based 3D genome structure analysis reveals driving forces in spatial genome organization. Proc Natl Acad Sci U S A **113**, E1663-1672 (2016).

70. Dai, C. et al. Mining 3D genome structure populations identifies major factors governing the stability of regulatory communities. Nat Commun **7**, 11549 (2016).

71. Yardimci, G.G. et al. Measuring the reproducibility and quality of Hi-C data. Genome Biol **20**, 57 (2019).

72. Li, G. et al. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. Cell **148**, 84-98 (2012).

73. Zhang, S., Chasman, D., Knaack, S. & Roy, S. In silico prediction of high-resolution Hi-C interaction matrices. Nat Commun **10**, 5449 (2019).

74. Zhang, Y. et al. Enhancing Hi-C data resolution with deep convolutional neural network HiCPlus. Nat Commun **9**, 750 (2018).

75. Fudenberg, G., Kelley, D.R. & Pollard, K.S. Predicting 3D genome folding from DNA sequence with Akita. Nat Methods **17**, 1111-1117 (2020).

76. Schwessinger, R. et al. DeepC: predicting 3D genome folding using megabase-scale transfer learning. Nat Methods **17**, 1118-1124 (2020).

77. Giorgetti, L. et al. Predictive polymer modeling reveals coupled fluctuations in chromosome conformation and transcription. Cell **157**, 950-963 (2014).

78. Qi, Y. & Zhang, B. Predicting three-dimensional genome organization with chromatin states. PLoS Comput Biol **15**, e1007024 (2019).

79. Brackley, C.A. et al. Predicting the three-dimensional folding of cis-regulatory regions in mammalian genomes using bioinformatic data and polymer models. Genome Biol **17**, 59 (2016).

80. Qi, Y. et al. Data-Driven Polymer Model for Mechanistic Exploration of Diploid Genome Organization. Biophys J **119**, 1905-1916 (2020).

81. Meluzzi, D. & Arya, G. Computational approaches for inferring 3D conformations of chromatin from chromosome conformation capture data. Methods **181-182**, 24-34 (2020).

82. Lin, X., Qi, Y., Latham, A.P. & Zhang, B. Multiscale modeling of genome organization with maximum entropy optimization. J Chem Phys **155**, 010901 (2021).

83. Moller, J. & de Pablo, J.J. Bottom-Up Meets Top-Down: The Crossroads of Multiscale Chromatin Modeling. Biophys J **118**, 2057-2065 (2020).

84. Di Stefano, M., Paulsen, J., Jost, D. & Marti-Renom, M.A. 4D nucleome modeling. Curr Opin Genet Dev **67**, 25-32 (2021).

85. Consortium, E.P. et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature **583**, 699-710 (2020).

86. Roadmap Epigenomics, C. et al. Integrative analysis of 111 reference human epigenomes. Nature **518**, 317-330 (2015).

87. Perry, M.W., Boettiger, A.N. & Levine, M. Multiple enhancers ensure precision of gap gene-expression patterns in the Drosophila embryo. Proc Natl Acad Sci U S A **108**, 13570-13575 (2011).

88. Tsai, A., Alves, M.R. & Crocker, J. Multi-enhancer transcriptional hubs confer phenotypic robustness. Elife **8** (2019).

89. Choi, J. et al. Evidence for additive and synergistic action of mammalian enhancers during cell fate determination. Elife **10** (2021).

90. Nord, A.S. et al. Rapid and pervasive changes in genome-wide enhancer usage during mammalian development. Cell **155**, 1521-1531 (2013).

91. Schoenfelder, S. & Fraser, P. Long-range enhancer-promoter contacts in gene expression control. Nat Rev Genet **20**, 437-455 (2019).

92. Vicente, C.T. et al. Long-Range Modulation of PAG1 Expression by 8q21 Allergy Risk Variants. Am J Hum Genet **97**, 329-336 (2015).

93. Martin, P. et al. Capture Hi-C reveals novel candidate genes and complex long-range interactions with related autoimmune risk loci. Nat Commun **6**, 10069 (2015).

94. Deng, W. et al. Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. Cell **149**, 1233-1244 (2012).

95. Ragoczy, T., Bender, M.A., Telling, A., Byron, R. & Groudine, M. The locus control region is required for association of the murine beta-globin locus with engaged transcription factories during erythroid maturation. Genes Dev **20**, 1447-1457 (2006).

96. Lettice, L.A. et al. A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. Hum Mol Genet **12**, 1725-1735 (2003).

97. Jeong, Y., El-Jaick, K., Roessler, E., Muenke, M. & Epstein, D.J. A functional screen for sonic hedgehog regulatory elements across a 1 Mb interval identifies long-range ventral forebrain enhancers. Development **133**, 761-772 (2006).

98. Sagai, T. et al. A cluster of three long-range enhancers directs regional Shh expression in the epithelial linings. Development **136**, 1665-1674 (2009).

99. Smemo, S. et al. Obesity-associated variants within FTO form long-range functional connections with IRX3. Nature **507**, 371-375 (2014).

100. Dryden, N.H. et al. Unbiased analysis of potential targets of breast cancer susceptibility loci by Capture Hi-C. Genome Res **24**, 1854-1868 (2014).

101. McGovern, A. et al. Capture Hi-C identifies a novel causal gene, IL20RA, in the pan-autoimmune genetic susceptibility region 6q23. Genome Biol **17**, 212 (2016).

102. Jager, R. et al. Capture Hi-C identifies the chromatin interactome of colorectal cancer risk loci. Nat Commun **6**, 6178 (2015).

103. Shlyueva, D., Stampfel, G. & Stark, A. Transcriptional enhancers: from properties to genome-wide predictions. Nat Rev Genet **15**, 272-286 (2014).

104. Buecker, C. & Wysocka, J. Enhancers as information integration hubs in development: lessons from genomics. Trends Genet **28**, 276-284 (2012).

105. Hoffman, M.M. et al. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. Nat Methods **9**, 473-476 (2012).

106. Ernst, J. & Kellis, M. Chromatin-state discovery and genome annotation with ChromHMM. Nat Protoc **12**, 2478-2492 (2017).

107. Pennacchio, L.A., Bickmore, W., Dean, A., Nobrega, M.A. & Bejerano, G. Enhancers: five essential questions. Nat Rev Genet **14**, 288-295 (2013).

108. Mumbach, M.R. et al. Enhancer connectome in primary human cells identifies target genes of disease-associated DNA elements. Nat Genet **49**, 1602-1612 (2017).

109. Gondor, A. & Ohlsson, R. Chromosome crosstalk in three dimensions. Nature **461**, 212-217 (2009).

110. Kvon, E.Z. et al. Progressive Loss of Function in a Limb Enhancer during Snake Evolution. Cell **167**, 633-642 e611 (2016).

111. Claussnitzer, M. et al. FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. N Engl J Med **373**, 895-907 (2015).

112. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. Science **295**, 1306-1311 (2002).

113. Zhao, Z. et al. Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. Nat Genet **38**, 1341-1347 (2006).

114. Dostie, J. et al. Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. Genome Res **16**, 1299-1309 (2006).

115. Mifsud, B. et al. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. Nat Genet **47**, 598-606 (2015).

116. Schoenfelder, S., Javierre, B.M., Furlan-Magaril, M., Wingett, S.W. & Fraser, P. Promoter Capture Hi-C: High-resolution, Genome-wide Profiling of Promoter Interactions. J Vis Exp (2018).

117. Fullwood, M.J. & Ruan, Y. ChIP-based methods for the identification of long-range chromatin interactions. J Cell Biochem **107**, 30-39 (2009).

118. Li, X. et al. Long-read ChIA-PET for base-pair-resolution mapping of haplotype-specific chromatin interactions. Nat Protoc **12**, 899-915 (2017).

119. Smith, E.M., Lajoie, B.R., Jain, G. & Dekker, J. Invariant TAD Boundaries Constrain Cell-Type-Specific Looping Interactions between Promoters and Distal Elements around the CFTR Locus. Am J Hum Genet **98**, 185-201 (2016).

120. Li, G. et al. ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing. Genome Biol **11**, R22 (2010).

121. Meuleman, W. et al. Index and biological spectrum of human DNase I hypersensitive sites. Nature **584**, 244-251 (2020).

122. Yen, A. & Kellis, M. Systematic chromatin state comparison of epigenomes associated with diverse properties including sex and tissue type. Nat Commun **6**, 7973 (2015).

123. Roy, S. et al. A predictive modeling approach for cell line-specific long-range regulatory interactions. Nucleic Acids Res **43**, 8694-8712 (2015).

124. Hait, T.A., Amar, D., Shamir, R. & Elkon, R. FOCS: a novel method for analyzing enhancer and gene activity patterns infers an extensive enhancer-promoter map. Genome Biol **19**, 56 (2018).

125. Gao, T. & Qian, J. EAGLE: An algorithm that utilizes a small number of genomic features to predict tissue/cell type-specific enhancer-gene interactions. PLoS Comput Biol **15**, e1007436 (2019).

126. Cao, Q. et al. Reconstruction of enhancer-target networks in 935 samples of human primary cells, tissues and cell lines. Nat Genet **49**, 1428-1436 (2017).

127. He, B., Chen, C., Teng, L. & Tan, K. Global view of enhancer-promoter interactome in human cells. Proc Natl Acad Sci U S A **111**, E2191-2199 (2014).

128. Whalen, S., Truty, R.M. & Pollard, K.S. Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin. Nat Genet **48**, 488-496 (2016).

129. Fishilevich, S. et al. GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. Database (Oxford) **2017** (2017).

130. Thurman, R.E. et al. The accessible chromatin landscape of the human genome. Nature **489**, 75-82 (2012).

131. Corradin, O. et al. Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. Genome Res **24**, 1-13 (2014).

132. Moore, J.E., Pratt, H.E., Purcaro, M.J. & Weng, Z. A curated benchmark of enhancer-gene interactions for evaluating enhancer-target gene prediction methods. Genome Biol **21**, 17 (2020).

133. Cao, F. & Fullwood, M.J. Inflated performance measures in enhancer-promoter interaction-prediction methods. Nat Genet **51**, 1196-1198 (2019).

134. Whitaker, J.W., Nguyen, T.T., Zhu, Y., Wildberg, A. & Wang, W. Computational schemes for the prediction and annotation of enhancers from epigenomic assays. Methods **72**, 86-94 (2015).

135. Nolis, I.K. et al. Transcription factors mediate long-range enhancer-promoter interactions. Proc Natl Acad Sci U S A **106**, 20222-20227 (2009).

136. Hnisz, D., Shrinivas, K., Young, R.A., Chakraborty, A.K. & Sharp, P.A. A Phase Separation Model for Transcriptional Control. Cell **169**, 13-23 (2017).

137. Quevedo, M. et al. Mediator complex interaction partners organize the transcriptional network that defines neural stem cells. Nat Commun **10**, 2669 (2019).

138. Maksimenko, O. & Georgiev, P. Mechanisms and proteins involved in long-distance interactions. Front Genet **5**, 28 (2014).

139. Li, Y. et al. The structural basis for cohesin-CTCF-anchored loops. Nature **578**, 472-476 (2020).

140. Beagan, J.A. et al. YY1 and CTCF orchestrate a 3D chromatin looping switch during early neural lineage commitment. Genome Res **27**, 1139-1152 (2017).

141. Weintraub, A.S. et al. YY1 Is a Structural Regulator of Enhancer-Promoter Loops. Cell **171**, 1573-1588 e1528 (2017).

142. Morgan, S.L. et al. Manipulation of nuclear architecture through CRISPR-mediated chromosomal looping. Nat Commun **8**, 15993 (2017).

143. Zhang, K., Li, N., Ainsworth, R.I. & Wang, W. Systematic identification of protein combinations mediating chromatin looping. Nat Commun **7**, 12249 (2016).

144. Wang, R. et al. Hierarchical cooperation of transcription factors from integration analysis of DNA sequences, ChIP-Seq and ChIA-PET data. BMC Genomics **20**, 296 (2019).

145. Kato, M., Hata, N., Banerjee, N., Futcher, B. & Zhang, M.Q. Identifying combinatorial regulation of transcription factors and binding motifs. Genome Biol **5**, R56 (2004).

146. Michaelis, C., Ciosk, R. & Nasmyth, K. Cohesins: chromosomal proteins that prevent premature separation of sister chromatids. Cell **91**, 35-45 (1997).

147. Tan, K., Shlomi, T., Feizi, H., Ideker, T. & Sharan, R. Transcriptional regulation of protein complexes within and across species. Proc Natl Acad Sci U S A **104**, 1283-1288 (2007).

148. Szklarczyk, D. et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res **47**, D607-D613 (2019).

149. Harrow, J. et al. GENCODE: producing a reference annotation for ENCODE. Genome Biol **7 Suppl 1**, S4 1-9 (2006).

150. Zhang, Y. et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol **9**, R137 (2008).

151. Amoutzias, G.D., Robertson, D.L., Van de Peer, Y. & Oliver, S.G. Choose your partners: dimerization in eukaryotic transcription factors. Trends Biochem Sci **33**, 220-229 (2008).

152. Dixon, J.R. et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature **485**, 376-380 (2012).

153. Akdemir, K.C. et al. Disruption of chromatin folding domains by somatic genomic rearrangements in human cancer. Nat Genet **52**, 294-305 (2020).

154. Chesi, A. et al. Genome-scale Capture C promoter interactions implicate effector genes at GWAS loci for bone mineral density. Nat Commun **10**, 1260 (2019).

155. Pugacheva, E.M. et al. CTCF mediates chromatin looping via N-terminal domain-dependent cohesin retention. Proc Natl Acad Sci U S A **117**, 2020-2031 (2020).

156. Vishwanathan, S.V.N., Borgwardt, K.M., Risi Kondor, I. & Schraudolph, N.N. arXiv:0807.0093 (2008).

157. Pons, P. & Latapy, M. physics/0512106 (2005).

158. Newman, M.E. Modularity and community structure in networks. Proc Natl Acad Sci U S A **103**, 8577-8582 (2006).

159. Hauenstein, S., Dormann, C.F. & Wood, S.N. arXiv:1603.02743 (2016).

160. Storey, J.D. A Direct Approach to False Discovery Rates. Journal of the Royal Statistical Society. Series B (Statistical Methodology) **64**, 479-498 (2002).

161. Huang da, W., Sherman, B.T. & Lempicki, R.A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc **4**, 44-57 (2009).

162. Consortium, G.T. et al. Genetic effects on gene expression across human tissues. Nature **550**, 204-213 (2017).

163. Gong, J. et al. PancanQTL: systematic identification of cis-eQTLs and trans-eQTLs in 33 cancer types. Nucleic Acids Res **46**, D971-D976 (2018).

164. Mumbach, M.R. et al. HiChIRP reveals RNA-associated chromosome conformation. Nat Methods **16**, 489-492 (2019).

165. Mumbach, M.R. et al. HiChIP: efficient and sensitive analysis of protein-directed genome architecture. Nat Methods **13**, 919-922 (2016).

166. Fulco, C.P. et al. Activity-by-contact model of enhancer-promoter regulation from thousands of CRISPR perturbations. Nat Genet **51**, 1664-1669 (2019).

167. Jiang, Y. et al. Genome-wide analyses of chromatin interactions after the loss of Pol I, Pol II, and Pol III. Genome Biol **21**, 158 (2020).

168. Dyson, N.J. RB1: a prototype tumor suppressor and an enigma. Genes Dev **30**, 1492-1502 (2016).

169. Marke, R., van Leeuwen, F.N. & Scheijen, B. The many faces of IKZF1 in B-cell precursor acute lymphoblastic leukemia. Haematologica **103**, 565-574 (2018).

170. Sarvagalla, S., Kolapalli, S.P. & Vallabhapurapu, S. The Two Sides of YY1 in Cancer: A Friend and a Foe. Front Oncol **9**, 1230 (2019).

171. Stengel, K.R. & Hiebert, S.W. Class I HDACs Affect DNA Replication, Repair, and Chromatin Structure: Implications for Cancer Therapy. Antioxid Redox Signal **23**, 51-65 (2015).

172. Losada, A., Hirano, M. & Hirano, T. Identification of Xenopus SMC protein complexes required for sister chromatid cohesion. Genes Dev **12**, 1986-1997 (1998).

173. Lee, T.C. & Ziff, E.B. Mxi1 is a repressor of the c-Myc promoter and reverses activation by USF. J Biol Chem **274**, 595-606 (1999).

174. Visel, A. et al. ChIP-seq accurately predicts tissue-specific activity of enhancers. Nature **457**, 854-858 (2009).

175. Lynch, C.J. et al. The RNA Polymerase II Factor RPAP1 Is Critical for Mediator-Driven Transcription and Cell Identity. Cell Rep **22**, 396-410 (2018).

176. Hu, Z., Killion, P.J. & Iyer, V.R. Genetic reconstruction of a functional transcriptional regulatory network. Nat Genet **39**, 683-687 (2007).

177. Albert, F.W., Bloom, J.S., Siegel, J., Day, L. & Kruglyak, L. Genetics of trans-regulatory variation in gene expression. Elife **7** (2018).

178. Brynedal, B. et al. Large-Scale trans-eQTLs Affect Hundreds of Transcripts and Mediate Patterns of Transcriptional Co-regulation. Am J Hum Genet **100**, 581-591 (2017).

179. Johanson, T.M. et al. Transcription-factor-mediated supervision of global genome architecture maintains B cell identity. Nat Immunol **19**, 1257-1264 (2018).

180. Ebert, A. et al. The distal V(H) gene cluster of the Igh locus contains distinct regulatory elements with Pax5 transcription factor-dependent activity in pro-B cells. Immunity **34**, 175-187 (2011).

181. Arvey, A. et al. An atlas of the Epstein-Barr virus transcriptome and epigenome reveals host-virus regulatory interactions. Cell Host Microbe **12**, 233-245 (2012).

182. Bult, C.J. et al. Mouse Genome Database (MGD) 2019. Nucleic Acids Res **47**, D801-D806 (2019).

183. Li, H., Quang, D. & Guan, Y. Anchor: trans-cell type prediction of transcription factor binding sites. Genome Res **29**, 281-292 (2019).

184. Keilwagen, J., Posch, S. & Grau, J. Accurate prediction of cell type-specific transcription factor binding. Genome Biol **20**, 9 (2019).