COMPUTATIONAL INTERROGATION OF CELL-TYPE SPECIFIC IMPACTS FOR NON-CODING GENETIC VARIANTS BASED ON 3D GENOME ORGANIZATION

By

Jiaxin Yang

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

2024

ABSTRACT

Genetic variations play a crucial role in the development of severe human diseases, but deciphering their genetic basis remains challenging due to the complex multi-scale genome structure and intricate molecular mechanisms. Genome-wide Association Studies (GWAS) have identified genotype-phenotype associations, but face limitations such as low statistical power and lack of causality. With advancements in Next Generation Sequencing techniques, researchers have gained insights into the vast non-coding genome, discovering millions of functional DNA elements that regulate cell type-specific gene expression, including promoters and enhancers. These *cis*-regulatory elements form gene regulatory networks (GRNs) that provide pathways for understanding the effects of genetic variants on diseases. However, leveraging GRNs in non-coding genetic variant analysis poses challenges, such as the vast number of potential enhancer-gene links, the influence of multilevel variabilities on chromatin interactions and 3D structures, and the need for comprehensive data integration methods. This thesis aims to address these challenges by developing machine learning, deep learning, and optimization-based models to discover novel disease-associated genes, enhance eQTL fine-mapping predictions, and investigate the multi-level variabilities of multi-scale 3D chromatin organization. By leveraging regulatory networks of long-range chromatin interactions, incorporating 3D chromatin organizations, and modeling the 3D structures, this work contributes to deep understanding on cell type-specific non-coding genetic variations and advancing precision medicine and clinical care.

This thesis is dedicated to my grandpa
Thank you for the unconditional love across time

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my advisor and committee chair, Dr. Jianrong Wang. His expert guidance was instrumental in helping me navigate the various stages of my projects, from setting specific goals to developing computational algorithms. I am profoundly grateful for his mentorship.

My appreciation also extends to my committee members, Dr. Jianliang Qian, Dr. Yuehua Cui, and Dr. Rongrong Wang, for their invaluable feedback on my research and their guidance in developing my professional skills.

I am thankful for the support and camaraderie of my lab colleagues, Dr. Hao Wang, Dr. Wenjie Qi, Sikta Das Adhikari, Dr. Binbin Huang, Dr. Jiwoong Kim, and Dr. Pronoy Kanti Mondal. The days we spent working together will always be cherished.

I owe a debt of gratitude to my friends, Dr. Wei Jin, Lei Yang, Yiming Zhou, Yiqian Wang, and Dr. Zhongjie Ji, for their understanding, support, and companionship throughout my graduate studies. Their presence made the challenging moments more bearable.

Most importantly, I wish to thank my family for their unwavering support and encouragement. Their belief in my abilities has been a constant source of strength. I would like to give a special mention to my grandfather, an incredible source of love, guidance, and inspiration. I am forever grateful for the foundation you laid in my life, and I love you more than words can express. As a tiny planet in this vast universe. you are the unseen dark matter embracing me. How I long to behold you again, yet I know our paths cannot cross anew. Still, your gravity lingers, a silent force. I cherish the moment our light cones intertwined, forever altering my celestial course. Though we may never meet again, you are the reason my galaxy remains whole, an eternal strand in the cosmic web to which I am bound.

TABLE OF CONTENTS

CHAPTER 1
INTRODUCTION

Genetic variations play a crucial role in the development of various severe human diseases, including leukemia, Alzheimer's disease, and cancers[1]. Understanding the relationship between genetic variants and human diseases is essential for advancing precision medicine and clinical care. Genome-wide Association Studies (GWAS) have revolutionized complex disease genetics by identifying genotype-phenotype associations through the analysis of millions of single-nucleotide polymorphisms (SNPs)[2,3]. However, GWAS faces limitations such as low statistical power[4] and lack of causality due to Linkage Disequilibrium (LD)[5]. Deciphering the genetic basis of human diseases remains challenging, considering the multi-scale genome structure and intricate underlying molecular mechanisms[6,7].

Traditionally, research focused on SNPs located in the coding regions of genes, which only account for 1.2% of the human genome[2]. With the advancements in Next Generation Sequencing (NGS) techniques[8], such as RNA-seq, DNase-seq, and ChIP-seq[9], scientists have gained insights into the functions of the vast non-coding genome[10]. Recent studies have discovered millions of functional DNA elements that regulate cell type-specific gene expression in *cis*, with promoters and enhancers being the two major *cis*-regulatory elements (CREs)[11]. Enhancers, bound by various transcription factors (TFs), activate target gene expression by forming chromatin loops with promoters. Gene regulatory networks (GRNs), formed by enhancers, genes, and TFs, provide clear pathways for understanding the effects of genetic variants on diseases[6]. However, leveraging GRNs in the analysis of non-coding genetic variants poses several challenges, including the vast number of potential enhancer-gene links, the need for comprehensive data integration methods, the unclear relationship between non-coding genetic variant-gene associations and enhancer-promoter associations, and the influence of multilevel variabilities on chromatin interactions and 3D structures.

To address these challenges, this thesis aims to develop machine learning, deep learning, and optimization-based models to (1) discover novel disease-associated genes by leveraging regulatory networks of long-range chromatin interactions, (2) enhance eQTL fine-mapping predictions by incorporating 3D chromatin knowledge, and (3)

investigate the multi-level variabilities of 3D chromatin organization and model the 3D structures.

Accurate prediction of cell type-specific enhancer-promoter interactions (EPIs) is crucial for constructing GRNs[12]. Computational algorithms for predicting chromatin interactions can be classified into unsupervised and supervised models[12]. Unsupervised models, such as the Activity-by-contact (ABC) model[13] and correlation-based models, have demonstrated superior performance compared to distance-based approaches. However, their performance is sensitive to feature selection and correlation calculation algorithms. With the development of chromosome conformation capture (3C)-based techniques, such as Hi-C[14], HiChIP[15], ChIA-PET[16], and Capture Hi-C[17], supervised models have been designed to identify potential EPIs by incorporating various 1D genomic/epigenomic features[18,19]. Deep learning-based models, such as SPEID[20] and DeepTACT[21], have further expanded the predictive capabilities by extracting information from reference DNA sequences using representation learning techniques.

Understanding the spatial organization and folding of the human genome within the cell nucleus is crucial for deciphering the complex systems of spatially coordinated transcriptional and epigenetic activities[7]. Hi-C[14] has been a driving force in studying 3D genome structures, revealing structural components such as chromatin loops, topologically associated domains (TADs), and chromatin compartments. However, accurately reconstructing high-resolution spatial conformations for all chromosomes is computationally challenging due to the large missing rate and high noise level in Hi-C data. Two categories of computational models have been developed to address this problem: (1) Hi-C simulation and prediction[22–24] and (2) 3D genome structure reconstruction[25–28]. Simulation-based models rely on physical mechanism assumptions, such as polymer modeling and phase separation, while data-driven deep learning-based models, like DeepC[23] and Akita[24], predict Hi-C contact maps using DNA sequences. In 3D structure reconstruction, observed Hi-C contact frequencies are converted into spatial distances[29], and consensus or ensemble structures are inferred by maximizing the similarity between predicted and observed Hi-C distances. The emerging single-cell Hi-C (scHi-C)[30–40] technologies have enabled the mapping of 3D chromatin structures in individual cells, revealing the fundamental genome structure and function connections at

single-cell resolution. However, the extremely low sequencing depth of single-cell chromatin contact maps poses challenges for studying high-resolution 3D chromatin structures. Methods like Higashi[41] have been developed to impute contact maps based on latent correlations among single cells, but there is still a lack of comprehensive methods to reconstruct single-cell 3D chromatin structures.

Identifying functionally relevant variants is a significant challenge in human genetics, particularly for non-coding variants[10,42]. Large-scale efforts, such as the ENCODE consortium[9] and the US National Institutes of Health Roadmap Epigenomics project[11], have provided data from various assays across the genome to help interpret non-coding variants. Tools like GWAVA[43] and FunSeq2[44] have been developed to annotate potential regulatory variants and predict their effects genome-wide. However, simply considering the overlap of a variant with annotations is insufficient due to the low resolution of publicly available data and the high false-negative rate of rare variants. Traditional machine learning-based model like Kmer-SVM[45] and gkm-SVM[46] emerged as pioneering models for predicting regulatory elements directly from DNA sequences, bypassing the need for existing annotated motifs. These models employ support vector machines (SVMs) trained on k-mer features to assess the likelihood of a sequence being a functional genomic regulatory element or a tissue-specific enhancer. Delta-SVM[47] incorporates gkm-SVM predictions to assess the disruptive impacts of genetic variants. However, the complexity and non-linearity of the underlying regulatory grammar in DNA sequences require further improvements in model performance.

In recent years, advanced machine learning, deep learning, and optimization-based models have achieved extraordinary performance across various scientific fields[48–68], including the prediction of genetic variant effects from DNA sequences. DeepSEA[69], Basset[70], and DanQ[71], have demonstrated the potential of convolutional neural networks (CNNs) and recurrent neural networks (RNNs) for predicting genetic variant effects based on DNA sequences. These models are trained on large-scale multi-omics datasets across different cell types from the reference genome and can predict the effects of genetic variants on transcription factor binding, chromatin accessibility, and gene expression. Further advancements, such as Basenji[72] and ExPecto[73], have expanded the scope of predictions to include a wider range of genomic signals and target gene expressions.

The Transformer architecture[74], originally developed for natural language processing, has shown remarkable success in capturing long-range dependencies and has been adapted for genomic applications. Models like Enformer[75], which combines CNNs and Transformers, excel in predicting functional genome profiles and offer improved interpretability through attention weights. Enformer's ability to capture distal regulatory elements up to 100kb away enables more accurate predictions of genetic variant effects. Foundation Models, such as DNABERT[76–78] and the Nucleotide Transformer[79], leverage self-supervised pre-training on unlabeled genomic sequences to capture the fundamental grammatical structures of DNA. These models have demonstrated remarkable efficacy across various downstream applications, including the detection of functional genetic variants. The increased model size and ability to process longer sequences have further enhanced their performance.

Despite the advancements in deep learning models for predicting genetic variant effects, challenges remain. The reliance on labeled data at the cell type level limits their capability to discern functional effects at the single-cell level. Quantifying the impact of genetic variation at the single-cell level has also been studied. The development of single-cell RNA-sequencing (scRNA-seq)[80] has enabled the simultaneous and unbiased estimation of cellular composition and cell type-specific gene expression, creating opportunities for mapping eQTLs across different cell types and in dynamic processes[81]. The single-cell eQTLGen consortium (sc-eQTLGen)[81] has been established as a large-scale international collaborative effort to identify the upstream interactors and downstream consequences of disease-related genetic variants in individual immune cell types. The integration of single-cell sequencing data, such as scRNA-seq[80,82], scATAC-seq[82,83], and scHi-C[30–40], presents opportunities for fine-tuning models with minimal data. Additionally, the training of current models based on the reference genome neglects the diversity and frequency of genetic variations across different genotypes. CRISPR[84–87] technology, which elucidates the causal and real effects of genetic variants, offers valuable insights beyond the reference genomic context and can help bridge the gap between model predictions and biological reality.

In conclusion, this thesis aims to address these challenges by developing machine learning, deep learning, and optimization-based models to discover novel disease-

associated genes, enhance eQTL fine-mapping predictions, and investigate the multi-level variabilities of multi-scale 3D chromatin organization. By leveraging regulatory networks of long-range chromatin interactions, incorporating 3D chromatin organizations, and modeling the 3D structures, this work contributes to deep understanding on cell type-specific non-coding genetic variations and advancing precision medicine and clinical care.

CHAPTER 2
DISCOVER NOVEL DISEASE-ASSOCIATED GENES BASED ON REGULATORY NETWORKS OF LONG-RANGE CHROMATIN INTERACTIONS

2.1 INTRODUCTION

Genome-wide association studies (GWAS) has been one of the major approaches to identify genetic variants, e.g. SNPs, that are associated with specific phenotypes, such as diabetes, neurodegenerative diseases, autoimmune diseases and cancer[3,88–92]. The statistically significant SNPs from GWAS can indicate genomic loci containing genes whose expression levels are functionally related with the disease status[93]. Although it has been successful in many studies, there are two major limitations of traditional GWAS analysis: 1) limited statistical power due to sample sizes and minor allele frequencies, and 2) the lack of mechanistic understandings of statistical associations. These limitations are especially challenging for some complex diseases, where multiple functionally coordinated genes and non-coding SNPs contribute to the observed phenotypic changes with mild effect sizes[94].

A key to address these two challenges is to systematically delineate the regulatory effects of non-coding SNPs[95] on gene expression. Since the vast majority of the human genome are non-coding, most of the significant SNPs from GWAS are located in non-coding regions instead of in genes[92]. As has been shown, non-coding GWAS SNP hits are enriched in regulatory elements, such as enhancers[96]. Algorithms are therefore needed to leverage the regulatory information from multi-omics dataset to discover more disease-associated genes that are regulated by the non-coding SNPs[97]. It also provides benefits to identify *cis*- and *trans*-factors of co-regulation to further understand the functional relationships among multiple disease-associated genes and SNPs[97,98], as well as mechanistic insights of genetic associations mediated by dysregulation of genes[99–101].

Based on the recognition of transcriptional dysregulation as one of the major mechanisms underlying disease-associated genetics[102], gene regulatory networks have been proposed to improve and interpret GWAS results[94,103–105]. But most network-based GWAS analysis algorithms are based on basic gene co-expression networks, where nodes are genes and edges represent correlated expressions[94,103]. Co-expression

networks are limited in improving GWAS analysis because only genes are modeled as nodes and non-coding SNPs are largely missing in the networks. Furthermore, the co-expression based edges treat the underlying regulatory mechanisms as black boxes without considering the specific *cis*- or *trans*- regulatory elements, such as enhancers and transcription factors (TFs), that mediate the observed correlated activities[103,104].

Due to the recent biotechnology developments, such as Hi-C[14,106], ChIA-PET[107] and Capture-C[108], high-throughput chromatin contact maps are being generated[14,109,110]. These chromatin contact maps provide information of three-dimensional (3D) chromatin structure[14,106–110] and reveal specific long-range chromatin interactions linking distal non-coding enhancers to target genes[14,106–110]. Unlike *cis*-regulatory links identified by eQTL calling[111], the chromatin interactions are based on evidence of physical interactions between enhancers and promoters and can capture longer *cis*-regulatory interactions (e.g. ~1Mb)[14,106–110]. Therefore, the incorporation of 3D chromatin contact maps into regulatory network construction is expected to extend the capability of analyzing distal non-coding GWAS SNPs and their regulatory impacts on specific target genes, which may participate in critical biological pathways associated with the disease. There have been several successful case studies of using long-range chromatin interactions to decode the underlying disease mechanisms, such as SNPs associated with prostate cancer, erythrocyte and triglyceride were found to be enriched in regulatory DNA regions and may disrupt TF binding[96,112,113]. Statistical methods have been built to jointly model SNPs and gene expression with respect to the impacts on disease risks[114,115]. Although the methods show promising results, further incorporation of chromatin interactions to link SNP with specific genes will substantially enlarge the SNP-sets associated with genes and improve the statistical power. A couple of algorithms[104,116] have been developed to utilize chromatin interactions to combine *cis*-regulatory elements with gene-gene networks for disease-association analysis, which can prioritize disease-associated non-coding SNPs. But these methods cannot aggregate the SNP-level signals to discover new genes that might be functionally associated with diseases.

Another layer of complexity of gene regulation comes from combinations of TF bindings to not only gene promoters but also linked distal enhancers[117,118]. Knowing the key TFs as trans-regulatory factors for specific genes can shed light on common

regulatory mechanisms shared by multiple genes and can also indicate how multiple enhancers coordinate together to regulate target genes[119,120]. In this regard, long-range enhancer-promoter interactions provide important information to aggregate TF binding patterns across promoters and enhancers. Combined with GWAS data, candidate master TF regulators of disease-associated genes can be inferred based on enrichment analysis from both promoters and enhancers[121–123].

Given the benefits of integrating 3D chromatin interactions into the study of gene regulation, we have developed a computational infrastructure to construct regulatory networks and make network-based predictions of novel disease-associated genes. The software, named as APRIL, combines cell-type specific epigenomics and transcriptomics datasets with public available chromatin contact maps to build expanded regulatory networks, which include long-range *cis*-regulation between non-coding enhancers and genes and trans-regulatory links of TFs. APRIL also provides functions to analyze co-occurring signatures of GWAS SNPs in the constructed networks, in order to obtain insights on functional coordination of disease-associated genetic variants. Furthermore, APRIL contains both unsupervised and supervised machine learning algorithms to predict new disease-associated genes, by leveraging the information from non-coding regulatory SNPs and the regulatory network structures. The application of APRIL on GWAS studies of leukemia demonstrates that not only new disease-associated genes can be reliably predicted but also regulatory mechanisms underlying diseases can be indicated.

2.2 MATERIALS AND METHODS

2.2.1 Dataset and annotations/definitions

For chromatin interaction datasets used in this study, the ChIA-PET dataset of two replicates are collected from GSE39495[124] for K562 cell-line. The two replicates are merged together, leading to ~100k interactions in total. Additional public-available high quality chromatin interaction datasets, including ChIA-PET[107], Hi-C[106], Capture-C[108] and computationally predicted enhancer-gene interactions, i.e. JEME[125] and IM-PET[18], can also be used as inputs.

GENCODE version 19[126] is used as the gene model annotations. Only protein coding genes are considered. The promoters for each gene are defined as the +/- 1kb

region centered at the transcription start sites (TSS). In total it includes ~20k protein-coding genes. The consensus enhancer annotation is collected from the Roadmap Epigenome and ENCODE consortia[96,124], using the version filtered by DNase-seq signals (-log10(p) >= 10). The enhancers are further annotated as cell-type specific enhancers based on chromatin segmentations learnt from chromHMM in all 127 cell lines/tissues based on histone modifications[96].

For epigenomics dataset, the imputed DNase-seq and RNA-seq data of 127 cell lines are collected from the Roadmap Epigenome and ENCODE consortia to quantify cell-type specific enhancer activities and gene expression[96,124]. For TF binding site annotations, TF motif hits predicted by Kheradpour *et al*[127] are collected. The version filtered by conservation scores across multiple species (>0.3) is used. This resulted in ~13.6M motif hits for ~500 TFs.

Known disease-associated genes are collected from DisGeNET[128]. This curated dataset includes gene-disease associations from multiple resources, including UNIPROT, CTD, ORPHANET, CLINGEN, GENOMICS ENGLAND, CGI and PSYGENET[129–132]. Immune-associated genes are identified based on keywords matching summarized in Supplementary Table (**Figure A. 1**). GWAS SNPs associated with immune are collected from two sources: Biobank[133] and EMBL EBI[134]. Significant eQTLs in whole blood are collected from GTEx V7[111]. The nominal p-values of eQTLs are provided by GTEx and are used in this paper.

2.2.2 Regulatory network construction based on long-range interaction

Construct 3D chromatin modules of long-range *cis*-regulation

Groups of genomic fragments that are inter-connected with each other by chromatin interactions, i.e. 3D chromatin modules, are first identified in the APRIL algorithm (**Figure 2.1 A**). Within each 3D chromatin module, nodes represent genomic fragments and edges represent long-range chromatin interactions. For each pair of nodes in a module, there exists at least one path connecting them. Every chromatin interaction has two linked genomic loci, i.e. interaction anchors, and different chromatin interactions may involve overlapping interaction anchors. To create a catalogue of unique indexes for interacting

genomic locations, consecutively overlapping anchors along the genome are merged into single fragments and are represented as single nodes in the network.

Genomic fragments (i.e. nodes) in 3D chromatin modules are further annotated based on their potential functions in gene regulation. The nodes are classified into three types: gene nodes, enhancer nodes and other-element nodes. A genomic fragment is classified as a gene node if it overlaps with a gene's promoter (i.e. +/-1kb from TSS) based on the provided gene model annotations and if the gene is expressed in the specific tissue or cell-type (log2(RPKM)>0) based on the provided transcriptome data. A genomic fragment is classified as an enhancer node if it contains enhancer-specific epigenetic signatures provided by the user. Enhancer-specific epigenetic signatures can be:

a) enhancer chromatin state called by chromHMM or Segway[135,136],

b) high chromatin accessibility such as DNase-seq peaks or ATAC-seq peaks,

c) enhancer-specific histone modification (e.g. H3K4me1 or H3K27ac) ChIP-seq peaks,

d) enhancer RNA (eRNA) signal peaks.

Users have the flexibility to select the specific enhancer epigenetic signatures to run APRIL. Genomic fragments that are not classified as genes or enhancers will be considered as other-elements. Based on these annotations, nodes will be color coded in the final network visualizations (**Figure 2.1 A**). Overall, the annotated 3D chromatin modules represent connected regulatory units of multiple genes and multiple enhancers. The chromatin interactions within each module represent long-range *cis*-regulation among pairs of enhancers, other-elements and gene promoters, which are usually co-located in chromatin domains[137].

Expand regulatory sub-networks with trans-regulation

The regulatory networks will be further expanded from 3D chromatin modules, by incorporating connectivities of trans-regulation. Although long-range *cis*-regulation within 3D chromatin modules are mainly generated by intra-domain interactions[14,137], trans-regulation by transcription factors (TFs) can coordinate expression of genes located in different chromatin domains. The APRIL algorithm builds expanded regulatory sub-networks by merging 3D chromatin modules that share common enriched TFs (**Figure 2.1 A, Figure A. 2**). Based on the genome-wide TF motif annotations provided by the

user, APRIL will first scan genomic fragments in every 3D chromatin module by counting the motif occurrences for each specific TF. The motif occurrence counts across all genomic fragments are then organized into a node-level TF matrix, where each row corresponds to a genomic fragment (i.e. nodes in the network) and each column corresponds to a specific TF. To reduce false positives of TF motif analyses, only TFs that are expressed (log2(RPKM) > 0) in the specific tissue or cell-type under study are included. Based on the node-level TF matrix, a module-level TF matrix is then constructed, where each row corresponds to a 3D chromatin module and each column corresponds to the averaged TF occurrence counts across nodes within the specific module. Hierarchical clustering is then applied on the module-level TF matrices to identify clusters of 3D chromatin modules whose TF occurrence profiles are significantly similar (**Figure A. 3**). Considering the large variances across different TFs, 'complete' mode of hierarchical clustering is used. Pearson correlation is used as the similarity metric. The default number of clusters is determined based on analyses of Within-cluster Sum of Squares and Averaged diameters of the resulting sub-networks (**Figure A. 3**). 3D chromatin modules belonging to the same clusters are linked together by adding 1) nodes to represent the common TFs shared by the chromatin modules, and 2) edges between TF nodes and genomic fragments in the 3D modules which contain motifs of the specific TFs. The resulting graph of multiple 3D chromatin modules linked by common TF nodes are used as the expanded regulatory sub-networks, which serve as the foundation for the subsequent disease-associated gene predictions.

In the meantime, based on the hierarchical clustering, clusters of TFs that are enriched within specific regulatory sub-networks are identified as candidate master TF regulators, and a clustered heatmap is generated to visualize the associations of candidate master TF regulators and specific regulatory sub-networks.

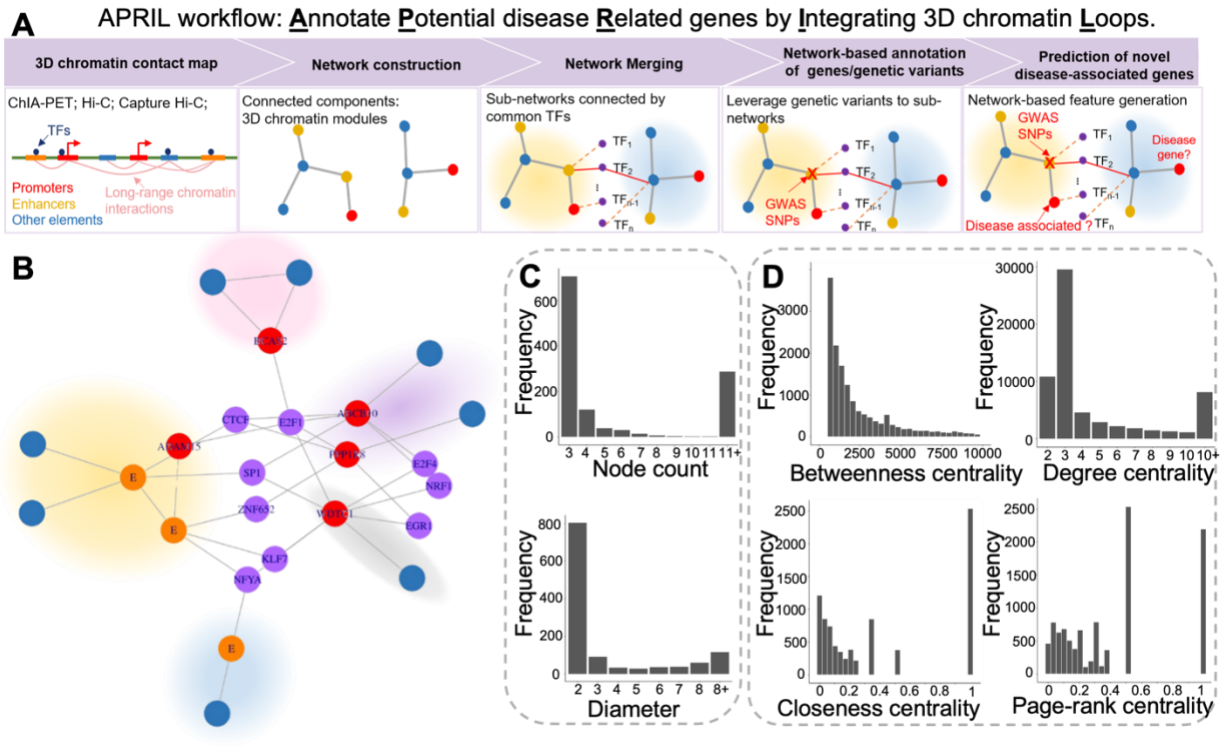

Figure 2.1. The workflow of APRIL and its application on K562 ChIA-PET dataset. (**A**) APRIL takes 3D chromatin contact maps and cell-type specific multi-omics dataset as inputs to construct 3D chromatin modules including gene nodes (red), enhancers nodes (orange) and other-elements nodes (blue). Enriched TFs based on motif analysis are added as TF nodes (purple) and are used to merge 3D chromatin modules into sub-networks. Based on the GWAS SNPs and the constructed regulatory sub-networks, novel disease-associated genes are predicted using different machine learning techniques. (**B**) One example of the constructed regulatory sub-network based on K562 ChIA-PET data. (**C**)-(**D**) Statistical distributions of network properties. Global graph properties (**C**) include Node count (upper panel) and Diameter (lower panel). Node centrality metrics (**D**) include Betweenness centrality (upper left panel), Closeness centrality (lower left panel), Degree centrality (upper right panel) and Page-rank centrality (lower right panel). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

2.2.3 Network-based enrichment analysis for genetic associations

Before predicting novel disease-associated genes, APRIL provides a series of functions to statistically test whether GWAS SNP hits are over-represented within specific regulatory sub-networks. Based on the provided GWAS summary statistics (e.g. association p-values of SNPs) from the user, APRIL first identifies genomic fragments (i.e.

nodes in regulatory sub-networks) that harbor those SNPs. Geodesic distances, which are defined as the lengths of shortest paths between two nodes harboring GWAS hits, are calculated to quantitatively demonstrate how close GWAS hits are located across the topology of the regulatory sub-networks. 1,000 random controls are generated by shuffling the labels of nodes (i.e. GWAS-hit labels and non-GWAS-hit labels) across all regulatory sub-networks. The observed distribution of geodesic distances based on real GWAS hits are then compared to the distributions from the 1,000 random control sets using Kolmogorov-Smirnov tests to statistically evaluate whether significant GWAS hits prefer to be located closer in the network.

For each regulatory sub-network, APRIL further statistically tests whether the specific sub-network has significantly higher frequency of observing GWAS hits. 1,000 random control sets are generated by shuffling the labels of nodes across all sub-networks (i.e. GWAS-hit labels and non-GWAS-hit labels), with the total number of nodes and edges in each sub-network controlled. For each sub-network, using the counts of GWAS-hit nodes in the 1,000 random sets as the null distribution, we calculated the empirical p-value using the formula: 1+No. of (GWAS-hit in random networks > observed GWAS-hit in APRIL sub-network)) / 1001. The regulatory sub-networks are then sorted based on their GWAS enrichment significance.

2.2.4 Prediction of novel disease-associated genes

APRIL employs both unsupervised and supervised machine learning algorithms to predict novel disease-associated genes, based on the constructed networks of long-range gene regulation. Users have the flexibility to choose the prediction methods and do comparative analysis on the results. The unsupervised prediction is based on label propagation[138] and the supervised prediction is based on random forest models.

Unsupervised label propagation to predict disease-associated genes

APRIL utilizes HotNet[138] algorithm to predict novel disease-associated genes. Different from traditional applications, APRIL incorporates three different types of nodes in the network for label propagation, i.e. gene nodes, enhancer nodes and TF nodes. As in the HotNet[138] diffusion framework, every regulatory sub-network is considered as a dynamic fluid system, where each node (except TF nodes) serves as a source of fluid with a

constant rate and fluid can diffuse from node to node within the sub-network. APRIL uses the GWAS association effect sizes of each node as the corresponding node-specific source fluid rates. Therefore, nodes harboring significant GWAS SNP hits are assigned with higher source fluid rates. If a node contains multiple SNPs, the maximum effect size is used for the node. Sub-networks with less than five gene nodes are removed from this analysis. After HotNet label propagation converges to the steady state, a predicted disease-association score is assigned for every node, especially for gene nodes. Top-ranked genes based on the association scores are highlighted as candidate disease-associated genes.

Supervised prediction of disease-associated genes using random forest models

To construct predictive models of disease-associated genes and identify key network features, APRIL provides a supervised mode of predictions based on random forests. For every gene node in the sub-networks, the binary label to predict is disease-associated or not. The features used in random forest to predict the disease-association labels for a specific gene node include five sets: 1) regulatory sub-network related features, such as node degree, betweenness, closeness, page rank centrality; 2) GWAS related features, such as the maximum and summation of effect sizes of neighboring nodes; 3) TF related features, such as TF motif occurrences in the gene node and neighboring nodes. For each specific TF, the motif occurrence counts from neighboring nodes are aggregated into a weighted summation score, where the weights are GWAS effect sizes assigned to each neighboring node; 4) eQTL features in neighboring nodes, including the count of eQTLs and the p-value of the most significant neighboring eQTL; and 5) cell-type specific activity features of genes and enhancers, such as the z-scores of gene expression and enhancer chromatin accessibility across multiple cell-types. Since different gene nodes may be connected to different numbers of enhancer nodes, APRIL uses quantile statistics of enhancer activities to uniformly describe activity distributions of neighboring enhancer nodes for every gene node. To train the random forest model, balanced negative training sets are generated from the pool of genes in regulatory sub-networks that are not identified as disease-associated based on traditional GWAS and do not share common neighboring nodes in the sub-networks with genes in the positive training sets. Random forest classifier is trained using the R package ‘randomForest’ with 50 trees. Ten-fold

cross-validation is used to evaluate the performance. Feature importance is calculated and provided to users.

2.2.5 Cell-type specific activity analysis

Based on user provided cell-type specific activity datasets for genes and enhancers, APRIL also provides functions to understand how cell-type specific activities are coordinated in regulatory sub-networks. To quantitatively evaluate whether genes, enhancers and TFs are significantly correlated, APRIL provides different versions of random backgrounds to carry out statistical tests. The first version of background is created by randomly pairing nodes in the same networks, with all other factors controlled. The second version is based on all potential pairs. For each version, 1,000 random samples are generated. Pearson correlations of observed node pairs are calculated based on quantile normalized cell-type specific activities. Similar correlation calculations are carried out for node pairs from random samples. The observed correlations are then compared with random background and are statistically tested (one-sided Student's t-test). Three types of node pairs are considered: 1) gene-gene pairs, 2) enhancer-gene pairs, and 3) TF-gene pairs. Since the regulatory sub-networks are constructed based on information of 3D chromatin interactions and TF motif occurrences, the cell-type specific activity correlations and the significance tests provide orthogonal information on functional coordination of gene regulation at systems-level. Furthermore, APRIL employs public available protein-protein interaction (PPI) databases[139] to evaluate whether TF-TF pairs in regulatory sub-networks are statistically enriched with PPIs based on empirical permutation tests. For each sub-network, APRIL generates 1,000 random sets of TF-TF pairs by shuffling TF names across all sub-networks. This procedure controls all other factors and can efficiently eliminate bias. The fractions of PPI-supported TF-TF pairs in the observed sub-networks are then compared with random samples to quantitatively test whether TFs co-regulating genes are more likely to physically interact with each other.

2.2.6 Network and prediction visualizations

The constructed regulatory sub-networks are visualized using the R package 'igraph'. In the network visualization, gene nodes are marked as red, enhancer nodes are marked as orange, other-element nodes are marked as blue, and TF nodes are marked as purple.

Edges of gene-gene pairs, enhancer-gene pairs and enhancer-enhancer pairs represent long-range chromatin interactions. Edges of TF-gene pairs and TF-enhancer pairs represent predicted trans-regulation. A plotting function with default setting above is provided to users. Along with the network visualization, additional plots can be generated, including 1) activities of genes, enhancers and TFs, 2) correlations of gene-gene pairs, enhancer-gene pairs, and TF-gene pairs, 3) PPI enrichment, and 4) clustered heatmap of enriched TFs across 3D modules. The plots are generated based on the same method as stated in the methods 2.2.4.

For predictions of novel disease-associated genes, results of the two machine learning methods are provided as a series of plots and tables. The predictions based on label propagation are provided as a table with three columns: gene names, predicted scores and sub-network indexes. Since the raw predicted scores from different regulatory sub-networks are not directly comparable, the mini-max normalized scores are provided for each gene to make it comparable across sub-networks. Along with the table, a boxplot of predicted score ranks between disease-associated genes and control gene sets is also provided to the user. The result can also be visualized in the regulatory sub-networks, where nodes are colored by the predicted scores of label propagation and candidate disease-associated genes are represented as stars. The predictions based on the random forest are provided as a table with two columns: gene names and the predicted disease-association probabilities. The ROC plot of ten-fold cross-validation is provided to the users. The predicted disease-associated genes can be visualized in the sub-networks and the newly discovered candidate genes are highlighted as red stars.

2.3 RESULTS

2.3.1 Expanded regulatory networks constructed based on 3D chromatin interactions.

To demonstrate the performance, APRIL is applied on a ChIA-PET chromatin interaction dataset of K562 cell-line[124]. The distances of chromatin interactions linking regulatory elements and genes can be longer than 1Mb, with a median distance of 25kb (**Figure A. 2**). The regulatory elements mainly include enhancers as characterized by enhancer-specific epigenetic signatures, including enhancer chromatin states and DNase-seq

signal peaks[96,124], along with other potential non-coding regulatory elements. They are termed as enhancer nodes hereafter in this paper. The interacting genomic locations that overlap with gene promoters[126] are annotated as gene nodes. In addition, APRIL identified ~402k DNA motif occurrences for 221 TFs located in interacting genomic locations from this ChIA-PET dataset. By integrating the long-range *cis*-regulation of chromatin interactions and trans-regulation mediated via TFs, APRIL constructed an expanded K562-specific regulatory network, including edges linking different types of nodes (i.e. gene, enhancer and TF nodes).

Overall, there are 10,458 gene nodes, 73,494 enhancer nodes (including enhancer elements and other regulatory elements) and 14,667 TF nodes. There are 741 sub-networks containing at least one gene node. In each sub-network, multiple 3D chromatin modules (i.e. groups of long-range *cis*-regulation) are connected by common enriched TFs (i.e. combinatorial trans-regulation). **Figure 2.1 B** shows one example of the constructed sub-network, where different types of color-coded nodes connected by *cis*- or trans-regulatory links. More examples are shown in **Figure A. 2**. The sub-networks have different sizes, as demonstrated by the number of nodes and the network diameters (**Figure 2.1 C**). In addition, indicated by a variety of centrality metric analyses (**Figure 2.1 D**), substantial fractions of sub-networks are organized in a hierarchical structure with node 'hubs' linking many neighbors in the graphs, emphasizing the importance of understanding gene regulation from a systems-level.

2.3.2 Cell-type specific regulatory signatures encoded in networks

The constructed regulatory sub-networks provide a platform to analyze K562 specific gene regulation, as the networks contain 83,952 genomic locations in total, such as gene promoters, enhancers and other regulatory elements, which participate in long-range gene regulation (**Figure 2.2 A**). In addition, the TF nodes in sub-networks represent abundant trans-regulation mediated by TF combinations. Compared to genome-wide background, genes included in the expanded regulatory networks have significantly higher expression levels in K562 cells (**Figure 2.2 B**). Similarly, enhancers and TFs included in the networks also show higher activities in K562 cells compared to other enhancers or TFs not in the networks (**Figure 2.2 C, D**). These results suggest the

functional importance of nodes contained in the constructed regulatory networks for cell-type specific activities.

To characterize the coordinated activities of genes, enhancers and TFs in specific sub-networks, correlations across different cell-types are calculated for gene-gene, enhancer-gene and TF-gene node pairs, based on DNase-seq and RNA-seq data from 127 cell-types/tissues in ENCODE and Roadmap Epigenomics consortia[96,124]. All three types of node pairs are found to have significantly higher activity correlations than random samples, where basic network topology properties are strictly controlled (**Figure 2.2 E-G**). These observed high correlations indicate that the constructed regulatory networks, which are built from long-range chromatin interactions and TF motif enrichments, can accurately capture the cell-type specific regulatory modules.

To obtain mechanistic understandings of potential multi-TF collaborations involved in gene regulation, APRIL further incorporates protein-protein interaction (PPI) data[139]. Compared to randomly shuffled TF node annotations across the regulatory networks, TF-TF node pairs within the same sub-networks are found to be highly enriched with PPIs (**Figure 2.2 H**), suggesting combinations of TFs involved in co-regulation of genes may have the potential to physically interact with each other.

As the constructed 3D chromatin modules represent *cis*-regulatory units with multiple enhancers and genes, they provide a better way to characterize combinatorial TF groups in gene regulation than analyzing single enhancers or promoters. For every 3D chromatin module, the TF motif counts from enhancer or gene nodes are aggregated together to create module-level TF enrichment profiles. Interestingly, clusters of TF combinations are found to be shared across specific subsets of 3D chromatin modules (**Figure 2.3**). For example, there are 14 3D chromatin modules enriched with TFs of the FOXO family, 8 3D chromatin modules enriched with TFs of the MAF family and 10 3D chromatin modules enriched with TFs of the ETS family. This observation suggests that the clusters of TFs are potential trans-regulators working together to co-regulate multiple genes and enhancers across different 3D chromatin interacting hubs.

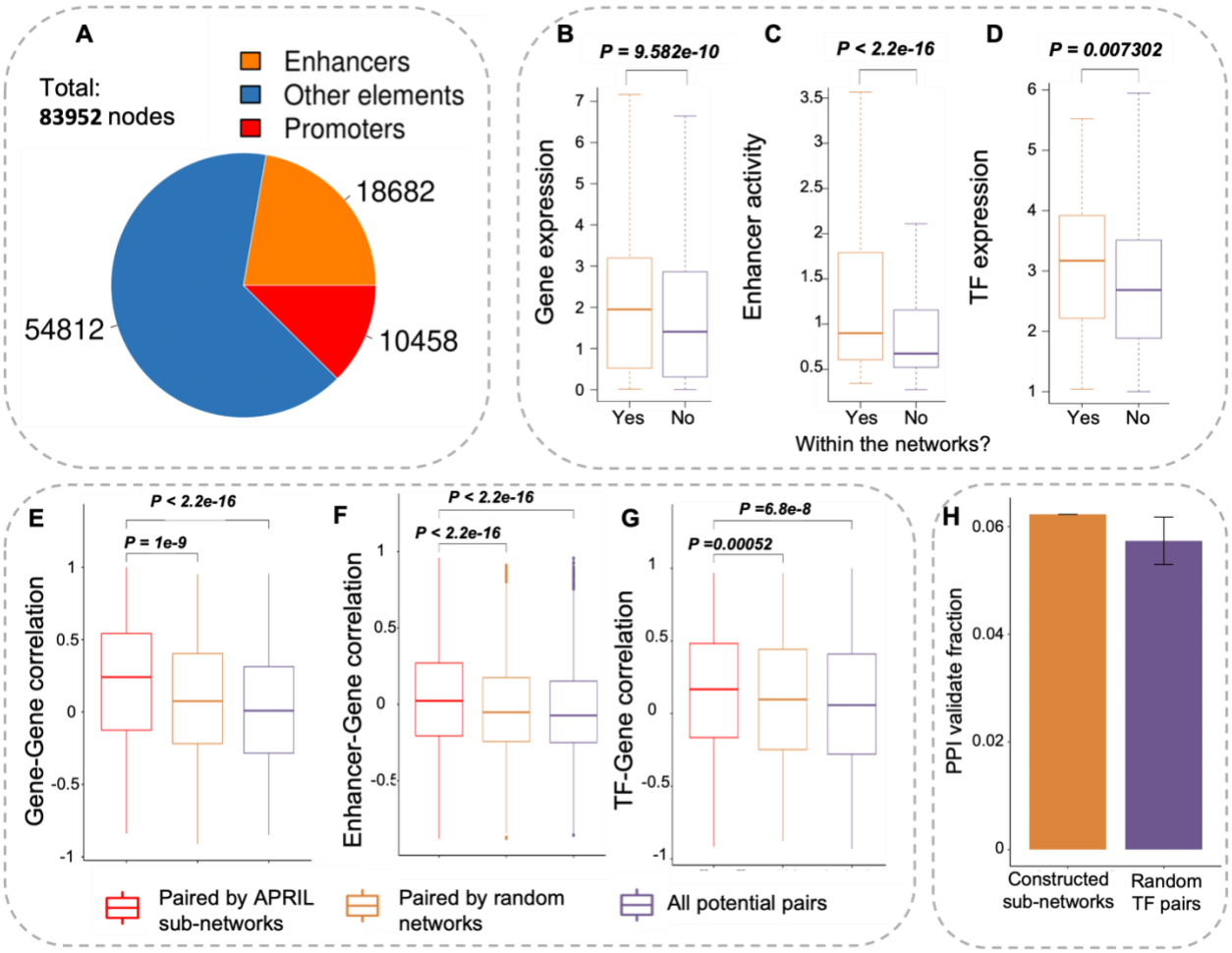

Figure 2.2. Genomic fragments in regulatory sub-networks tend to share similar cell-type specific regulatory activities. (**A**) Pie-chart summary of node annotations. (**B-D**) Regulatory sub-networks constructed by APRIL capture highly active (**B**) genes (based on RNA-seq), (**C**) enhancers (based on DNase-seq) and (**D**) TFs (based on RNA-seq, P = 0.007302) in K562 cells. The controls are genes, enhancers and TFs that are not covered by the regulatory sub-networks. P-values are calculated by the one-sided Student's t-test. (**E**)-(**G**) The constructed sub-networks capture highly co-active (**E**) gene-gene pairs, (**F**) enhancer-gene pairs and (**G**) TF-gene pairs across cell-types (red) based on quantile normalized activities, compared with controls of shuffled networks (orange) and random pairs (purple). P-values are based on the one-sided Student's t-test. (**H**) The TF-TF node pairs within the same sub-networks are enriched with PPIs. Controls are generated by randomly shuffled TF names on the same networks. Error bars are the standard errors based on 1,000 random controls. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

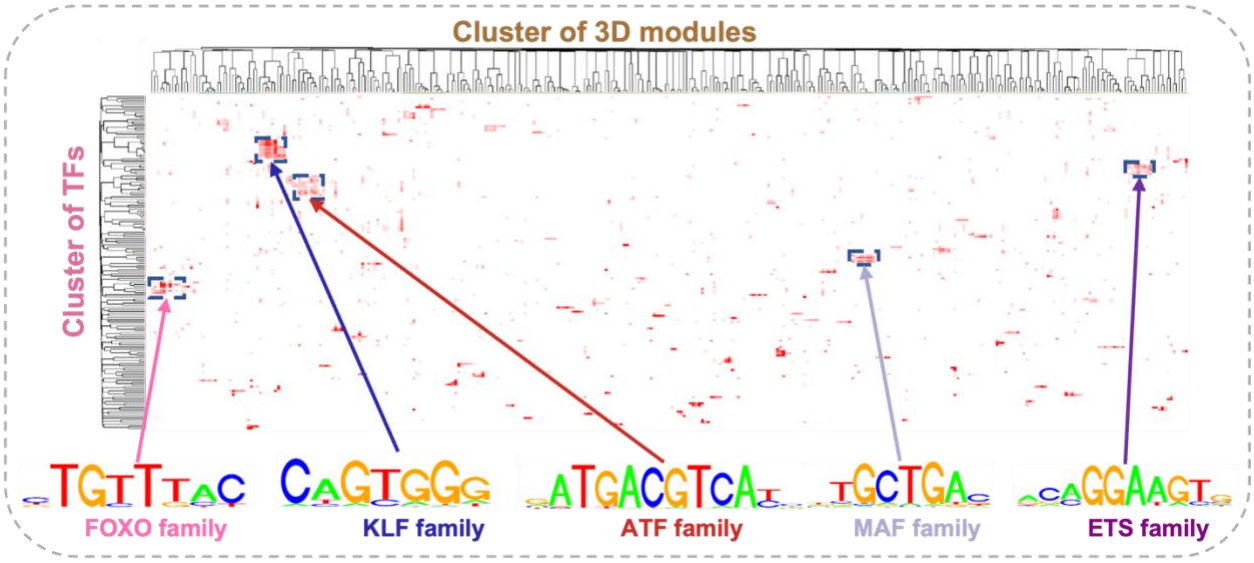

Figure 2.3. 3D chromatin modules form clusters with common TF combinations. Rows represent different TFs and columns represent 3D chromatin modules. TF enrichment within a specific 3D chromatin module is calculated based on motif occurrences in promoters and enhancers. All five highlighted TF combinations enriched in different clusters of 3D chromatin modules are known leukemia or immune related TF families.

2.3.3 Long-range regulatory networks can characterize the relationship among disease-associated genes.

To show how disease-associations are encoded in the constructed K562-specific regulatory sub-networks, the co-occurring patterns of GWAS SNP hits are statistically tested with respect to both topological closeness and enrichment of over-representation. Considering that K562 is a blood cancer cell-line, immune-associated GWAS datasets from the UK BioBank and EMBI databases[133,134] are used for the analysis (see Methods). After overlaying the immune-associated GWAS SNPs to the regulatory networks based on genomic location overlaps, the pairwise geodesic distances for each pair of gene nodes containing significant GWAS SNPs, termed as disease-gene nodes hereafter, are calculated (**Figure 2.4 A**). Compared to randomly shuffled networks, the geodesic distances between disease-gene nodes are significantly shorter (p-value< 2.2x10-16, Kolmogorov-Smirnov test), suggesting that disease-associated genes are closer to each other based on the topology of the constructed networks. This observation also indicates the functional relationship of GWAS SNP hits in the process of gene regulation, which

may mediate the association between the genetic variants and traits through regulatory paths in the networks. Consistent with this hypothesis, the immune-associated GWAS SNPs contained by nodes of the regulatory network demonstrate more stringent GWAS p-values than SNPs that are not contained by the network (**Figure 2.4 B, C**). In addition, disease-associated genes in regulatory networks tend to have relatively higher fractions of neighboring nodes containing eQTLs (**Figure A. 4**).

Furthermore, a subset of regulatory sub-networks are found to be statistically enriched with GWAS genes (**Figure 2.4 D**), compared to randomly shuffled controls using empirical permutation tests with the network topologies maintained. As one of the examples, **Figure 2.4 E** shows a K562 regulatory sub-network containing five genes (red nodes) linked together by a number of long-range *cis*-regulation and six TFs. Three out of the five genes are found to be significantly associated with immunity based on the GWAS analysis (**Figure 2.4 F**), which is significantly more than randomly shuffled controls (p-value=0.008, empirical permutation tests). These sub-networks enriched with GWAS genes suggest that disease-associated genes and the corresponding regulatory elements are functionally inter-connected with each other.

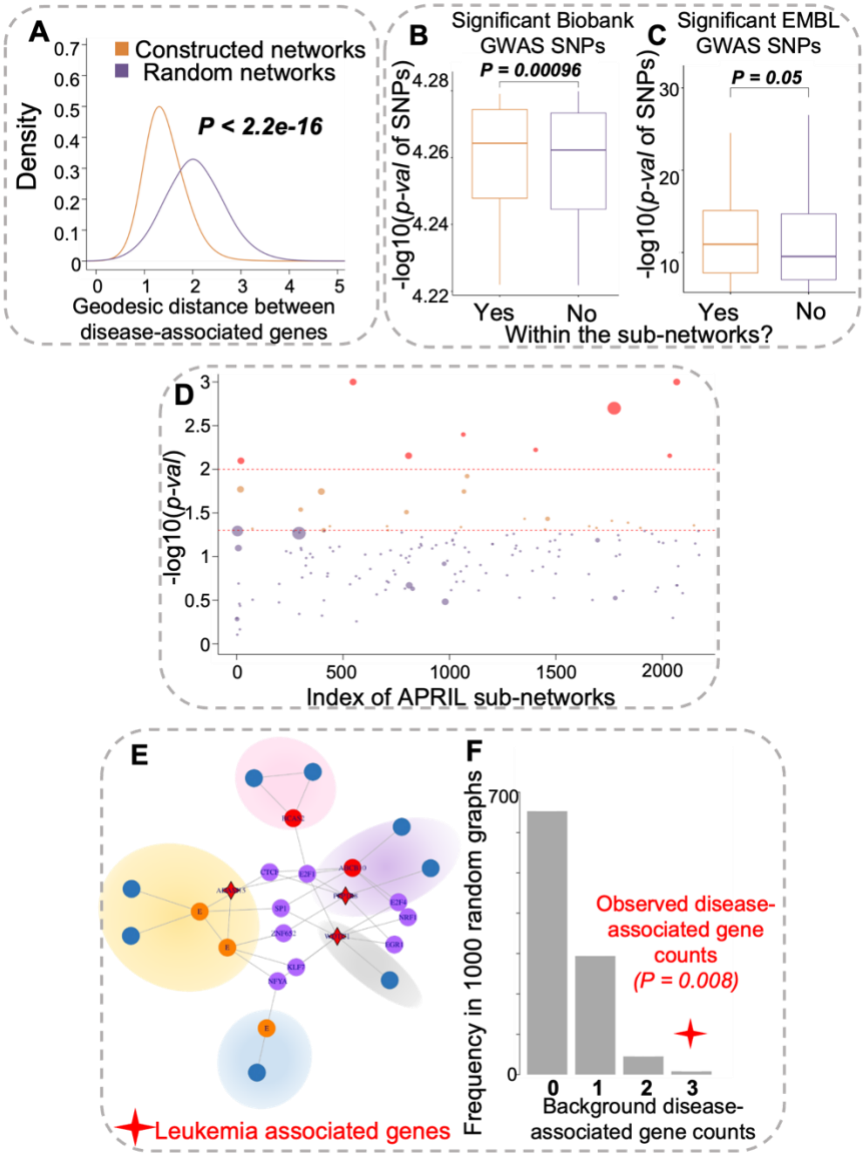

Figure 2.4. Connectivity and enrichment of disease-associated genes and GWAS SNPs in regulatory sub-networks. (**A**) Geodesic distances between disease-gene nodes in sub-networks are significantly shorter compared to random networks (P < 2.2e-16, Kolmogorov-Smirnov test). (**B-C**) GWAS SNPs covered by the regulatory sub-networks are more significantly associated with disease based on data from (**B**) the UK Biobank (P = 0.00096) and (**C**) EMBL database (P = 0.05). The P-value is calculated using one-sided Student's t-test). (**D**) A subset of sub-networks are enriched with disease-associated genes. P-values are based on the permutation tests against random graphs. (**E**) One example of disease-associated genes enriched in a regulatory sub-network in K562 cells. Disease-associated genes are represented as stars. (**F**) Distribution of disease-associated gene counts based on random samples of shuffled GWAS labels (disease-gene and non-disease-gene nodes) across the networks.

2.3.4 Unsupervised disease-associated gene discovery using long-range regulatory networks

As supported by the observed co-occurring GWAS SNP hits in regulatory sub-networks, predictive machine learning algorithms are applied to predict novel disease-associated genes based on network features. As every node in the networks is assigned with a score of association with immunity, based on the GWAS SNPs from UK BioBank[133,134] that overlap with the nodes, HotNet label propagation algorithm[138] is employed to aggregate the scores of connected nodes (**Figure 2.5 A**). Using the GWAS scores from gene nodes alone to do the label propagation, the disease-associated genes from GWAS are significantly top-ranked compared to control gene sets (**Figure 2.5 B**), which is expected. Strikingly, without using any GWAS scores from gene nodes, the disease-associated genes are also significantly top-ranked, based on GWAS scores from enhancer nodes that are linked to the regulatory networks by long-range chromatin interactions with genes (**Figure 2.5 B**). This finding suggests that distal genetic variants in linked regulatory elements are predictive for disease-association of target genes. By combining GWAS scores from both genes and distal enhancers, APRIL can achieve a better separation of disease-associated genes and control genes (**Figure 2.5 B**).

As an interesting example, the gene CTPS1 is predicted to be immune-associated based on label propagation on K562 regulatory networks (**Figure 2.5 C**). The association of CTPS1 with immunity is found to be supported by DisGeNet[128]. As shown in the regulatory sub-network where nodes are color-coded by predicted association scores, CTPS1 is predicted mainly due to a neighboring enhancer node (**Figure 2.5 C**), which has long-range chromatin interaction with CTPS1's promoter based on K562 ChIA-PET dataset. Therefore, the high association score of this particular enhancer with immunity accurately predicts CTPS1's functional relevance in immune systems. Furthermore, this specific enhancer contains multiple SNPs that are also eQTLs of CTPS1 in whole blood tissues (**Figure 2.5 C**), providing cellular level evidence of the genetic association. In addition, CTPS1 has a neighboring TF node: NRF1, which is supported by the ChIP-seq signal peak of NRF1 in the promoter region of CTPS1 (**Figure 2.5 C**). Considering the known functional roles of NRF1 in cancer[140], this observation further indicates potential mechanistic insights of CTPS1's association with immunity.

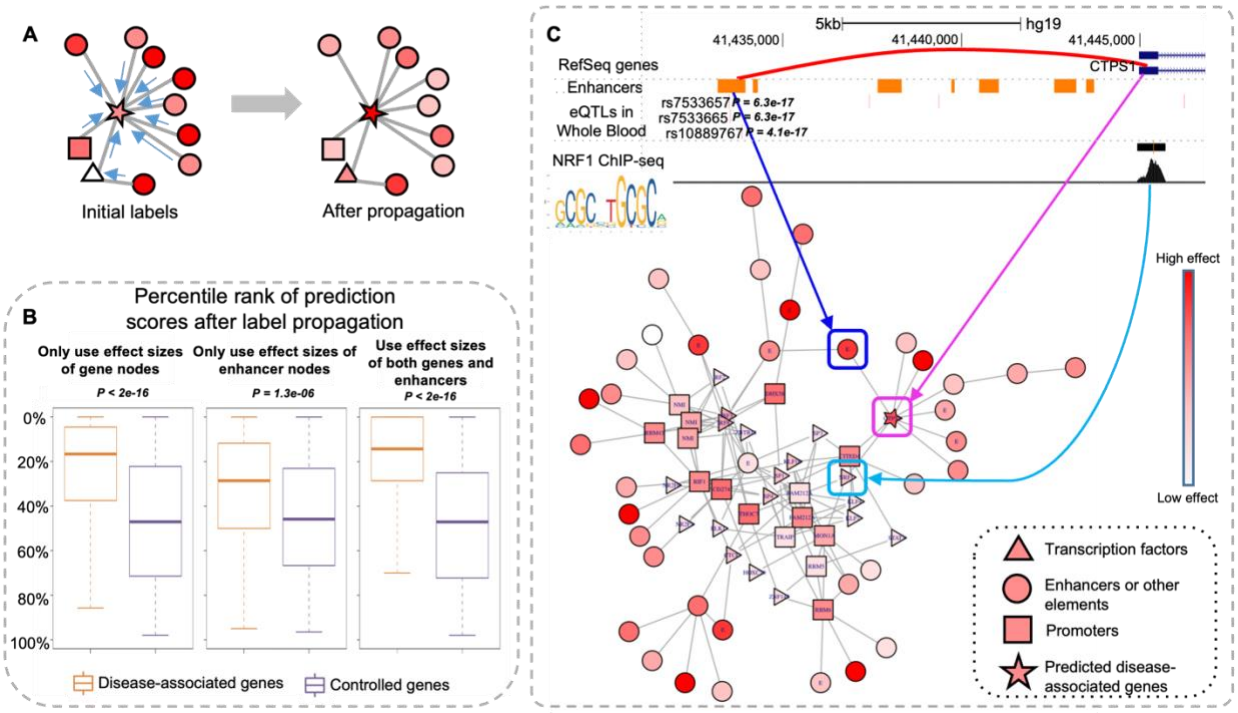

Figure 2.5. Predict disease-associated genes using regulatory sub-networks based on label propagation. (**A**) Schematic figure of label propagation. Disease association scores are propagated along the sub-networks so that every gene node can borrow disease-association information from neighboring nodes. (**B**) Disease-associated genes have higher percentile ranks of predicted scores than control genes (one-sided Student's t-test), suggesting that effect sizes of genetic associations from other nodes in the network can help to improve disease-associated gene prediction. (**C**) Example of newly discovered disease-associated gene CTPS1, which is validated by DisGeNET. The color of nodes corresponds to the predicted disease-association score. The neighboring TF node (NRF1) is inferred based on NRF1 motif in the promoter region of CTPS1, which is supported by the ChIP-seq data of NRF1 binding. Also, the neighboring enhancer node in the network is linked to CTPS1 by long-range K562 ChIA-PET data. The enhancer element contains three significant GTEx eQTLs in whole blood tissues.

2.3.5 Supervised prediction based on aggregated SNP information from neighboring nodes

The random forest based supervised prediction of disease-associated genes is also applied on the constructed K562 regulatory networks, which provides complementary advantages compared to unsupervised predictions. As explained in the methods section, multiple sets of features in gene promoters, linked enhancers and network topology are included (**Figure 2.6 A**). Using the database of DisGeNet[128], genes are labeled as

immune-associated or not associated to train the random forest classifiers. Based on ten-fold cross validation, the random forest models achieve high accuracy (average AUC=0.87) in predicting immune-associated genes (**Figure 2.6 A**). By analyzing the calculated feature importance, the top-ranked predictive features include network-topology related features of genes (e.g. centrality metrics), gene expression levels, specific TFs, and most interestingly, the GWAS effect sizes of linked enhancers in the regulatory networks (**Figure 2.6 B, Figure A. 5, Figure A. 6**). Since the distal enhancers are linked to genes by long-range chromatin interactions, this finding strongly highlights the importance of non-coding genetic variants that may induce gene expression dysregulation in complex diseases. As an example (**Figure 2.6 C**), the gene ABCA7 is predicted to be associated with immunity, consistent with its high expression in whole blood tissues. This gene is located in a large regulatory sub-network with multiple genes, enhancers and TFs densely connected, and it has been found to be associated with autoimmune diseases[141,142]. One of the contributing features for this prediction is a non-coding regulatory enhancer element linked to ABCA7 promoter through chromatin interactions. In support of dysregulation of ABCA7's expression from SNPs in this enhancer, there is an eQTL discovered in whole blood tissue which is correlated with the expression changes of ABCA7 ($p<7.8x10-28$). Other contributing features for this prediction include neighboring TF nodes of KLF16 and SP1, both have been demonstrated to be related with immunity, especially leukemia[143–146]. A similar example of predicted immune-associated genes can be found in **Figure A. 7**, which also leads to insights on the predictive features of *cis*- and *trans*-factors underlying disease-associations.

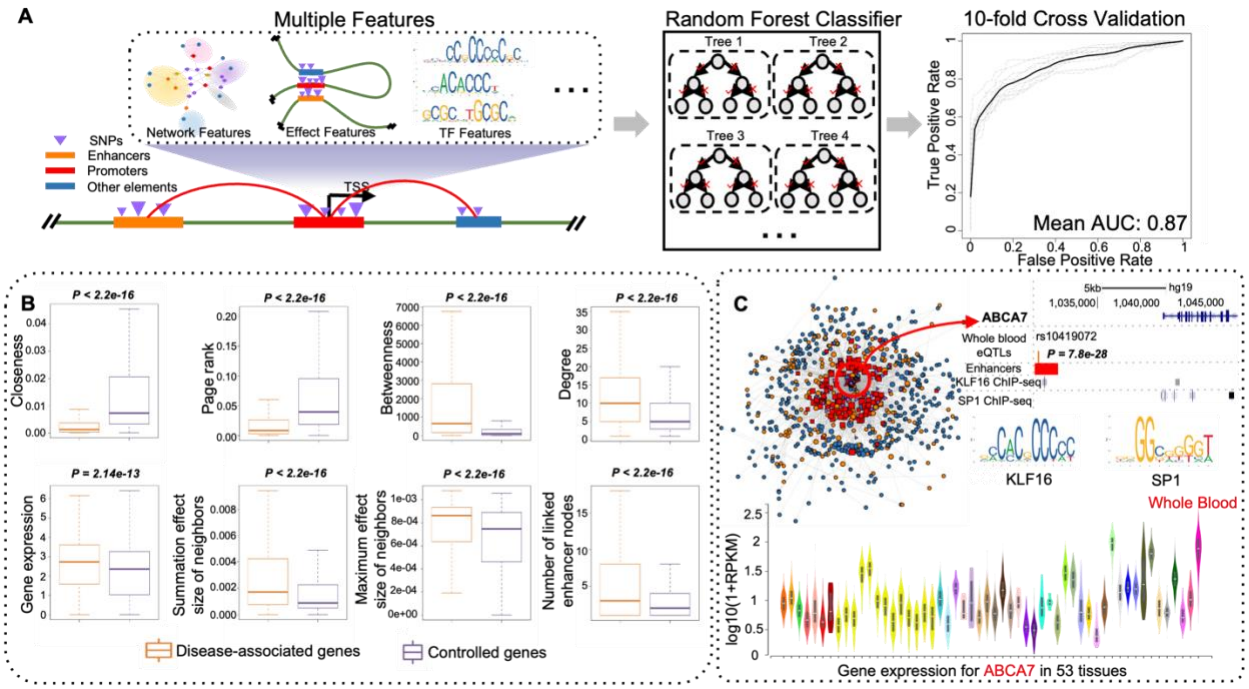

Figure 2.6. Supervised prediction of disease-associated genes based on regulatory sub-networks. (**A**) Overview of features used in the random forest model, including network features, effect sizes, TF features and activity features. The supervised prediction of disease-associated genes achieves an averaged AUC = 0.87 based on ten-fold cross validation. (**B**) Boxplots of top ranked features in predicted disease-associated genes compared to control genes (one-sided Student's t-test). (**C**) Example of a newly discovered disease-associated gene ABCA7. ABCA7 has high expression in whole blood tissues based on GTEx dataset. ABCA7 has two neighboring TF nodes: SP1 and KLF16, based on their motif occurrences. They are validated by ChIP-seq dataset of SP1 and KLF16 bindings. Both TFs are highly ranked with respect to the feature importance calculated by the random forest model. Also, a neighboring enhancer node linked to ABCA7 contains a highly significant eQTL in whole blood tissue based on GTEx dataset (p-value = 7.8e-28).

CHAPTER 3

3DVariantVision: MULTIMODAL FRAMEWORK TO DECODE GENETIC VARIATION BASED ON 3D CHROMATIN

3.1 INTRODUCTION

Decoding the functional impacts of genetic variants plays pivotal roles in revealing the underlying mechanisms of complex human diseases, such as Alzheimer's disease[1,147,148], autoimmune diseases[121,149] and cancer[150–152]. Although conventional genome-wide association studies (GWAS) have enabled the identifications of specific single-nucleotide polymorphisms (SNPs) associated with different diseases, its analytical framework imposes significant limitations, such as the low statistical power, the ambiguity among neighboring SNPs in LD, the restricted capability of pinpointing distal non-coding SNPs, and the lack of mechanistic interpretations[4,5]. These challenges are collectively caused by the limited sample sizes in disease genetics, moderate SNP effect sizes, the burden of genome-wide multiple hypothesis testing, the LD blocks, and, particularly, the simplified black-box treatment of molecular-level SNP effects, such as the functional dysregulation of cell-type specific transcription[4,5].

To address these challenges, one promising strategy is to integrate the multi-omics information under specific cellular contexts into the analysis and leverage the molecular-level phenotypes as mediators to decipher the disruptive impacts of SNPs. This strategy, depending on the specific modeling architectures, has led to a series of important discoveries, including molecular-level genetic associations (e.g. eQTLs, histone QTLs, and meQTLs)[153–156], transcriptome-wide association studies (TWAS)[157], and *de novo* machine learning predictions of cell-type specific SNP effects (e.g. delta-SVM, GWAWA, FunSeq)[43,44,47]. More recently, given the fast advancements in deep learning, a variety of natural language processing (NLP) models have been successfully adapted to predict non-coding SNP effects at single-nucleotide resolution. Based on Convolutional Neural Networks (CNN) and Bidirectional Long Short-Term Memory (BLSTM), DeepSEA[69], Basenji[72], DanQ[71] and BPNet[158] can predict cell-type specific transcription factor (TF) binding sites, chromatin accessibility, histone marks and nearby gene expression, which are further used to predict the disruptive effects induced by specific SNPs. Also based on

CNN architecture, SPEID[20] and [21] can predict enhancer-promoter links, while Akita[24] and DeepC[23] characterize Hi-C chromatin interactions. As the most recent development beyond the CNN-centric framework, Enformer[75] employs the attention mechanisms[74] and extends the genomic window of input features to be 100kb centered around gene promoters, in order to learn the long-range regulatory effects of non-coding SNPs, which yields promising results. Overall, these models substantially expand the functional annotation from coding regions to non-coding genetic variants and provide mechanistic hypotheses on potentially functional SNPs. However, existing approaches face challenges: 1) Integrating long-range and local effects remains unaddressed, and 2) the gap between chromatin and gene expression effects persists. 3) lacks cell-type specificity in chromatin interactions and fails to illustrate the relationship between chromatin and gene expression effects. As a result, a holistic comprehension of genetic variant impacts on local profiles, long-range chromatin, downstream target genes, and associated diseases remains incomplete.

Here we introduce 3DVariantVision, a cutting-edge deep learning-based multimodal framework, designed to offer a comprehensive view of genetic variant effects, spanning from genotype to phenotype. This innovative approach integrates diverse genome-wide datasets, including ChIP-seq[9], Hi-C[14], and eQTL summary statistics[159–161]. Initially pre-trained on predicting cell-type-specific 1D genomic and epigenomic profiles, 3DVariantVision crafts informative embeddings solely from DNA sequences, employing cross-attention in communicative learning to discern intricate enhancer-promoter relationships. Fine-tuning eQTL prediction tasks, it bridges the knowledge gap between 3D chromatin structure and gene expression. Remarkably, relying solely on DNA sequences, 3DVariantVision excels in predicting functional genetic variants, deciphering their disruptive effects on TF bindings, histone modifications, enhancer-promoter interactions, and gene expression. This enhanced capability leads to improved eQTL discoveries. By capturing the intricate, nonlinear regulatory grammar across multiple feature levels and genomic distances, 3DVariantVision accurately identifies the distal target genes of non-coding variants, unveils novel TF combinations, and provides fresh insights into trait-associated variants, transcending conventional GWAS and TWAS studies.

3.2 MATERIALS AND METHODS

3.2.1 Model framework of 3DVariantVision

We introduce 3DVariantVision, an innovative multimodal framework tailored to deciphering the effects of genetic variations using a 3D chromatin context. The framework operates across three distinct stages to unravel the effects of these variants: 1) Disruption of upstream TF binding effects; 2) Influence of chromatin interaction effects; 3) Manifestation of downstream gene expression effects, encompassing eQTLs. Through these three intricately connected stages, 3DVariantVision achieves a comprehensive and nuanced understanding of the complex interplay between genetic variations and their functional consequences within the context of 3D chromatin architecture.

Epigenomics prediction based on representation learning

The 3DVariantVision approach employs a configuration consisting of three CNN blocks, succeeded by two fully connected blocks. This architecture is applied to predict 139 distinct ChIP-seq peaks within the human genome, encompassing TF binding sites, histone modifications, and DNase peaks. As input data, the one-hot-encoded DNA sequence is utilized, where nucleotides A, C, G, and T are represented as [1,0,0,0], [0,1,0,0], [0,0,1,0], [0,0,0,1], respectively, while N represents an unresolved base as [0,0,0,0]. The input sequences are of length 2,000 base pairs.

The employment of CNNs in this context derives from their capability to effectively capture local sequence patterns. The CNN-based architecture serves to generate a representative embedding of the DNA sequence. Each individual CNN block comprises a convolutional layer, which is subsequented by Batch Normalization, Rectified Linear Unit (ReLU) activation, and Max Pooling operations. This orchestrated sequence of layers within each CNN block contributes to the network's capacity to discern and encode essential features from the DNA sequence data. Batch Normalization is a fundamental technique in machine learning that enhances training stability by normalizing activations within each mini batch, mitigating issues related to internal covariate shifts and accelerating convergence. ReLU is a widely used non-linear activation function in neural networks, facilitating the incorporation of essential non-linearity by outputting the input directly if it's positive, while transforming negative inputs to zero, enhancing the network's

ability to capture complex relationships within data. Max pooling is a pooling operation, where the input data is partitioned into non-overlapping regions and the maximum value from each region is selected to form a downsampled representation, aiding in feature extraction and spatial hierarchy preservation. The second and third CNN blocks utilized dilated CNN that introduces gaps or "dilation" between filter elements, effectively expanding the receptive field while preserving computational efficiency. This allows dilated CNNs to capture both local and global features from input data, making them particularly useful for tasks involving hierarchical or multiscale information.

The embedding resulting from the CNN blocks is transformed into a 1D vector through a flattening process. Subsequently, this vector is fed into a sequence of two fully connected blocks, which together facilitate the prediction of binary ChIP-seq peaks. The initial block encompasses a Linear layer, accompanied by Batch Normalization, a ReLU activation function, and Dropout operations. This composite arrangement serves to project the embedded features into a 200-dimensional space, enabling enhanced discriminative capabilities. The subsequent block, which follows, comprises a Linear layer designed to produce output vectors of dimensionality 139. These vectors undergo a final transformation using the Sigmoid activation function. This transformation generates probabilities indicative of peak bindings for the 139 distinct categories. The utilization of the Sigmoid function allows for the effective modeling of these binding probabilities, encapsulating the predictive capacity of the network regarding the presence of ChIP-seq peaks across the target genomic regions.

In training the model, we employed the Binary Cross-Entropy as the objective function to gauge the loss. The derivatives of this objective function with respect to the model parameters were computed through a standard backpropagation algorithm. To optimize the objective function and facilitate efficient convergence, we harnessed the power of the Adam optimizer.

For the input preparation in the representation learning stage of 3DVariantVision, we divided the entire human genome into non-overlapping 2kb regions and computed labels for all 139 epigenomic peaks. Training labels were derived from uniformly processed data releases from ENCODE[124], encompassing 127 TF ChIP-seq, 11 Histone ChIP-seq, and 1 DNase-seq datasets specifically from the GM12878 cell type. Each

epigenomic peak was labeled as 1 if the corresponding 2kb bin overlapped with the peak region and 0 otherwise. Our focus was on the set of 2kb bins that contained at least one peak, resulting in a total of 382,742 sequences.

To facilitate model training, we partitioned the data into training, validation, and testing sets, following an 8:1:1 ratio. In summary, each sample within the dataset comprises a 2kb sequence extracted from the human hg19 reference genome, paired with a corresponding label vector encompassing information for all 139 epigenomic peaks. This meticulously curated dataset forms the foundation for training and evaluating the performance of 3DVariantVision in decoding genetic variations within the context of 3D chromatin architecture.

Chromatin interaction prediction based on communicative learning

Through its training for epigenomic peak prediction, 3DVariantVision acquires the capability to generate embeddings that effectively encapsulate the sequence information of any given DNA sequence. These embeddings are derived from the CNN blocks and can be directly leveraged for the prediction of chromatin interactions between two DNA fragments. It's noteworthy that existing methods for chromatin interaction prediction commonly adopt a simplistic approach of concatenating the features or embeddings of two DNA fragments, thereby neglecting to model the joint features that emerge from their interaction. To address this limitation, 3DVariantVision introduces a novel approach known as communicative learning. This approach is designed to foster a deeper understanding of the relationship between enhancer-promoter pairs, leveraging the power of cross-attention mechanisms. In doing so, 3DVariantVision goes beyond the conventional concatenation method by actively learning and capturing the intricate interplay and dependencies between enhancer-promoter pairs, thereby enhancing the accuracy and interpretability of chromatin interaction predictions.

Cross-attention mechanisms, also known as inter-attention, are a crucial component in transformer-based models[74], differentiating themselves from self-attention mechanisms. While self-attention focuses on capturing relationships within a single sequence, cross-attention extends its scope to establish connections between elements in different sequences or entities. This means cross-attention allows the model to weigh and attend to information not just within one sequence but across multiple sequences,

enabling a more comprehensive understanding of relationships and dependencies in broader contexts, making it particularly valuable in tasks involving multiple sources of information or different modalities, such as translation, summarization, or multi-modal understanding.

Derived from the CNN, the one-hot encoding representations of enhancers S_e and promoters S_p undergo a transformation into richer embeddings, denoted as E_e and E_p, respectively. These embeddings are structured as $E \in \mathbb{R}^{K \times L}$, where K signifies the embedding dimension, which corresponds to the number of CNN channels, while L stands for the sequence length of the enhancer or promoter after pooling processes. In this arrangement, each column within the matrix symbolizes the embedding of a specific minuscule region. Positional encoding is strategically employed to incorporate positional information into the embeddings of both enhancers and promoters.

In the context of implementing cross-attention, we have three fundamental components at play: query (X_q), key (X_k), and value (X_v). The query represents the information we are actively seeking, the key provides the context or reference, and the value encapsulates the content we are searching through. The query and the key are multiplied together to produce the attention scores, which are then used to compute the weighted sum of the values.

In our communicative learning framework, the promoter embedding takes on the role of the query $X_Q = E_p$, while the enhancer embedding assumes the roles of both key and value $X_K = X_V = E_e$. Each of these components possesses its unique set of weights for linear transformations within distinct representation subspaces. Queries $Q = W_Q X_Q$, in essence, represent the promoter-specific information at each minuscule region, whereas keys $K = W_K X_K$ encapsulate enhancer-specific details, enabling each minuscule region to focus on relevant information during the attention process. The scaled dot-product of these queries and keys gives rise to the attention matrix A, where $a_{ij} = softmax(\frac{q_i {k_j}^T}{\sqrt{K}})$. Here q_i and k_j represent the ith and jth minuscule region of query and key, respectively.

Values $V = W_V X_V$, on the other hand, encapsulate the enhancer-derived information. Each individual attention head computes its output as a weighted summation across all input positions by ${E_p}^{reweighted} = AV$. This enables each promoter minuscule

region to glean insights from the entirety of its corresponding enhancer minuscule region. Crucially, the multiple attention heads operate with independent parameters. The outputs from these multiple heads are concatenated to form the final layer output, which is subsequently subjected to a linear layer for aggregation and combination. This comprehensive approach effectively integrates positional and contextual information, enhancing the model's ability to capture intricate relationships between enhancers and promoters in chromatin interactions. In our cross-attention mechanism, we employed four heads, each with a value size of 200 and a key/query size of 200. To effectively capture the intricate, non-linear relationships between enhancers and promoters, we integrated three layers of cross-attention modules. In this architecture, the input to the current module consists of the embeddings of enhancers E_e and the reweighted promoters $E_p{}^{reweighted}$ obtained from the preceding module.

Following the establishment of the intricate relationship between enhancers and promoters, the cross-attention mechanism facilitates the creation of a reweighted embedding of promoters, influenced by their paired enhancers. The attention matrix A generated in the final layer of the cross-attention mechanism is referred to as the "communication map." This communication map serves as a representation of the detailed relationships between enhancer-promoter pairs and is then transformed into a 1D vector through a flattening process. Subsequently, this vector is channeled through two fully connected blocks. These blocks serve to process the information and generate binary predictions pertaining to chromatin interactions. In essence, this series of steps consolidates the knowledge acquired from enhancer-promoter associations and leverages it to make informed predictions about the presence or absence of chromatin interactions.

To construct the enhancer-promoter dataset, we sourced chromatin interactions from Capture-C data within the GM12878 cell type[108]. These interactions were generated by overlapping fragment pairs with active enhancer and promoter regions associated with genes, thus forming the positive enhancer-promoter interaction dataset. To maintain a balanced dataset, we created a negative set by randomly pairing enhancers and promoters while preserving the same distance distribution observed in the positive set. For partitioning the data into training, validation, and testing sets, a careful approach was

undertaken. We divided the data by chromosomes, ensuring strict non-overlapping partitions to mitigate overfitting concerns. Chromosome 9 was reserved for validation, whereas chromosomes 10 and 11 were allocated for testing purposes. The remaining chromosomes constituted the training dataset.

Similar to the representation learning stage in the first phase, we employed the Adam optimizer to optimize the target function, which in this case was Binary Cross-Entropy.

eQTL prediction based on transfer learning

The final stage of our model focuses on predicting eQTLs by leveraging information gleaned from both 1D epigenomic signals and 3D chromatin interactions. These insights are captured through the preceding stages of representation learning and communicative learning. To accomplish this, each SNP-gene pair is examined, and the genomic region spanning 2kb centered around the SNP is designated as the "enhancer." Consequently, we establish the corresponding enhancer-promoter pairs derived from these SNP-gene associations. Both reference enhancers and alternative enhancers are constructed based on the underlying genetic variants present in these pairs. These enhancer-promoter pairs, representing both reference and alternative scenarios, are individually fed into 3DVariantVision. After undergoing the cross-attention process, two reweighted promoter embeddings are obtained—one for reference enhancer-promoter pairs and the other for alternative enhancer-promoter pairs. These embeddings collectively encode the features characterizing the SNP-gene pairs.

In the final stage of 3DVariantVision, we employ a Random Forest Classifier. This model is used to predict whether the SNP-gene pairs exhibit an eQTL effect or not, drawing on the information encapsulated in the embeddings. This approach allows us to make informed predictions about the regulatory impact of genetic variants on gene expression, contributing valuable insights to the study of eQTLs.

3.2.2 eQTL dataset preparation

To compile the SNP-gene pair dataset, we aggregated eQTL data from an extensive ensemble of 31,684 blood samples sourced from 37 eQTLGen Consortium cohorts. Given our primary focus on chromatin-mediated eQTLs, we first filtered out SNP-gene

pairs with distances smaller than 5kb and retained only those SNPs situated within enhancer regions. To enhance the statistical robustness of our analysis, we concentrated on SNP-gene pairs with a sample size exceeding 20,000.

To distinguish significant SNP-gene pairs from non-significant ones, we employed a Bonferroni correction threshold of P-value < 0.05 for significance, and P-value > 0.1 for non-significance. It's important to note that one gene may be associated with multiple SNPs based on this criteria.

Creating the positive dataset of eQTLs from the significant SNP-gene pairs involved two key steps: 1) Addressing linkage disequilibrium (LD) concerns, we retained only the most significant SNP within each SNP island. SNP islands were defined by grouping significant SNPs with adjacent distances of less than 2kb. 2) To balance the genetic influence of each gene, we retained a maximum of five of the most significant SNPs for each gene. This ensures a balanced representation of SNP-gene associations. To establish a balanced negative set for our dataset, we followed the same distance distribution as the positive set. This meticulous approach to dataset construction ensures that our analysis is rooted in rigorous statistical criteria and represents a comprehensive understanding of chromatin-mediated eQTLs.

3.2.3 Performance comparison

In the first stage of 3DVariantVision, we conducted a comparative analysis of epigenomic predictions with DeepSEA[69], a cutting-edge CNN-based model renowned for its performance in this domain. To ensure an equitable evaluation, both 3DVariantVision and DeepSEA were trained and tested on identical datasets. We evaluated the predictive performance by calculating the Area Under the Receiver Operating Characteristic Curve (AUROC) for each epigenomic track in the testing set, separately for each model. This rigorous approach allowed us to objectively assess and compare the effectiveness of both models in predicting epigenomic features.

In the second stage of 3DVariantVision, we conducted a comprehensive comparison of enhancer-promoter interaction predictions with previous state-of-the-art models. These prior models included tree-based models utilizing engineered features, as well as deep learning-based models that based on DNA sequences. However, it's essential to note that many of these models merely concatenated enhancer and promoter

information or features without explicitly modeling their joint relationships. One notable model in this context is ProTECT[162], a tree-based model that excels at capturing joint features between enhancers and promoters, specifically through Protein-Protein Interaction (PPI) bindings. Both ProTECT and 3DVariantVision possess the ability to model these joint features, prompting a direct comparison. To ensure fairness in our evaluation, ProTECT was trained and tested on the same datasets as 3DVariantVision, focusing on subsets with PPI occurrences. It's worth mentioning that ProTECT utilizes only PPIs as features, aligning with the approach adopted by 3DVariantVision, which exclusively uses joint features for predictions. Unlike the original ProTECT, 3DVariantVision does not leverage additional information such as genomic distance, activity levels, or epigenomic correlations across different cell types. To gauge prediction performance, we employed ROC curves and quantified performance metrics on the testing set.

Additionally, we assessed the model's capacity to learn joint features by utilizing Capture-C supported enhancer-promoter interactions. To create background data, we randomly shuffled the enhancer-promoter pairs while preserving the same enhancer and promoter information but altering the joint features. We then calculated the odds ratio for both ProTECT and 3DVariantVision based on different percentages of enhancer-promoter pairs predicted as interactions by each model. A higher odds ratio signifies more accurate predictions, highlighting the effectiveness in learning joint features.

In the third and final stage of eQTL prediction, we conducted a comprehensive evaluation of our model's performance by comparing it with the state-of-the-art model known as Enformer[75]. Enformer, trained on DNA sequences, is adept at predicting multi-chromosome gene expression by integrating long-range interactions, thereby demonstrating competence in eQTL prediction. To facilitate this comparison, we retrieved the predicted features generated by Enformer and subsequently trained a Random Forest model on the same datasets used for 3DVariantVision. In assessing the performance, we leveraged ROC curves and compared the performance on the same testing set.

Furthermore, to further substantiate our findings, we amassed three orthogonal fine-mapped eQTL datasets as gold-standards[161,163,164]. We then meticulously constructed balanced negative sets while controlling for genomic distances separately.

Once again, we employed ROC curves to evaluate model performance and conducted a comparative analysis to ascertain the effectiveness of our approach in comparison to existing state-of-the-art methods.

We compared 3DVariantVision predicted eQTLs with the traditional statistical eQTL calling method (GTEx)[160]. We assembled a gold-standard eQTL dataset by aggregating eQTLs from Muther *et al*[163], Battle *et al*[161], and Geuvadis *et al*[164], resulting in a dataset comprising 11,210 entries. To facilitate a rigorous comparison, we generated a background dataset ten times larger by randomly selecting SNP-gene pairs while controlling for distance. Both 3DVariantVision and GTEx were tasked with predicting the same number (11,210) of positive eQTLs. Our analysis focuses on three distinct subsets: 1) 'Recalled_by_3DVariantVision': This subset comprises eQTLs that were successfully discovered by 3DVariantVision but missed by GTEx, representing the intersection between GTEx's False Negatives and 3DVariantVision's True Positives. 2) 'TP_by_GTEx': In this subset, GTEx effectively identified eQTLs, representing True Positives according to GTEx's predictions. 3) 'All_eQTL': This subset encompasses the entirety of gold-standard eQTLs used as the benchmark. We conducted a comparative analysis involving genomic distance and Minor Allele Frequency (MAF) among these three groups.

3.2.4 Comprehensive genetic variant insights

3DVariantVision offers a multifaceted and comprehensive perspective on genetic variants, encompassing a wide array of insights. These include understanding the impacts of genetic variants on TF binding, elucidating chromatin effects, predicting eQTLs, uncovering novel TF combinations, identifying target genes, and explaining associations with various diseases.

Genetic impacts on TF binding

In the case of a SNP, we gather the 2kb DNA fragment centered around the SNP. From this fragment, we construct both a reference sequence and an alternative sequence, reflecting the genetic variants present. Within 3DVariantVision, predictions for epigenomic peaks are made separately for these two sequences. The disparities between these predictions are then quantified as "TF disrupting values," signifying the influence of

the genetic variant on different TF binding patterns. These values serve as indicators of how the variant affects the binding of various TFs.

Chromatin effects

In the context of SNP-gene pairs, 3DVariantVision enables the quantification of chromatin interaction effects by assessing the distinctions between the reference genome and the alternative genome. To elaborate, consider a specific SNP-gene pair: the genomic region spanning 2kb centered around the SNP is designated as the "enhancer." From this, we establish enhancer-promoter pairs that correspond to these SNP-gene associations.

For both reference and alternative scenarios, we construct enhancer-promoter pairs based on the underlying genetic variants inherent in these associations. Subsequently, we individually input these enhancer-promoter pairs into the second stage of 3DVariantVision to obtain separate predictions for chromatin interaction scores. To quantify the chromatin effects, we calculate chromatin changing ratios, which reflect the extent of change and are computed as the ratio between the difference in alternative chromatin interaction and reference chromatin interaction and the reference chromatin interaction itself. Furthermore, the sign of the chromatin changing ratio indicates whether the SNP enhances or diminishes the chromatin interaction. These ratios provide a meaningful representation of the impact of genetic variants on chromatin interactions.

Functional TF combinations for chromatin

3DVariantVision employs cross-attention mechanisms to effectively model the intricate relationships among minuscule regions situated between enhancers and promoters. This enables the identification and prioritization of vital pairs of these minuscule regions, thereby unveiling crucial functional TF combinations that play a role in chromatin interactions.

For each enhancer-promoter pair, we select the top-priority minuscule region pair based on the cross-attention weights. Subsequently, we cross-reference these selected pairs with TF ChIP-seq datasets to determine how many TF combinations are bound to these prioritized minuscule region pairs. To establish a rigorous baseline, we generate background data by shuffling the top-priority minuscule region pairs 1,000 times. We

calculate Z-scores for each TF combination relative to this background. TF combinations are then prioritized based on a p-value threshold of 0.05.

To ascertain the biological significance of these prioritized TF combinations, we evaluate Protein-Protein Interaction (PPI)[139] enrichments. Specifically, we examine the number of TF combinations supported by PPI data, comparing the counts between the prioritized TF combinations identified by 3DVariantVision and a set of randomly shuffled pairs. This analysis serves to validate the relevance of the identified TF combinations in the context of chromatin interactions.

Downstream gene discovery

3DVariantVision exhibits remarkable proficiency in eQTL prediction, a capability that can be harnessed to unveil prospective downstream target genes linked to genetic variants. The eQTL prediction scores, generated by 3DVariantVision, can be computed for any candidate SNP-gene pairs. By identifying the gene within each pair with the highest prediction score, we can effectively prioritize target genes.

This targeted gene prioritization approach goes beyond conventional Genome-Wide Association Studies (GWAS)[3] and Transcriptome-Wide Association Studies (TWAS)[157], offering a pathway to uncover novel mechanistic insights into variants associated with specific traits. It provides a powerful means of understanding the functional implications of genetic variants and their influence on gene regulation, shedding new light on the genetic basis of complex traits and diseases.

3.2.5 Genetic Variant and target gene prioritization

Prioritize the SNP within CRISPRi-perturbed enhancer

We devised a strategy to prioritize the causal SNP among CRISPR-QTLs[13,165], specifically those originating from CRISPRi-perturbed enhancers. For each CRISPR-QTL, we collect all SNPs with a MAF greater than 0.05 located within perturbed enhancers. These SNPs are then paired with their corresponding target genes. Next, we calculate eQTL prediction scores based on models like 3DVariantVision. The SNP-gene pair with the highest score is prioritized, and its associated score is deemed representative of the CRISPR-QTL.

We utilize CRISPR-QTL data to corroborate the eQTL predictions made by 3DVariantVision. To establish a baseline, we generate a background dataset for CRISPR-QTLs by assembling random enhancer-gene pairs with matching enhancer lengths and genomic distance distributions. Employing the aforementioned strategy, we calculate predicted scores for each pair independently using both 3DVariantVision and Enformer models. To visually compare the score disparities between CRISPR-QTLs and the background, we construct boxplots for each model, thus facilitating a thorough analysis. Additionally, a similar analysis is performed based on the ABC model for comprehensive insights.

Prioritize target gene from neighbors based on fine-mapped eQTL

We compiled a dataset of four fine-mapped eQTLs specific to blood cell types, utilizing four distinct fine-mapping methodologies: SuSIE[166], DAP-G[167], CAVIAR[168], and CaVEMaN[169]. For each eQTL, we meticulously generated corresponding control datasets by matching the same SNP to the nearest gene, ensuring that these genes had a False Discovery Rate (FDR) greater than 0.1. Moreover, we imposed an additional criterion for these control genes, requiring that the distance between them and the actual target genes exceeded 1kb. This step was taken to eliminate any potential overlaps in sequence features. Subsequently, we computed prediction scores using 3DVariantVision, and to assess the performance of our model, we employed ROC curves. These ROC curves provided a robust means of quantifying model performance for each fine-mapping dataset individually, thereby allowing for a comprehensive evaluation of 3DVariantVision's predictive capabilities in the context of fine-mapped eQTLs.

3.2.6 Local effect predictions of genetic variant

Variant effect predictions on Saturation Mutagenesis data

In 3DVariantVision, we employ TF disturbing scores as a metric to quantitatively assess the genetic impact on epigenomic signals. This metric serves as a valuable resource for identifying high-effect genetic variants, a task facilitated by integrating insights from MPRA (Massively Parallel Reporter Assay) experiments[170]. The MPRA experiments offer detailed insights into genetic variation effects within ten enhancer and ten promoter regions, providing data at a single-base pair resolution. Building on this data, we crafted

a classification task aimed at identifying high-effect variants amid a broader background. To define positive variants, we employed the following criteria: 1) a p-value threshold of less than 1e-5, and 2) selection of the top N-ranked variants based on their effect sizes. Conversely, for negative variants, we applied these criteria: 1) a p-value greater than 0.05, and 2) a log2(effect size) threshold below 0.05. To ensure the creation of a balanced dataset, we sampled the data accordingly.

Next, we employed a RandomForest model and executed a rigorous 5-fold cross-validation procedure using the TF disturbing scores derived from 3DVariantVision for each variant. Our model underwent training and evaluation separately for the top 200, 500, and 1000 effects on both promoter and enhancer regions. Evaluation of model performance was carried out using ROC curves, providing a robust assessment of the model's ability to discriminate high-effect genetic variants from the background.

Decipher TFs' role as activators and repressors in chromatin

3DVariantVision offers insights into chromatin effects and their directions by computing both the ratio and sign of chromatin changes. These metrics provide a valuable foundation for investigating the intricate relationship between chromatin effects and expression effects. To facilitate this exploration, we partitioned the consolidated eQTL dataset into groups defined by distinct TF ChIP-seq peak bounds on the corresponding enhancers. Within each group, we computed the Spearman correlation between the absolute values of the chromatin changing ratio and the absolute values of eQTL expression effects.

For a deeper understanding of a TF's role as an activator or repressor in chromatin, we devised a sorting mechanism based on the absolute values of expression effects. We introduced a metric to gauge the consistency of effect direction between chromatin and expression, measured as the fraction of concordant signs among the top N eQTLs with the most pronounced expression effects:

$$Consistency(n) = \frac{1}{n}\sum_{i=1}^{n} I(c_i = e_i)$$

where c_i, e_i represent the sign of the chromatin changing ratio and expression effects, respectively. Subsequently, we constructed a plot illustrating the evolving

consistency as N varies. A pivotal point of reference is the consistency at N=100, which is instrumental in determining whether the TF functions as an activator or repressor. This determination hinges on whether the consistency surpasses 0.5, signifying activation, or falls below 0.5, indicating repression.

3.2.7 Disease association evaluation

3DVariantVision delivers a comprehensive understanding of genetic variants and their implications for disease associations. In our analysis of eQTL predictions using 3DVariantVision, we bifurcated the dataset based on whether the SNPs were substantiated by GWAS. To elucidate the distinctions between these groups, we compared the eQTL prediction scores through the use of boxplots.

Furthermore, we curated TWAS data[171] related to lipid traits, identifying 28 statistically significant SNP-gene associations in whole blood cell types with p-values less than 0.05. For each SNP-gene pair, we gathered all coding genes within a 500kb radius from the TSS of the TWAS-targeted gene, constituting the background. Subsequently, we leveraged 3DVariantVision to prioritize genes from this background. To assess performance, we computed the rank of the TWAS target gene within the background based on the 3DVariantVision prediction score. We created a control scenario by randomly selecting one gene from the background and determining the corresponding ranks, thus allowing for robust quantification of 3DVariantVision's prioritization capabilities in the context of TWAS data.

3.3 RESULTS

3.3.1 Comprehensive decoding of genetic variation through 3D chromatin with 3DVariantVision

Recent advances in genome-wide chromatin conformation capture methods have unveiled the crucial role of long-range chromatin interactions in bringing distant *cis*-regulatory elements and promoters into close proximity, thereby exerting influence over gene expression[160]. When a genetic variant resides within an enhancer and disrupts TF bindings within that enhancer, it often emerges as the causal eQTL for the target gene associated with that enhancer (**Figure 3.1 A**). This assumption is substantiated by the

significant enrichment of chromatin interactions observed in fine-mapped eQTL datasets. Specifically, the three eQTL datasets, CaVEMaN[169], DAP-G[167], and CAVIAR[168], exhibit markedly higher proportions of chromatin interaction support in comparison to the genomic distance-controlled background, showing fold changes of 2.75, 2.28, and 1.73, respectively (**Figure 3.1 B**). As an illustrative case on chromosome 13, rs9517725 emerges as an eQTL for the gene ENSG00000125304 in whole blood tissues. Remarkably, this SNP is situated within an enhancer that interacts with the promoter of the same target gene, as supported by Capture-C data[108]. Both the SNP region and the promoter region are bound by critical TFs, which may play a mediating role in gene expression (**Figure 3.1 C**). With the aim of harnessing 3D chromatin information for comprehensive long-range prediction of genetic variant effects, we introduce 3DVariantVision, offering an in-depth understanding of genetic variants' impact.

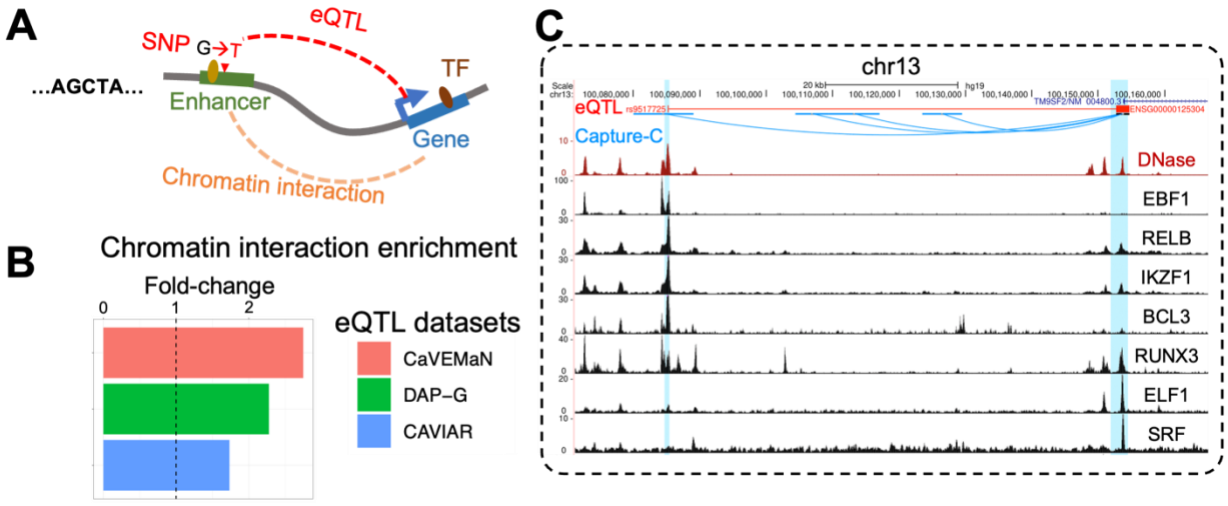

Figure 3.1. Genetic variants impact gene expression through chromatin regulation. (**A**) Chromatin interaction is a key mechanism for genetic variants to regulate gene expression. (**B**) Fine-mapped eQTLs (CaVEMaN, DAP-G, CAVIAR) are enriched in Capture-C supported chromatin interactions comparing the genomic distance controlled background. (**C**) An example of chromatin interaction supported eQTL in the blood cell line.

3DVariantVision is a multimodal framework designed for the comprehensive assessment of genetic variant effects on both local and distant regions within the *cis*-regulatory domain of target genes (**Figure 3.2 A**). This encompassing understanding of

genetic variants necessitates the integration of three distinct types of multi-omics datasets (**Figure 3.2 A Left**): 1) Genome-wide cell-type-specific genomic and epigenomic peaks, comprising 127 TF ChIP-seq, 1 DNase-seq, and 11 Histone ChIP-seq datasets in GM12878[9]. 2) Chromatin interactions, derived from enhancer-promoter interactions sourced from Capture-C data in GM12878[108]. 3) An expansive eQTL dataset in whole blood tissue, meticulously curated through meta-analysis involving 31,684 blood samples across 37 eQTLGen Consortium cohorts[159], thus maximizing statistical power and enabling robust model training. The 3DVariantVision framework serves to quantify variant impacts through multiple lenses by (**Figure 3.2 A Right**): 1) Predicting genetic influences on TF binding sites within the local vicinity. 2) Revelating chromatin effects on long-range chromatin interactions in *cis*, along with insights into their connection with expression effects. 3) Forecasting whether a given SNP-gene pair qualifies as an eQTL or not. 4) Unraveling the roles played by TF combinations within the context of 3D chromatin. 5) Identifying potential distal target genes affected by genetic variants. 6) Shedding light on associations between genetic variants and diseases.

The architecture of 3DVariantVision unfolds through three distinct stages, each contributing to a comprehensive understanding of genetic variants (**Figure 3.2 A Middle**, **Figure B. 1**):

1) Genomic Track Peak Prediction based on Representation Learning (**Figure B. 1 Left**): In the initial stage, we embark on the task of pre-training an embedding model. This model's purpose is to craft informative representations for any given DNA sequence. Here, 1D Convolutional Neural Networks (CNNs) come into play, extracting local features from DNA sequences and predicting genomic and epigenomic profiles within the GM12878 cell line. The input consists of one-hot encoded 2kb DNA sequences, while the output manifests as a binary vector signifying peak occurrences, including ChIP-seq, DNase-seq, and Histone ChIP-seq. Through rigorous training for peak predictions in this first step, the CNN module is subsequently frozen, empowering 3DVariantVision to adeptly generate embeddings that effectively encapsulate sequence information from any given DNA sequence.

2) Chromatin Interaction Prediction based on Communicative Learning (**Figure B. 1 Middle**): The second stage delves into modeling 3D chromatin information by predicting enhancer-promoter interactions. When presented with enhancer-promoter pairs, 3DVariantVision leverages the embeddings of enhancers and promoters, which are generated based on the now-frozen CNNs trained in the initial step. Unlike many existing approaches that simply concatenate features of enhancers and promoters, 3DVariantVision introduces communicative learning. This innovative approach capitalizes on cross-attention mechanisms, inspired by self-attention in Transformer-based models. Cross-attention establishes intricate connections between elements across different DNA fragments, enabling the model to weigh and prioritize information spanning enhancers and promoters. This facilitates a more profound comprehension of relationships and dependencies among enhancer-promoter pairs. After traversing through three layers of cross-attention, a communication map emerges, depicting the intricate relationships between enhancers and promoters. This map is subsequently deployed to predict the existence or absence of interactions.
3) eQTL Prediction based on Fine-Tuning (**Figure B. 1 Right**): In the final stage, the focus shifts to eQTL prediction based on large-scale eQTL datasets. For every SNP-gene pair under consideration, two corresponding enhancer-promoter pairs are constructed based on the reference genome and alternative genome separately, contingent on whether the genetic variant occurs or not. Utilizing the cross-attention block pre-trained in the second stage, two communication maps are generated. These maps serve as the foundation for predicting whether the SNP-gene pair qualifies as an eQTL.

In summation, the 3DVariantVision architecture unfolds across three stages, amalgamating both local features and 3D chromatin information. This holistic approach enables accurate eQTL predictions and a profound comprehension of genetic variants.

Taking the representative long-range eQTL instance of rs12413588, 3DVariantVision predicts it as an eQTL for the gene ENSG00000170525, validated by the fine-mapped eQTL dataset from Battle *et al*[161] (p-value=9.54e-14, distance=258kb) (**Figure 3.2 B**). Furthermore, this predicted eQTL is substantiated by experimentally

verified chromatin interactions identified through Capture-C. Intriguingly, 3DVariantVision reveals that this SNP significantly disrupts the binding sites of crucial TFs, including SMC3, RAD21, CTCF, RUNX3, and PAX5. This disruption leads to a noteworthy 0.02% reduction in chromatin interaction with the target gene, offering insightful explanations for the underlying eQTL mechanisms. In addition, rs79259450 emerges as an eQTL prediction for the gene ENSG00000185112 (**Figure B. 2 A**). This prediction aligns with validation from the fine-mapped eQTL dataset provided by Geuvadis *et al*[164] (p-value=3.70e-11, distance=460kb) and corroborated by Capture-C chromatin interactions. Notably, this genetic variant is anticipated to disrupt the binding sites of TBP, NBN, SPI1, among others, resulting in a 0.06% decrease in chromatin interactions. Similarly, rs813000 surfaces as an eQTL prediction for the gene ENSG00000257923, with validation stemming from Battle *et al*.'s[161] fine-mapped eQTL dataset (p-value=5.03e-37, distance=458kb) and Capture-C chromatin interactions (**Figure B. 2 B**). This SNP is predicted to disturb the binding sites of EBF1, RUNX3, and others, resulting in a substantial 0.17% reduction in chromatin interactions. Collectively, these examples underscore 3DVariantVision's remarkable capacity to predict eQTLs and elucidate the intricate effects of genetic variants spanning both local and distal genomic regions.

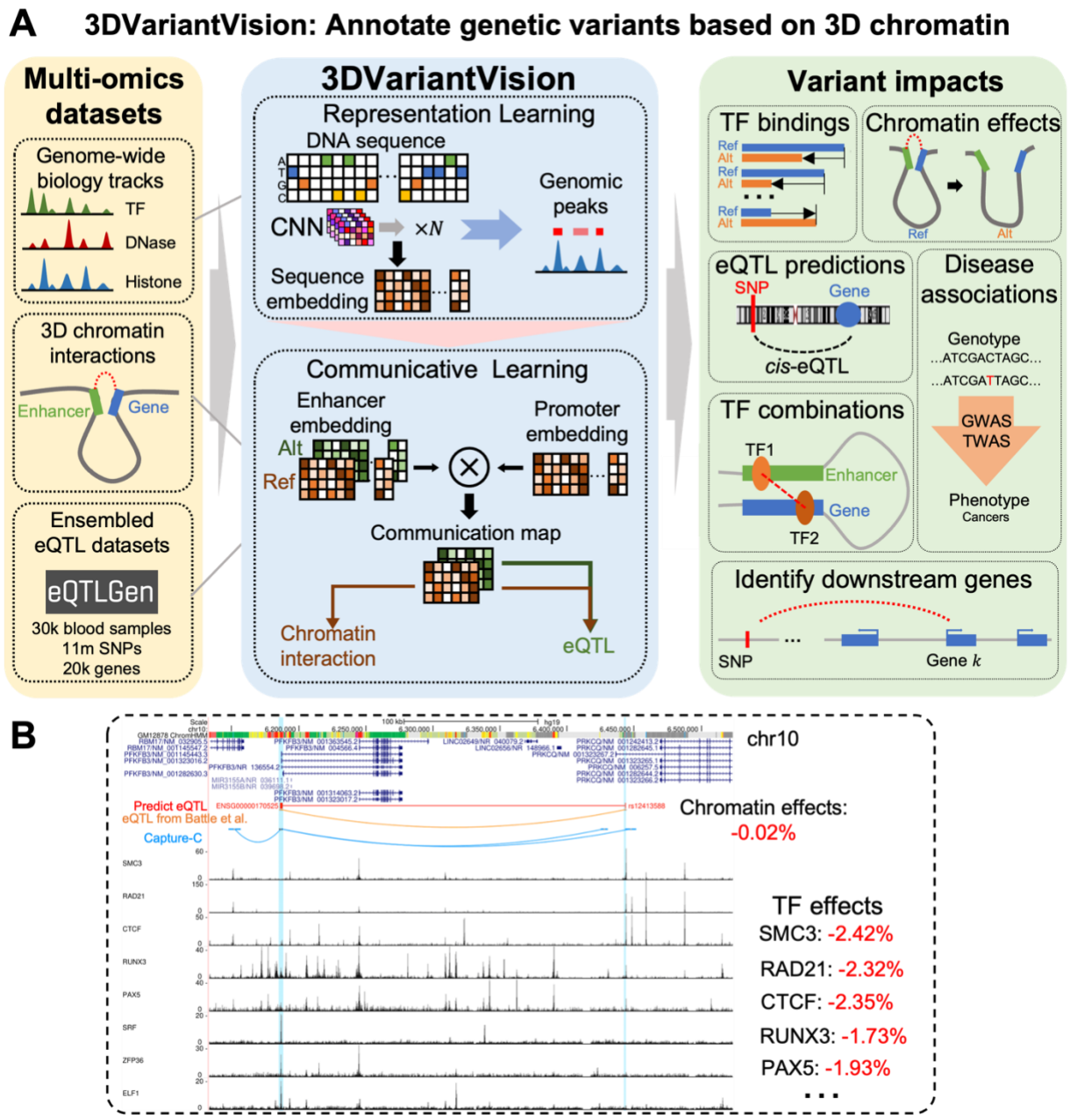

Figure 3.2. Overview of the 3DVariantVision framework. (**A**) 3DVariantVision mainly utilized 3 multi-omic datasets in training, including epigenomic peaks (TF binding sites, open chromatin, histone modifications), chromatin interaction pairs, and ensembled eQTLs (left). Representation Learning was used to generate embedding for DNA sequences by predicting epigenomic peaks (middle). (**B**) One example of predicted eQTLs based on 3DVariantVision in the whole blood cell line. *rs12413588* is validated by Battle *et al*. eQTL dataset with p-value = 9.54e-14 and also supported by Capture-C chromatin interactions. The right panel shows the predicted effects of chromatin interaction effects and represents TF bindings of the corresponding SNPs.

3.3.2 Superior performance on 1D genome profile and long-range chromatin prediction

3DVariantVision significantly outperforms the previous state-of-the-art model, DeepSEA[69], in the prediction of 1D genome profiles, underscoring its proficiency in generating DNA sequence embeddings during the initial phase. To ensure a rigorous comparison, both 3DVariantVision and DeepSEA were meticulously trained and evaluated using the same balanced dataset. 3DVariantVision consistently exhibits higher AUROC and AUPR scores in comparison to DeepSEA across most predictions for TF ChIP-seq peaks and Histone ChIP-seq peaks (**Figure 3.3 A-B**, **Figure B. 3**). This heightened prediction accuracy is further affirmed through qualitative assessments when visualizing the predicted peaks alongside observed ones within the genome. For instance, consider a representative case on 1Mb region in chromosome 17 (**Figure 3.3 C**). The peaks predicted by 3DVariantVision not only align with peaks identified by traditional peak calling methods at significant TF binding sites and histone modifications but also unearth previously undiscovered peaks. These novel peaks, although overlooked by traditional peak calling methods, find support in raw experimental signals, underscoring the enhanced predictive capabilities of 3DVariantVision.

Beyond its proficiency in predicting local genome profiles, 3DVariantVision excels in accurately forecasting enhancer-promoter interactions, shedding light on the intricate distal 3D chromatin mechanisms in stage 2. We gathered enhancer-promoter interactions from Capture-C data specific to the GM12878 cell type and created a balanced negative dataset while controlling for genomic distance. To mitigate the risk of overfitting, 3DVariantVision underwent training and testing on distinct chromosomes. Anchored by the cross-attention module, a backbone of communicative learning within 3DVariantVision, this module meticulously models the nuanced relationships between minuscule regions within enhancers and promoters. This approach equips the model with the capability to glean joint features shared between these two genomic fragments. To evaluate this aspect, 3DVariantVision's performance was benchmarked against ProTECT[162], another model harnessing protein-protein interaction features to encode joint features within genomic fragments. Both ProTECT and 3DVariantVision exhibit the ability to model these joint features, warranting a head-to-head comparison. In the interest

of fairness and accuracy, ProTECT was subjected to the same dataset for training and testing as 3DVariantVision, with a specific focus on subsets featuring protein-protein interactions (PPIs). The superior AUROC scores achieved by 3DVariantVision in chromatin interaction prediction, compared to ProTECT, underscore its enhanced performance (**Figure 3.3 D**).

To further assess 3DVariantVision's prowess in learning joint features within the communicative learning framework, we turned to enhancer-promoter interactions supported by Capture-C. To ensure consistency in enhancer and promoter information while modifying joint features, we generated background pairs by shuffling the corresponding enhancer-promoter pairs. We computed predictions using both ProTECT and 3DVariantVision, followed by the calculation of odds ratios at various thresholds. 3DVariantVision consistently exhibited higher odds ratios, affirming its superior ability to capture joint information between enhancers and promoters (**Figure 3.3 E**, **Figure B. 4**). This robust performance ensures the model's capacity to decipher intricate chromatin mechanisms within minuscule regions, such as those proximal to SNPs, and their influence on target genes.

In a compelling example from the testing set, 3DVariantVision adeptly predicted a long-range enhancer-promoter interaction within chromosome 4, a prediction that was subsequently validated by Capture-C experiments. This interaction was found to be associated with crucial TF binding sites, including HCFC1, SIX5, EED, YY1, and RAD21 (**Figure 3.3 F**). Impressively, 3DVariantVision went beyond validation and unearthed novel interactions that had eluded detection by Capture-C. For instance, in chromosome 8, our model effectively uncovered an enhancer-promoter interaction corroborated by Hi-C, an independent experimental dataset within the same cell type. This interaction featured the binding of BATF, SPI1, RUNX3, YY1, TAF1, and ELF1(**Figure 3.3 G**). Additionally, 3DVariantVision successfully unveiled enhancer-promoter interactions, supported by Hi-C data, in chromosome 5 and chromosome 21, all of which had been previously missed by Capture-C. These newly discovered interactions were also found to be associated with crucial TF binding sites (**Figure B. 5**). These remarkable findings underscore 3DVariantVision's capacity to not only validate existing chromatin interactions

but also to uncover previously unrecognized interactions, thereby enhancing our understanding of the complex 3D chromatin architecture.

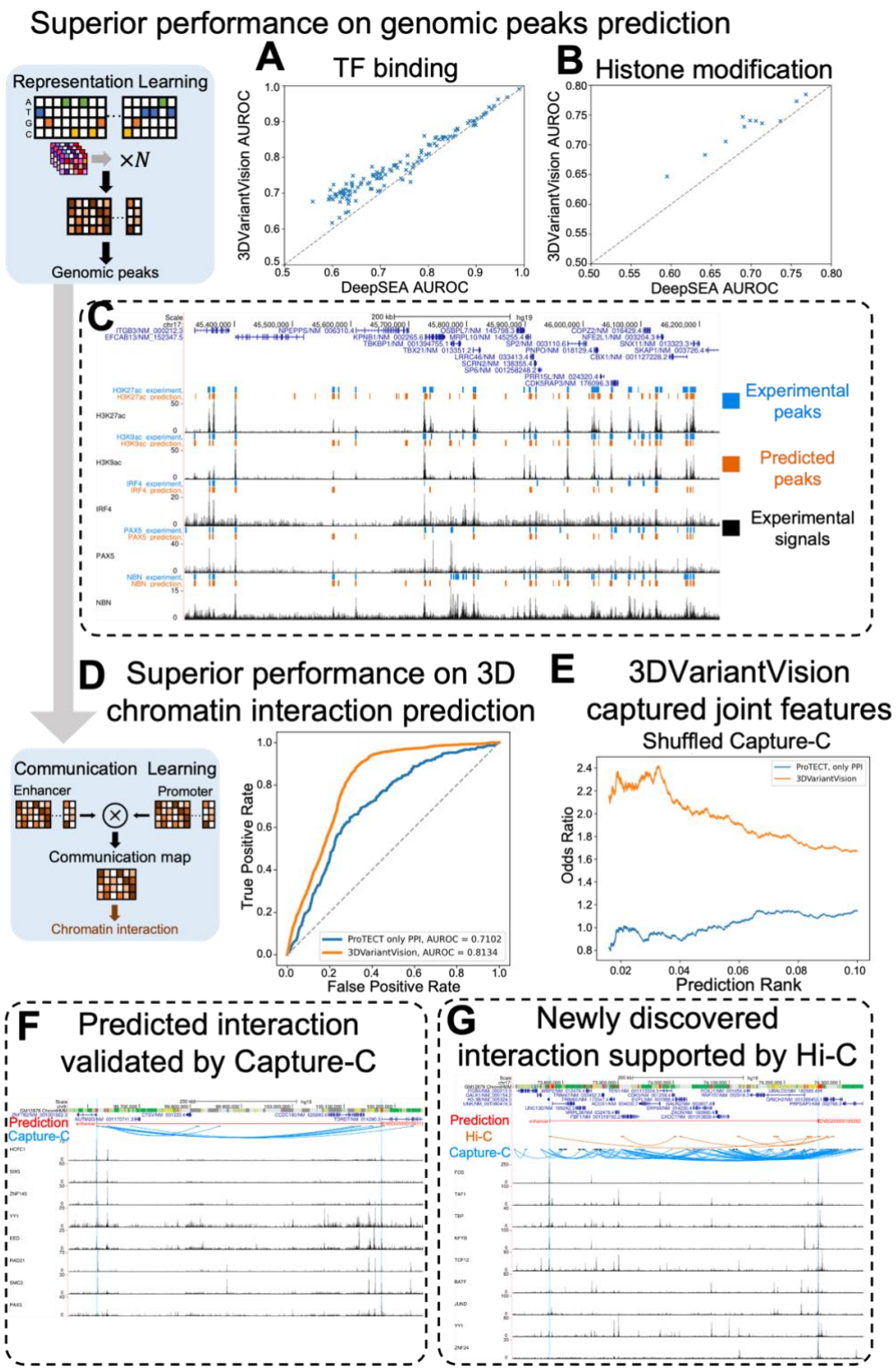

Figure 3.3. 3DVariantVision achieved superior performance in TF binding sites (**A**) and histone modification (**B**) predictions compared with the previous state-of-the-art model DeepSEA. Each point represents the AUROC of 3DVariantVision and DeepSEA in one

Figure 3.3 (cont'd)

epigenomic peak prediction. (**C**) Examples of epigenomic peak predictions in chr17:45,314,322-46,234,870. 3DVariantVision successfully predicted the annotated peaks and discovered the novel peaks supported by the signals. (**D**) ROC comparison of chromatin interaction prediction between 3DVariantVision and ProTECT. (**E**) Odds ratio comparison at top 10% predictions when shuffling the pairs of enhancer-promoter interactions. (**F**-**G**) Examples of predicted chromatin interaction by 3DVariantVision. (**F**) is validated by Capture-C, while (**G**) is a newly discovered interaction supported by Hi-C but missed by Capture-C.

3.3.3 3DVariantVision prioritizes casual SNP on eQTL prediction

In the final phase of our model, we shift our focus towards eQTL prediction, harnessing insights accumulated from the earlier stages of representation learning and communicative learning. These critical insights are the result of a meticulous process that incorporates 1D epigenomic signals and 3D chromatin interactions (**Figure 3.4 A**). For each SNP-gene pair, two corresponding enhancer-promoter pairs based on the reference genome and alternative genome are generated. Concurrently, communicative maps, hinging on these two enhancer-promoter pairs, are meticulously calculated. These maps, crafted through communicative learning trained during stage 2, encapsulate the distinctive features of each SNP-gene pair, serving as a foundation for our eQTL prediction framework. The construction of balanced training, validation, and testing sets is achieved through the large-scale eQTL dataset sourced from blood tissues. To quantify our eQTL prediction performance, we rely on the AUROC metric in the testing dataset. In a rigorous comparative analysis, we pitted the performance of 3DVariantVision against Enformer[75], a cutting-edge model adept at gene expression prediction, leveraging long-range interactions through self-attention mechanisms. Notably, 3DVariantVision achieved a commendable AUROC compared to Enformer, thereby underscoring its superiority in eQTL prediction (**Figure 3.4 A**). Furthermore, we conducted a comprehensive performance evaluation by comparing 3DVariantVision and Enformer using four independent fine-mapped eQTL datasets as gold standards. This orthogonal assessment aimed to provide additional insights into the predictive capabilities of both models across various scenarios. In each of these datasets – SuSIE[166], DAP-G[167], CAVIAR[168], and CaVEMaN[169] - 3DVariantVision consistently outperformed Enformer in terms of AUROC

(**Figure 3.4 B**). These consistently higher AUROC values for 3DVariantVision across all datasets underscore its robust and superior performance in eQTL prediction, regardless of the dataset characteristics and complexities. These results further solidify the efficacy of 3DVariantVision as a state-of-the-art tool for eQTL prediction.

While chromatin effects and expression effects share a related connection, it's imperative to distinguish between these two distinct molecular phenomena. This distinction underscores the necessity of the fine-tuning procedure in Stage 3 of 3DVariantVision, where we transition from capturing chromatin interaction information to making accurate eQTL predictions. Notably, some existing predictive models are trained solely on reference genomes to forecast chromatin interactions and subsequently derive eQTL effects by contrasting alternative genomes with reference genomes. However, this direct approach may struggle to bridge the gap between chromatin-level information and expression-level outcomes. To substantiate this notion, we embarked on an experiment to assess the feasibility of employing chromatin interaction changes from Stage 2 of 3DVariantVision as prediction scores for eQTLs. Utilizing SuSIE fine-mapped eQTLs as gold standards, we observed an AUROC of 0.5606, which falls short of the AUROC achieved in Stage 3 of 3DVariantVision (0.7605) (**Figure B. 6**). This observation underscores that understanding the intricate dynamics of genetic variants' effects on gene expression necessitates a dedicated model fine-tuned for eQTL prediction. Such a model can meticulously consider the complex interplay between genetic variations, chromatin structures, and gene expression patterns. This approach allows us to navigate the intricate relationships between these factors, ultimately culminating in heightened accuracy for eQTL predictions.

Traditional statistical eQTL calling methods, such as GTEx[160], face inherent limitations stemming from their restricted sample sizes and incomplete understanding of underlying mechanisms. These constraints render the identification of long-distance and rare eQTLs challenging. To underscore the advantages of 3DVariantVision in eQTL prediction, we conducted a comparative analysis between 3DVariantVision and GTEx. Our comparision focuses on three distinct subsets: 1) 'Recalled_by_3DVariantVision' representing eQTLs successfully discovered by 3DVariantVision but missed by GTEx. 2) 'TP_by_GTEx' representing true positives according to GTEx's predictions. 3) 'All_eQTL

representing the entirety of gold-standard eQTLs used as the benchmark. Our findings reveal several compelling insights. Firstly, 3DVariantVision excels at recalling long-range eQTLs that eluded GTEx (**Figure 3.4 C**). Secondly, 3DVariantVision demonstrates a knack for pinpointing rare eQTLs missed by GTEx (**Figure 3.4 D**). These observations underscore the superior capacity of 3DVariantVision to unveil elusive eQTLs, particularly those located at extended genomic distances and characterized by low allele frequencies, thereby enhancing our understanding of the regulatory landscape underlying gene expression. Illustratively, consider rs17824742, predicted as an eQTL affecting gene ENSG00000100577, an instance successfully recalled by 3DVariantVision but overlooked by GTEx (**Figure 3.4 E**). Notably, this serves as a paradigmatic long-range eQTL, characterized by a Minor Allele Frequency (MAF) of 0.09 and an extensive genomic separation spanning 233kb. What lends further credence to this prediction is its substantiation through chromatin interactions gleaned from both Hi-C and Capture-C data, shedding light on the potential regulatory mechanisms underpinning this eQTL.

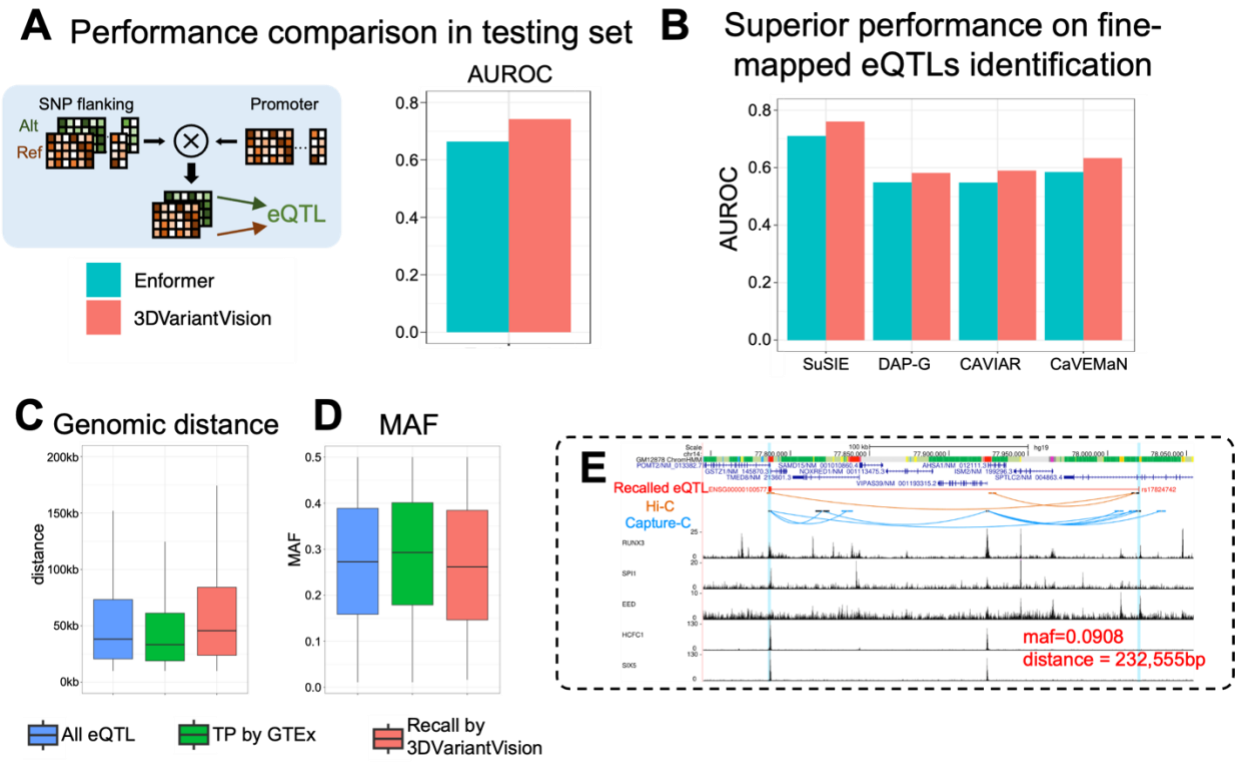

Figure 3.4. 3DVariantVision prioritizes casual SNP on eQTL prediction. (**A**) 3DVariantVision achieved better performance in eQTL prediction at the testing set

Figure 3.4 (cont'd)

compared with Enformer. (**B**) Consistently better performance in fine-mapped eQTL predictions based on three different datasets. (**C**-**D**) 3DVaraintVision also recalled eQTLs which were missed by GTEx, the statistical method in a smaller population. Recalled eQTLs have longer genomic distances (**C**) and smaller MAF (**D**). (**E**) One example of recalled long-range eQTL supported by chromatin interactions with MAF < 0.1.

3.3.4 3DVariantVision identifies TF effects and TF grammar in chromatin

The multifaceted 3DVariantVision framework adeptly encompasses local genomic and epigenomic profiles, long-range chromatin interactions, and eQTL predictions, providing a comprehensive outlook on genetic variants that spans a wide spectrum of insights. To gauge the local impacts, we utilize the TF Disturbing Score, which quantitatively measures the perturbation of genetic signals by computing the difference in TF profile prediction scores between the reference genome and the alternative genome. To validate the efficacy of the TF Disturbing Score, we devise an unsupervised algorithm aimed at genetic variant effect prediction, leveraging data from Massively Parallel Reporter Assays (MPRA)[170]. MPRA experiments grant us insights into the effects of genetic variations at a single-base pair resolution across ten enhancer and ten promoter regions. Employing the Principal Component 1 (PC1) of TF Disturbing Score predictions derived directly from 3DVariantVision, without further training, yielded commendable results in the prediction of genetic variant effects. To illustrate this effectiveness, consider an enhancer region in IRF4. 3DVariantVision exhibited a notably high Pearson correlation (r=0.295) with MPRA experimental effects (**Figure 3.5 A**). Impressively, 3DVariantVision adeptly pinpointed two loci with significant scores, which were subsequently validated as disrupting the binding motifs for two crucial TFs, namely, RUNX1 and NR2F1. These observations underscore the capacity of 3DVariantVision to faithfully capture the effects of genetic variants without the need for additional training or refinement, further demonstrating its utility in genetic variant effect prediction.

Furthermore, we devise a classification task aimed at identifying high-impact genetic variants. We select the topN-effect variants as the positive set based on MPRA-derived p-values and effect scores. To create a balanced background set for each positive set, we employ random sampling. We experiment with various values of N, specifically

200, 500, and 1000, constructing multiple datasets for enhancer and promoter variants separately. We employ a RandomForest classifier, utilizing TF Disturbing Scores generated by 3DVariantVision for each variant. Remarkably, our observations reveal high AUROC scores during testing when identifying top-effect variants in both promoters and enhancers. Specifically, we attain AUROC scores of 0.7199, 0.6743, and 0.6484 when considering the top 200, 500, and 1000 affected promoter variants (**Figure 3.5 B),** and AUROC scores of 0.8393, 0.7974, and 0.7499 when evaluating the top 200, 500, and 1000 affected enhancer variants (**Figure 3.5 C**). These results underscore 3DVariantVision's proficiency in discerning high-impact genetic variants with precision and confidence.

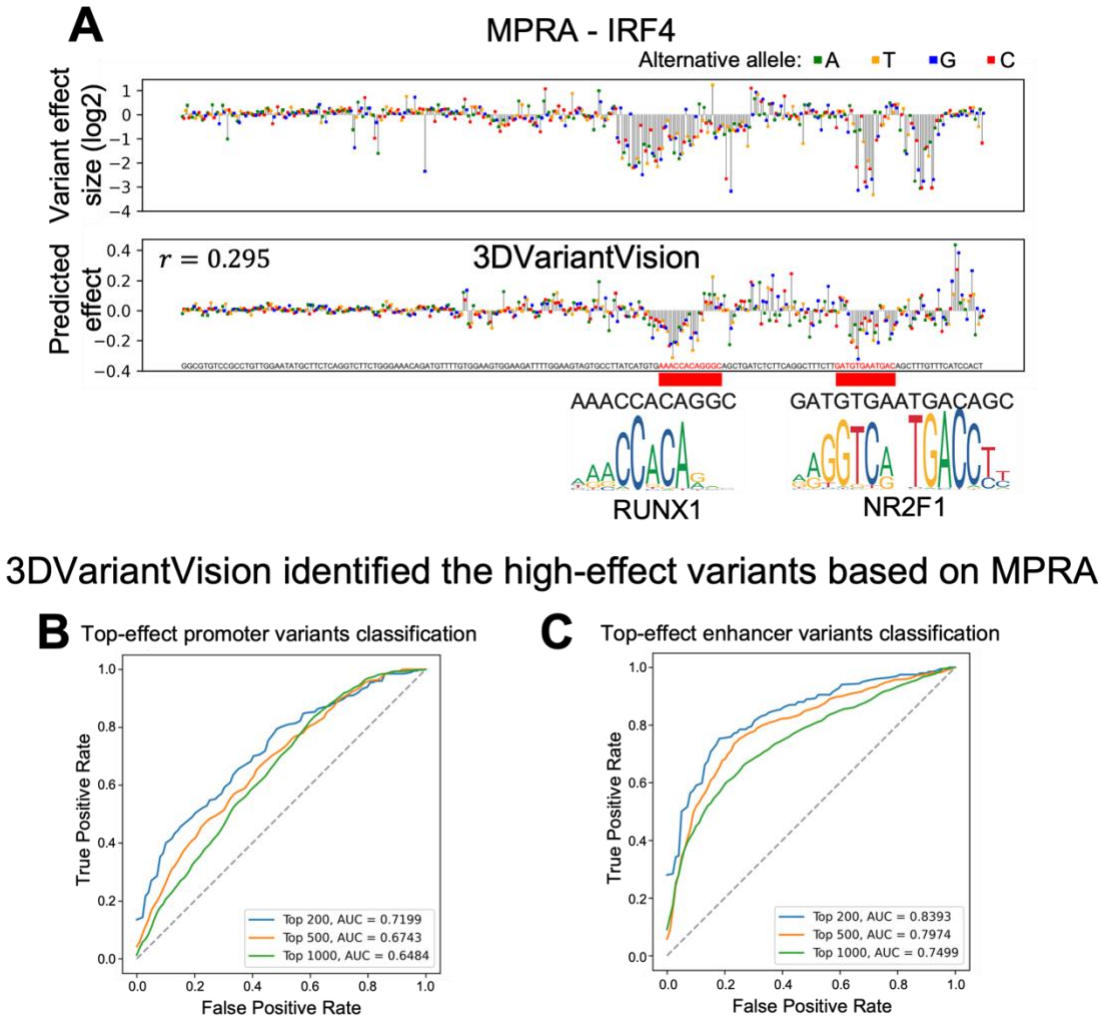

Figure 3.5. 3DVariantVision identified the high-effect variants based on MPRA. (**A**) One example showed the accurate prediction of variant effects based on 3DVariantVision using unsupervised learning. The top panel represents the variant effects in IRF4 loci based on MPRA, while the bottom panel represents the predicted effects based on the PC1 of TF disrupting values of 3DVariantVision. The Spearman correlation between the predicted effects and observed effects is 0.295. High variant effects were observed on the TF binding sites (highlighted). (**B**-**C**) A Random Forest model was trained based on TF disrupting values from 3DVariantVision to predict high-effect variants from the balanced background. The ROC curves showed that the 3DVariantVision's ability to identify the top effect variants against the background in enhancer regions (**B**) and promoter regions (**C**), respectively.

3.3.5 3DVariantVision prioritized the SNP within CRISPRi-perturbed enhancers and the target gene of fine-mapped eQTLs

3DVariantVision's outstanding capability in *de novo* eQTL prediction opens up new avenues for prioritizing causal SNPs among the vast pool of nearby SNP candidates, a challenge previously hindered by LD in GWAS and traditional QTL calling methods. When multiple SNP candidates in enhancer are linked to the same target gene, 3DVariantVision calculates eQTL prediction scores, ultimately prioritizing the SNP-gene pair with the highest score (**Figure 3.6 A**). The validity of this approach is demonstrated using CRISPR-QTL datasets[165], which are not constrained by LD and offer insights into potential causal genetic variants. CRISPR-QTL mapping is akin to conventional human eQTL studies, with individual humans replaced by individual cells, genetic variants substituted by unique combinations of 'unlinked' guide RNA (gRNA)-programmed perturbations per cell, and tissue-level RNA-seq of many individuals replaced by scRNA-seq of many cells. We applied both Enformer and 3DVariantVision to prioritize causal SNPs among CRISPRi-perturbed enhancers. To establish a baseline, we generated a background dataset for CRISPR-QTLs by assembling random enhancer-gene pairs with matching enhancer lengths and genomic distance distributions. eQTL prediction scores were calculated for the prioritized SNP-gene pairs in both CRISPR-QTLs and the background, using 3DVariantVision and Enformer separately. 3DVariantVision predicted a higher scores of CRISPR-QTL compared to scores of the background (**Figure 3.6 B**). In contrast, there is no significant difference between CRISPR-QTL and background based on Enformer predictions (**Figure 3.6 B**). These results highlight 3DVariantVision's significant discrimination in CRISPR-QTL compared to Enformer. Similar comparisons were carried out for enhancer-gene interactions (ABC-EPI) predicted by the Activity-by-Contact (ABC) model[13], where 3DVariantVision consistently exhibited greater discrimination than Enformer. 3DVariantVision predicted a higher score of ABC dataset compared to scores of the background (**Figure 3.6 C**). In contrast, there is no significant difference between ABC dataset and background based on Enformer predictions (**Figure 3.6 C**).

By adeptly modeling the joint features between enhancers and promoters, 3DVariantVision not only excels at predicting genetic variant effects in the local genomic

region but also possesses the capability to predict downstream target genes accurately. For any SNP of interest, 3DVariantVision generates eQTL prediction scores for all candidate genes paired with that SNP. The gene with the highest prediction score is prioritized as the target gene (**Figure 3.6 D**). To validate 3DVariantVision's proficiency in target gene identification, we employed four distinct fine-mapping datasets: SuSIE[166], DAP-G[167], CAVIAR[168], and CaVEMaN[169]. Balanced control backgrounds were meticulously generated, ensuring that each SNP matched the nearest gene to the actual target genes of eQTLs. The resulting AUROC values for these fine-mapped eQTLs are as follows: 0.7410 for SuSIE, 0.7173 for DAP-G, 0.7200 for CAVIAR, and 0.7679 for CaVEMaN (**Figure 3.6 E**), underscoring the superior performance of 3DVariantVision in target gene identification. To further enhance stringency, we curated a more selective background by exclusively choosing the nearest coding genes of the actual target genes of eQTLs. Remarkably, 3DVariantVision continues to exhibit commendable accuracy across all four fine-mapped eQTL datasets, with AUROC values of 0.5529, 0.5346, 0.5310, and 0.5870, respectively (**Figure B. 9**). These results affirm the precision and reliability of 3DVariantVision in target gene identification.

For instance, consider the eQTL rs6985508, where 3DVariantVision successfully pinpointed ENSG00000022567 as the target gene, distinguishing it from the nearby ENSG00000204882 (**Figure 3.6 F**). This eQTL is further supported by chromatin interactions from Capture-C data, and the target gene is bounded by pivotal TFs such as RUNX3, ELF1, HCFC1, NRF1, and MAZ. Additionally, in the case of eQTL rs1545837, 3DVariantVision precisely identified ENSG00000147439 as the target gene, effectively discriminating it from the nearby ENSG00000179388 (**Figure 3.6 G**). Notably, both the target gene and the background gene are associated with important TFs. However, a crucial distinction emerges - the actual target gene is bound by ELF1, while the background gene is not. Intriguingly, the occurrence of RUNX3 in the SNP region implies a potential mechanism of ELF1 repression, as RUNX3 has been proven to repress ELF1 in the CD8-TC cell line[172]. These two illustrative examples underscore 3DVariantVision's remarkable proficiency in identifying the target genes of genetic variants, shedding light on the intricate interplay between genetic variations, TFs, and gene expression regulation.

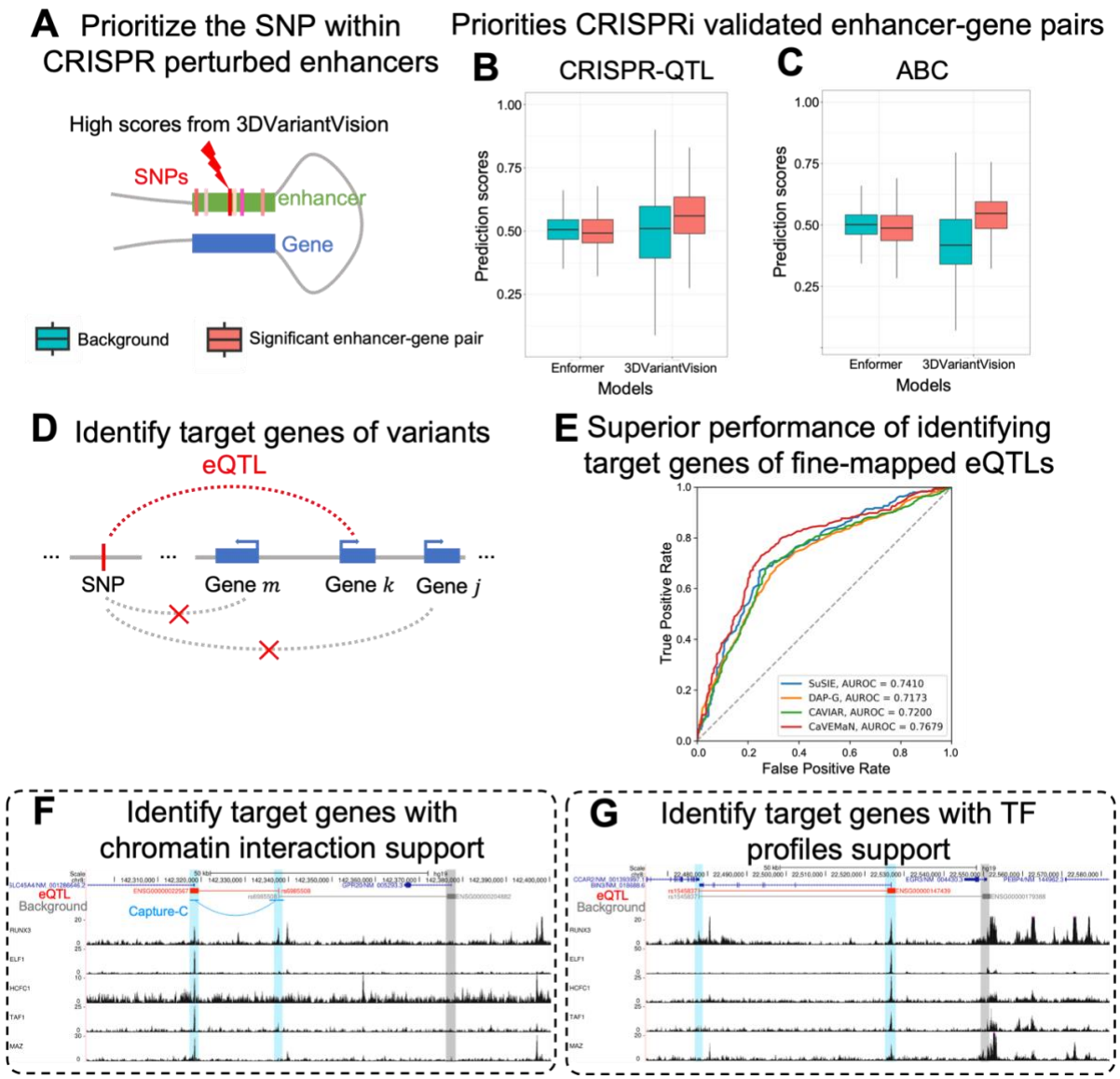

Figure 3.6. 3DVariantVision prioritized the SNP within CRISPRi-perturbed enhancers and the target gene of fine-mapped eQTLs. (**A**) Schematic figure of validation by CRISPR-QTL. The eQTL prediction scores of all the potential SNP-gene pairs located in the *cis* enhancer-gene pair are calculated. The SNP-gene pair with the highest prediction score is prioritized to represent the score of the corresponding CRISPR-QTL. (**B-C**) 3DVariantVision showed significantly higher prediction scores on two orthogonal datasets, CRISPR-QTL (**B**) and ABC (**C**), compared with genomic distance-controlled SNP-gene pairs. (**D**-**E**) 3DVariantVision prioritized the downstream target genes of genetic variants. (**F**) For the SNPs with interests, 3DVariantVision calculated the eQTL prediction scores with surrounding genes and prioritized the top effected one. (**E**) 3DVariantVision successfully identified target genes of 4 fine-mapped eQTL datasets from the closest gene to the real target genes with distance > 1kb. (**F**-**G**) Two eQTL examples of target gene identifications supported by chromatin interaction (**F**) and TF profiles (**G**).

3.3.6 3DVariantVision deciphers the relationships of effects between chromatin and expression

3DVariantVision excels in modeling long-range chromatin interactions, providing a means to quantify the chromatin effects of SNP-gene pairs by contrasting the reference genome with the alternative genome. While the preceding eQTL prediction section underscores the related yet distinct nature of chromatin interactions and gene expression associations, this section delves into quantifying their relationship by calculating correlations between chromatin effects and expression effects. GTEx supplies the slopes for all significant SNP-gene associations, serving as proxies for expression effects. Chromatin effects, in turn, are determined by assessing the chromatin changing ratio between the alternative and reference genomes, as derived from 3DVariantVision. The Spearman correlation of absolute values between expression effects and chromatin effects yields 0.1759, revealing a positive relationship wherein elevated chromatin effects correspond to increased expression effects. Notably, this correlation is even higher (0.2126) for a subset of SNP-gene associations characterized by genomic distances exceeding 500kb. To comprehensively explore this connection, we segment the significant SNP-gene associations into 20 evenly-distributed groups based on genomic distance. Strikingly, a statistically significant higher correlation emerges in the associations with longer genomic distances (p-value=0.0022) (**Figure 3.7 A**), suggesting that chromatin interaction may mediate long-range eQTLs.

Continuing our investigation, we delve into the directionality of chromatin effects and expression effects. While the absolute values of these effects exhibit positive correlations, the directions reveal a more intricate, non-correlative relationship. We posit that distinct roles of TFs might underlie this phenomenon. Specifically, eQTLs influenced by activator TFs may exhibit consistent directions between chromatin effects and expression effects, whereas those influenced by repressor TFs may demonstrate opposite directions. To put this hypothesis to the test, we segregate significant SNP-gene associations from GTEx into groups based on different TF bindings at the SNP sites, referred to as TF-mediated groups. We assess the fraction of consistent directions between chromatin effects and expression effects for each TF-mediated group based on the top N effect values derived from expression effects. N varies from 20 to 2000. The

direction consistency is plotted against N for each TF-mediated group. At the top-effect eQTLs, some TF-mediated groups indeed exhibit consistent directions, while others manifest opposite directions (**Figure B. 8**). However, the consistency tends to approach neutrality when the expression effects of eQTLs are less pronounced. To glean insights into the roles of TFs, whether they activate or repress, we employ the consistency at the top 100 expression effects as a representation (**Figure 3.7 B**). Notably, all TF-mediated groups exhibit positive Spearman correlations between the absolute values of chromatin effects and expression effects. Two well-established examples, EBF1 as an activator[173] and TBX21 as a repressor[174], exemplify consistent and opposite directions between chromatin effects and expression effects, respectively (**Figure 3.7 B**). These results underscore 3DVariantVision's capacity to unveil the roles of TFs in both chromatin and expression regulation.

Communicative Learning effectively modeled the intricate relationships among minuscule regions situated between enhancers and promoters, allowing for the prioritization of vital functional TF combinations that contribute to chromatin interactions. To assess the significance of these TF combinations, we focused on Capture-C supported enhancer-promoter interactions in heldout chromosomes (chr9, chr10, and chr11). We counted the occurrences of TF combinations within the top-priority minuscule region pairs using TF ChIP-seq peaks, and we created corresponding backgrounds. Subsequently, we calculated z-scores for each TF combination and visualized them as a heatmap (**Figure B. 7**). Notably, some of these enriched TF combinations were validated by Protein-Protein Interactions (PPIs)[139], such as SIN3A-ZBTB40, YY1-JUNB, CHD1-SRF, and STAT3-ARNT, highlighting the role of PPIs in chromatin interactions (**Figure 3.7 C**). To comprehensively assess the significance of the TF combinations prioritized by 3DVariantVision, we conducted PPI enrichment analysis by comparing them against shuffled backgrounds. This analysis revealed a statistically significant enrichment of PPIs among the prioritized TF combinations, as confirmed by the z-test (p-value = 1.03e-2) (**Figure 3.7 D**). These findings collectively underscore 3DVariantVision's capacity to discover and prioritize crucial TF pairs essential for chromatin interactions.

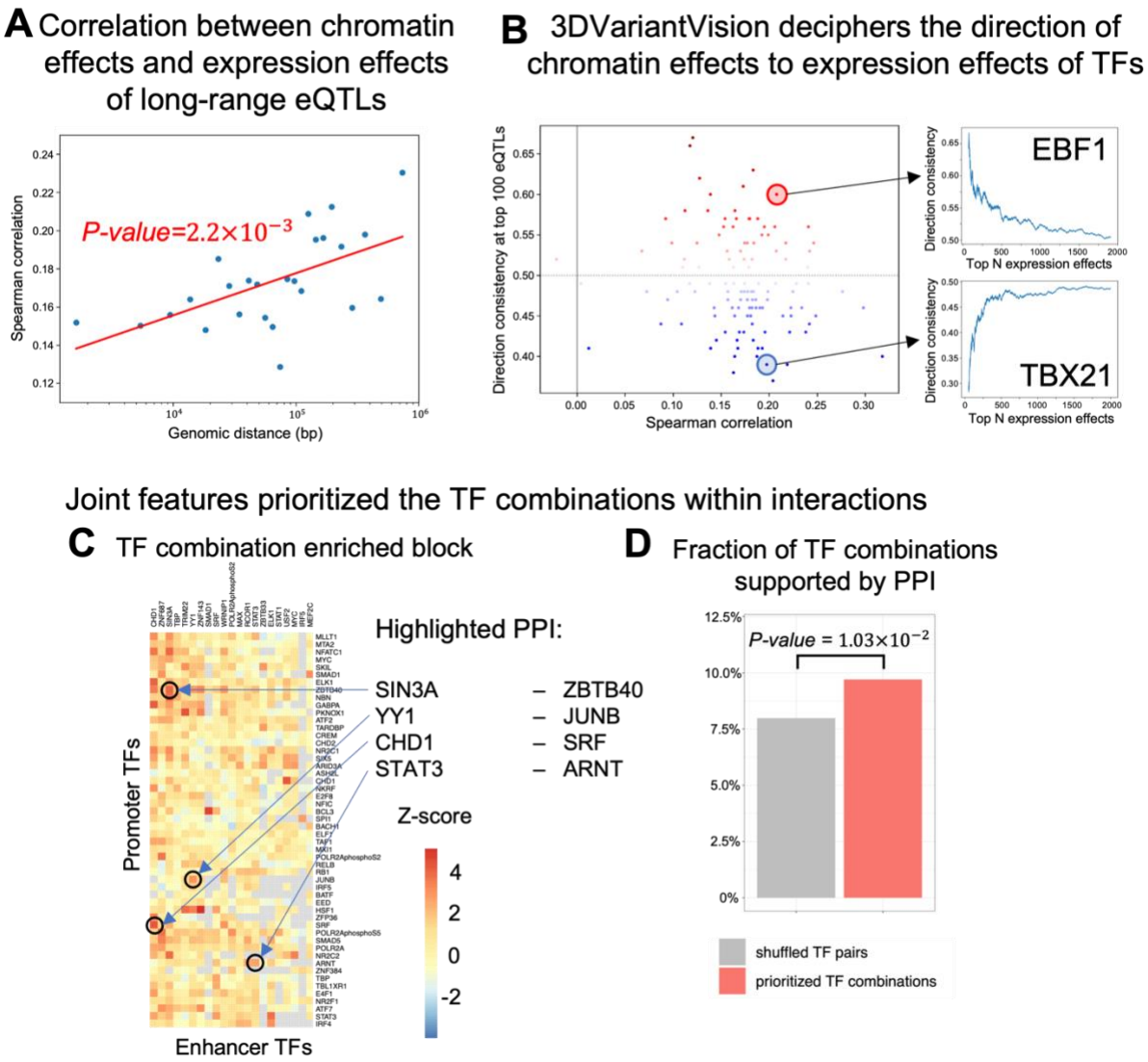

Figure 3.7. 3DVariantVision deciphers the relationships of effects between chromatin and expression. (**A**-**B**) 3DVariantVision deciphered the chromatin effects of eQTLs. (**A**) The eQTLs were divided into 20 groups based on genomic distance and the Spearman correlation between chromatin effects and expression effect sizes was calculated separately. The eQTLs with longer genomic distances showed a higher correlation between effect sizes and chromatin effects. (**B**) The eQTLs were grouped by different harboring TFs located near the SNPs. The direction consistency on the top 100 expression effected eQTLs was calculated by the fraction of the consistent direction between expression effects and chromatin effects. The red dots in the scatter plot represent TFs with directions while the blue dots represent TFs with opposite directions. All of them have a positive Spearman correlation between chromatin effect sizes and expression effect sizes. The right panel highlights two TFs with high consistency and high opposite directions. (**C**-**D**) 3DVariantVision prioritized the TF combinations of chromatin interactions. For the minuscule regions of the Capture-C enhancer-promoter interactions

Figure 3.7 (cont'd)

in chr9, chr10, chr11, the harboring TF combinations were counted based on TF ChIP-seq. The Z-score was calculated for each TF combination by comparing the top important minuscule region pairs based on the attention weights of 3DVariantVision and the random minuscule region pairs from the same enhancer-promoter interaction. The enriched TF combination block was shown in (**D**). (**E**) PPIs are enriched on the prioritized TF combinations (p-value of Z-score < 0.05), compared with the randomly shuffled TF pairs.

3.3.7 3DVariantVision helps to understand the disease associations

Understanding Genome-Wide Association Studies (GWAS)[3] and Transcriptome-Wide Association Studies (TWAS)[157] is of paramount importance in unraveling the genetic underpinnings of complex diseases and advancing personalized medicine. To assess 3DVariantVision's capacity to elucidate the implications of genetic variants for disease associations within the context of GWAS, we categorized GTEx eQTLs into two distinct groups based on whether the associated SNPs were corroborated by GWAS findings. Notably, for the eQTL group supported by GWAS, 3DVariantVision exhibited markedly higher eQTL prediction scores in comparison to the group without GWAS support (**Figure 3.8 A**). Furthermore, TWAS aims to identify genes whose expression levels are influenced by genetic variants. To comprehensively assess 3DVariantVision's potential in prioritizing target genes associated with TWAS, we focused on a set of 28 statistically significant SNP-gene associations linked to lipid traits in whole blood cell type[171]. For each of these associations, we generated a corresponding background set by aggregating all coding genes located within a 1Mb region centered on the TSS of the TWAS-targeted gene. Subsequently, we calculated the rank percentile of the TWAS target gene within this background, utilizing the 3DVariantVision prediction score. Remarkably, our results revealed that the TWAS target genes consistently obtained significantly higher rank percentiles compared to all the coding genes in the 1Mb background (**Figure 3.8 B**). Those observations underscore the potential of 3DVariantVision to shed light on the intricate relationship between genetic variants, gene expression, and disease associations, offering valuable insights into the genetic basis of complex diseases and the development of precision medicine approaches.

In this demonstration, we present notable examples of successful target gene prioritization for four distinct traits - high-density lipoprotein cholesterol (HDL-C), low-

density lipoprotein cholesterol (LDL-C), total cholesterol (TC), and triglycerides (TG). Consider the genetic variant rs12050262, associated with the HDL-C trait. Remarkably, 3DVariantVision effectively prioritized SETD3, the genuine target gene, out of a pool of 7 coding genes within the surrounding 1Mb background region (**Figure 3.8 C**). Similarly, for rs17599675, linked to the LDL-C trait, our model accurately pinpointed FBXO38 as the target gene among 11 coding genes within the 1Mb background region (**Figure 3.8 D**). Furthermore, rs11578696, associated with the TC trait, was successfully linked to TMEM79, the actual target gene, from a set of 38 coding genes located within the 1Mb background region (**Figure B. 10 A**). Lastly, for rs10889347, a genetic variant associated with the TG trait, 3DVariantVision precisely prioritized USP1 as the target gene, out of 6 coding genes situated in the 1Mb background region (**Figure B. 10 B**). These compelling examples underscore 3DVariantVision's efficacy in accurately identifying genes associated with specific traits, facilitating a deeper understanding of the genetic basis of complex phenotypes, and offering insights for potential therapeutic interventions.

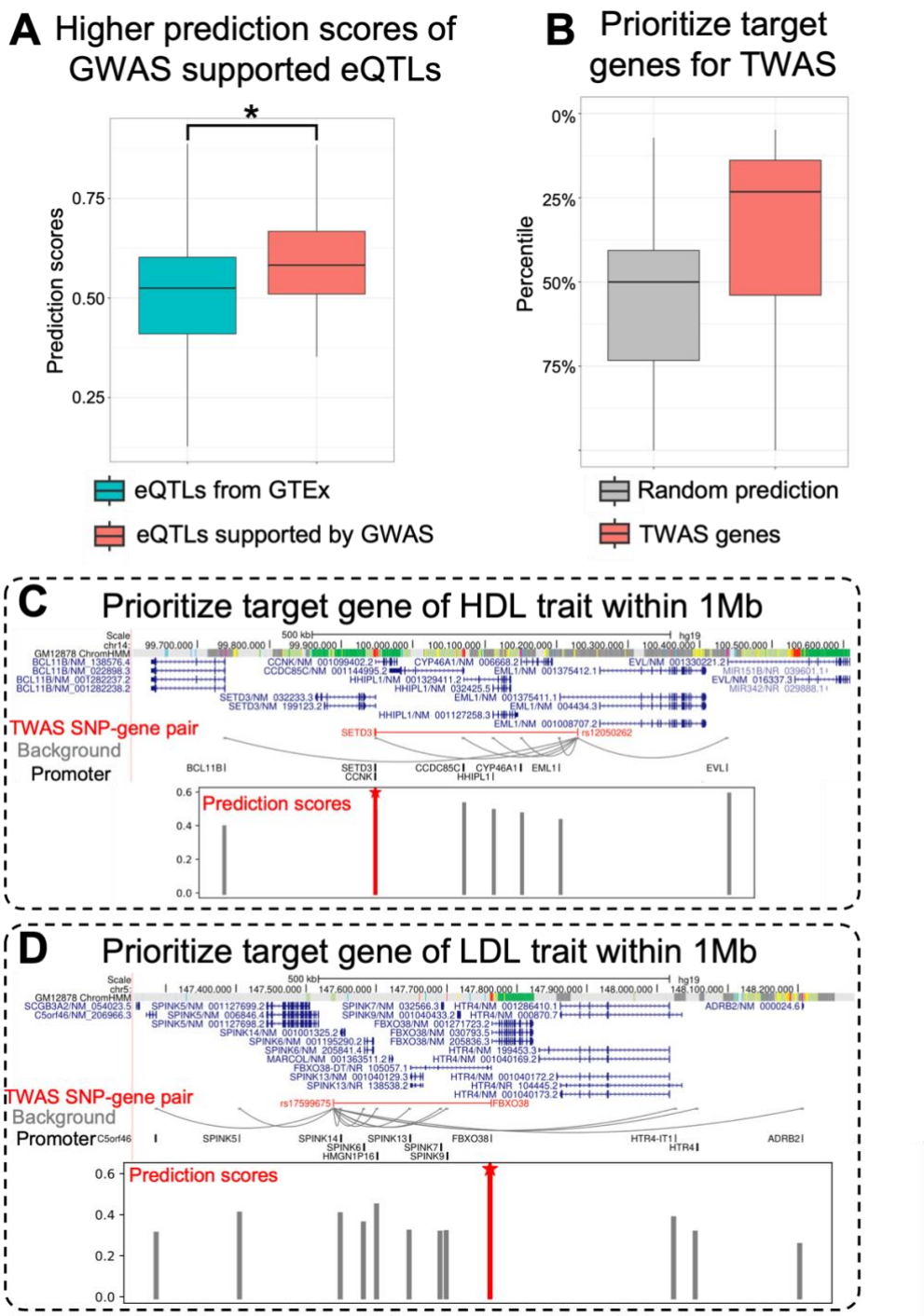

Figure 3.8. 3DVariantVision interprets disease-associated traits. (**A**) eQTLs supported by GWAS showed significantly higher prediction scores based on 3DVariantVision compared with the background eQTLs. (**B**) For 28 TWAS associations with blood cell traits, all the coding genes within 2MB from the TSS of the TWAS targeted gene were collected as the background, and the corresponding eQTL prediction scores were

Figure 3.8 (cont'd)

generated by 3DVariantVision. The y-axis represents the percentile rank based on eQTL prediction scores, showing 3DVariantVision prioritized the TWAS associations against the background. (**C**-**D**) Two examples of prioritizing target genes of TWAS SNPs based on 3DVariantVision within 1Mb. The barplot at the bottom panel shows the eQTL prediction scores. (**C**) and (**D**) are the examples of HDL and LDL trait associations, respectively.

CHAPTER 4
SYSTEMATIC DELINEATION OF MULTI-LEVEL STRUCTURAL VARIABILITIES OF SPATIAL GENOME CONFORMATIONS

4.1 INTRODUCTION

The three-dimensional (3D) organization of chromatin plays a crucial role in regulating gene expression, DNA replication, genome stability, and tissue differentiation[175–177]. In recent years, the rapid development of chromatin conformation capture assays, such as ChIA-PET[107], Capture-C[108], and Hi-C[14], has enabled the quantitative analysis of 3D chromatin interactions, including enhancer-promoter interactions and genome-wide intra-chromosomal and inter-chromosomal contacts across various species and cell types. These studies have revealed structural components of chromatin at multiple scales, including chromatin loops, topologically associating domains (TADs), and chromatin compartments[14]. However, bulk-tissue measurements can only represent the average chromatin conformation, failing to capture the heterogeneity among millions of individual cells (**Figure 4.1 A**). The advent of single-cell chromatin conformation capture methods, such as single-cell Hi-C[30–40], has opened new frontiers in the field, allowing researchers to investigate 3D genome structure at the single-cell level and unveil the dynamics and variability of chromatin organization across cells[30–40].

The analysis of 3D genome structures encompasses both intra-chromosomal and inter-chromosomal interactions. Within a chromosome, researchers investigate chromatin loops, TADs, and chromatin compartments[14]. When studying the inter-chromosomal structure of a diploid cell, researchers examine the relative positions of chromosomes and their spatial organization within the nucleus[41,178–180]. These multi-scale structures, both intra- and inter-chromosomal, are heavily influenced by variability at multiple levels (**Figure 4.1 B**). As cells differentiate from stem cells, cell-type variability leads to the emergence of distinct functions and shapes of different tissues[181]. Even within the same cell type, cell-level variability arises from differences in cell age and individual-level variations[182]. Moreover, within a single diploid cell, the two copies of the same chromosome, one from the paternal genome and the other from the maternal genome, can exhibit structural differences in allele-level, such as the active and inactive states of

chromosome X[183,184]. This multi-level variability poses significant challenges for both computational and experimental analyses of 3D chromatin regulation.

Based on bulk-tissue unphased Hi-C data, which is more widely available compared to single-cell Hi-C, researchers have developed numerous methods for contact normalization[185,186], imputation[187], and interaction calling[188], as well as further structure reconstruction[26–28,189–192]. These methods often employ a 'beads on a string' polymer model, where each chromosome is represented as a chain of 'beads' consisting of DNA fragments or loci, and the pairwise spatial distances between genomic loci are inferred from Hi-C contacts[29]. Structure reconstruction methods can be broadly categorized into two types: consensus methods and ensemble methods (**Figure 4.1 C**). Consensus methods, such as FLAMINGO[28] and GAM-FISH[192], reconstruct a single consensus structure for each chromosome by treating the intra-chromosomal Hi-C contact map as an average representation of pairwise distances between genomic loci without further justification[29]. However, these methods overlook cell-level and allele-level variability[30,182]. In contrast, ensemble methods[25,193] decompose the bulk-tissue Hi-C contact map matrix into a large tensor, with an additional dimension representing individual cells, thereby generating thousands of different structures. While ensemble methods attempt to capture cell-to-cell variability, the resulting structures mix multi-level variability and are challenging to interpret due to the lack of a clear correspondence between the generated structures and the individual cells in the Hi-C experiment[25,193]. Consequently, the strategies and underlying assumptions of these methods require further systematic evaluation to provide guidance for future research.

In this study, we provide a comprehensive analysis of the multi-level variability in 3D chromatin organization by leveraging a wide range of bulk-tissue and single-cell Hi-C datasets[14,34,35] across different cell types, developmental stages, and alleles. Our findings reveal several key biological insights. First, we demonstrate that intra-chromosomal chromatin organization is more stable within the same cell type compared to the dynamics observed across cell types, with the large-scale skeleton being stable at low resolution and the short-scale detailed organization captured at higher resolutions. Second, we show that inter-chromosomal chromatin exhibits a higher degree of complexity and variability that cannot be captured by a single consensus structure, emphasizing the

importance of considering cell-to-cell variability in inter-chromosomal interactions. Third, we uncover an ordered hierarchy of structural variability, with tissue-level variability being more pronounced than age-level variability, followed by individual cell-level and allele-level variability for autosomes, while chromosome X exhibits a distinct order with allele-level variability being more prominent than individual cell-level variability. These findings highlight the importance of considering cell type-specific factors, developmental dynamics, and allele-specific differences when studying 3D chromatin organization and its relationship to gene regulation and cellular function. From a computational perspective, our analysis provides guidance for selecting appropriate resolutions and read depths when studying chromatin organization across different genomic scales, considering the trade-off between resolution and genomic distance. We also emphasize the need for advanced computational methods that can capture the complexity and variability of inter-chromosomal interactions beyond single consensus structures. Furthermore, we propose a framework for modeling and analyzing chromatin structures at different levels of inquiry based on the ordered hierarchy of variability sources. Our study highlights the limitations of existing methods, such as consensus and ensemble approaches, in addressing the multi-level variability in 3D chromatin organization and underscores the need for integrating multiple experimental techniques to remove biases and obtain a more accurate representation of chromatin interactions. In conclusion, our comprehensive analysis of multi-level variability in 3D chromatin organization provides valuable insights into the complex interplay between cell type-specific factors, developmental dynamics, and allele-specific differences, guiding future experimental designs and computational method development in this rapidly evolving field.

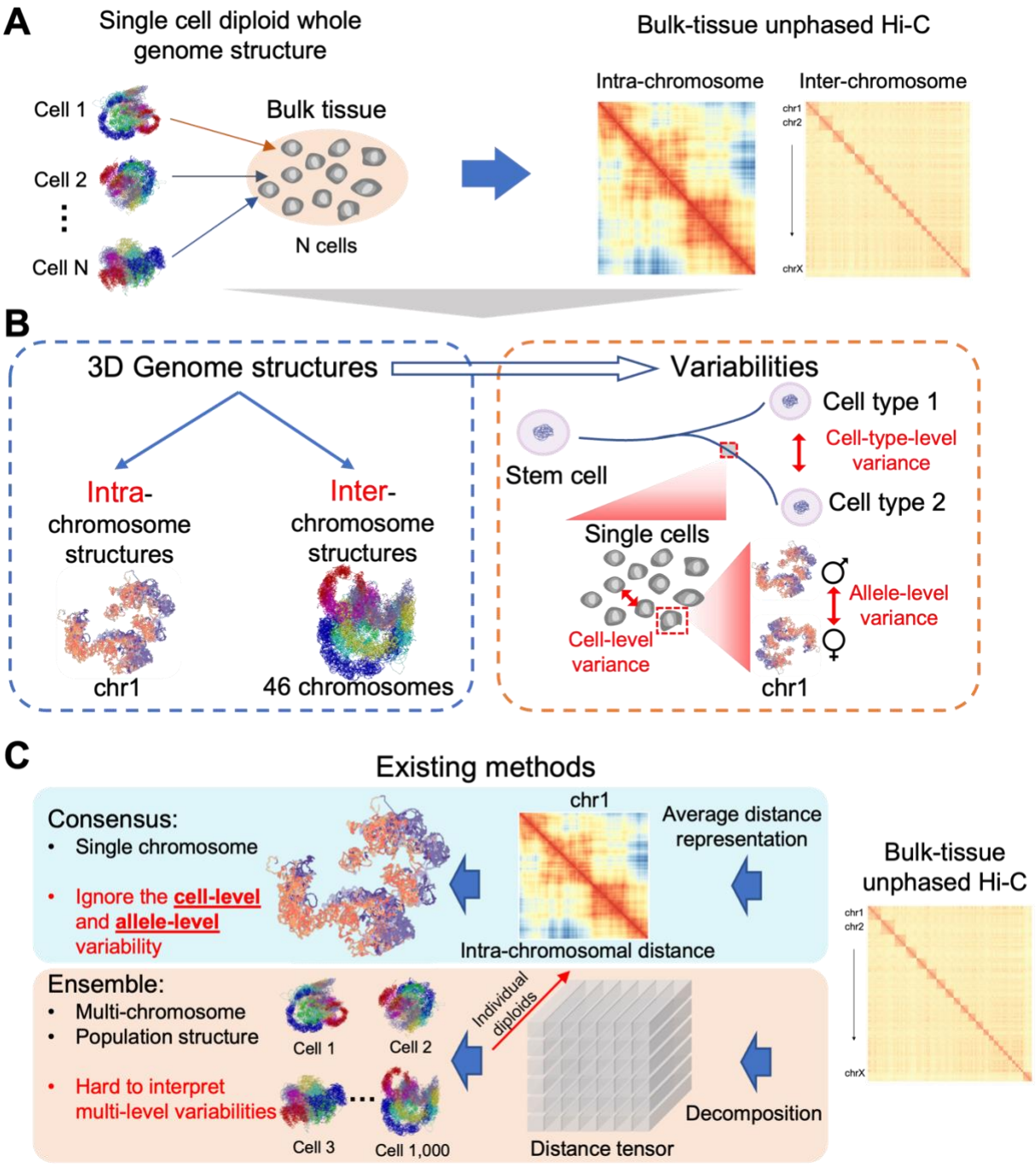

Figure 4.1. Overview of 3D chromatin structure modeling and analysis based on Hi-C. (**A**) Bulk-tissue Hi-C provides an average representation of both intra-chromosomal and inter-chromosomal chromatin organizations from millions of cells. The heatmap shows an example of a bulk-tissue Hi-C contact matrix, with warmer colors indicating higher interaction frequencies between genomic regions. (**B**) Multi-level variability in analyzing 3D genome structures. Intra-chromosomal contacts represent the 3D structure of a single chromosome, while inter-chromosomal contacts represent the interactions between

Figure 4.1 (cont'd)

multiple chromosomes in diploid cells, which contain 23 maternal and 23 paternal alleles. The analysis of 3D genome structure is influenced by variability at different levels, including cell-type level, individual cell level, and allele level. (**C**) Summary of existing methods for 3D genome structure reconstruction. Consensus methods based on bulk-tissue unphased Hi-C data ignore cell-level and allele-level variability to reconstruct the 3D structure of a single chromosome using the intra-chromosomal distance matrix derived from Hi-C. Ensemble methods decompose the distance tensors to generate thousands of multi-chromosomal 3D structures, but the resulting models can be difficult to interpret biologically.

4.2 MATERIALS AND METHODS

4.2.1 Datasets and pre-processing

Hi-C datasets from seven bulk tissues (GM12878, K562, KBM7, HUVEC, IMR90, HMEC, and NHEK) obtained from Rao *et al*[14] were utilized to compare intra-chromosomal 3D chromatin organization across different cell types. For the analysis of inter-chromosomal interactions, both SPRITE[194] and Hi-C data from the GM12878 cell type were employed. Single-cell Hi-C datasets from Dip-C experiments in human (GM12878)[34] and mouse (MOE and retina) cells[35] were used to investigate intra-chromosomal interactions at the single-cell level. All Hi-C datasets underwent mapping and normalization procedures as described in their respective original publications[14,34].

4.2.2 Bulk-level Intra-chromosomal Hi-C cross cell-type analysis

To assess cell-type level variability, Spearman correlation was calculated across different cell types. The Hi-C data was partitioned based on genomic distance bands to control for confounding factors such as missing rate and genomic distance. Let g represent the genomic distance between two genomic fragments. The Hi-C data was divided into the following bands: (0,1Mb], (1Mb,2Mb], (2Mb,5Mb], and (5Mb,10Mb]. For Hi-C data from two cell types, the union set of available data was considered, and Spearman correlation was calculated within each chromosome separately. Various Hi-C resolutions were examined, including 1Mb, 500kb, 250kb, 50kb, 25kb, and 10kb. However, the correlation for $g \in$ (0,1Mb] at 1Mb resolution was not available. Additionally, the correlation for $g <$ 2Mb at 10kb resolution was not computed due to low data quality resulting from a high

missing rate. The primary replicate of GM12878 was used as the query cell type, and Spearman correlation was calculated between the query cell type and the target cell types, which included K562, KBM7, HUVEC, IMR90, HMEC, NHEK, and another replicate of GM12878.

Spearman correlations between the primary replicate of GM12878 and another cell type were calculated for specific chromosomes, resolutions, and genomic distance bands. These correlations from all chromosomes were then grouped based on the corresponding resolution, genomic distance band, and target cell type, representing the overall correlation between GM12878 and the other cell type under specific resolution and genomic distance band settings. The cell type with the highest median correlation was used as the cell-type-specific boundary for a given resolution and genomic distance band, as all median correlations between GM12878 and other cell types fall below this threshold. This cell type is considered the most similar to GM12878 in terms of chromatin organization under specific resolution and genomic distance band settings.

To quantify the influence of read depth on the data, a single replicate of GM12878 was used, and the original mapped reads were downsampled. For a given chromosome and resolution, the corresponding Hi-C matrix was obtained. Original reads were randomly downsampled at the following rates: 80%, 50%, 30%, 20%, 10%, 8%, 5%, 3%, 2%, and 1%. Based on the sampled reads, downsampled Hi-C matrices were constructed. The downsampling process was performed once for each specific chromosome and resolution combination. Spearman correlation between the original Hi-C matrix and the downsampled Hi-C matrices was then calculated separately for different genomic distance bands.

4.2.3 Bulk-level inter-chromosomal Hi-C data normalization

To remove the bias in Hi-C interaction frequencies, two transformation functions were applied between Hi-C and SPRITE data. For the GM12878 cell type, both Hi-C and SPRITE data were collected. All interaction frequencies at 10Mb resolution were divided into intra-chromosomal and inter-chromosomal categories for Hi-C and SPRITE separately. First, for intra-chromosomal interaction frequencies, a linear transformation function $g(x) = \beta_1 x + \beta_0$ was used to convert SPRITE interaction frequencies to Hi-C

interaction frequencies using the equation $IF_{HiC_{intra}} = g(IF_{SPRITE_{intra}})$, where $IF_{HiC_{intra}}$and $IF_{SPRITE_{intra}}$ represent the intra-chromosomal interaction frequencies in Hi-C and SPRITE, respectively, and β_1, β_0 are the learned parameters. Second, for inter-chromosomal interaction frequencies, a log transformation function $f(x) = a_1(\alpha_1\ log\ (x + 1) + \alpha_0) + a_0$ was employed to convert Hi-C interaction frequencies to SPRITE interaction frequencies using the equation $IF_{SPRITE_{inter}} = f(IF_{HiC_{inter}})$, where a_1, a_0, α_1 and α_0 are the learned parameters. The log function was chosen over the linear function due to its better performance. The learned transformation functions were applied to normalize inter-chromosomal Hi-C interaction frequencies by removing the bias using the equation:

$$IF_{HiC_{intra}} = g(f(IF_{HiC_{inter}}))$$

The normalized Hi-C interaction frequencies were transformed into 3D Euclidean distances using the exponential function: $d_{ij} = IF_{ij}{}^{\eta}$, where d_{ij} represents the pairwise 3D distance between DNA fragments i and j, IF_{ij} represents the interaction frequency between fragments i and j, and η is set to 0.25, as suggested by previous literature. A standard distance metric satisfies the triangle inequality: $d_{ij} < d_{ik} + d_{jk}$, which can be used to quantify the quality of spatial distances transformed from inter-chromosomal interaction frequencies. For all genomic fragments in the whole genome at 10Mb resolution, a triplet of fragments i, j, and k was randomly sampled to check whether the triangle inequality holds. This sampling process was repeated 1000 times to generate 1000 triplets, and the fraction of cases in which the triangle inequality held was calculated. The entire procedure was performed 100 times to obtain the distribution of these fractions. A higher fraction indicates a better quality of the distance from Hi-C.

4.2.4 Low-rank property evaluation

The pairwise distance matrix is biologically generated by the underlying low-rank coordinate matrix of DNA fragments with rank ≤ 3. Thus, Due to the property of ranks for matrix addition, the observed Euclidean distance matrix has a rank ≤ 5. Singular value decomposition (SVD) was performed on the observed distance matrix transformed from the Hi-C interaction frequency. The SVD is given by:

$$D = U\Sigma V^T = \sum_i^n \sigma_i u_i {v_i}^T$$

$$|\sigma_1| \geq |\sigma_2| \geq \ldots \geq |\sigma_n|,$$

Here, D represents the observed distance matrix, σ_i are the singular values (SVs), u_i and v_i are the left and right singular vectors, respectively, and n represents the dimension of D. According to the low-rank property of the distance matrix, the top 5 SVs and corresponding vectors are used to reconstruct the approximation of the distance matrix:

$$\widehat{D} = \sum_i^5 \sigma_i u_i {v_i}^T$$

To assess the low-rank property in the observed distance matrix, we first calculated the explanation ratio of the top k SVs:

$$r_k = \frac{\sum_{i=1}^k |\sigma_i|}{\sum_{i=1}^n |\sigma_i|} \in [0,1]$$

The ideal explanation ratio of the top 5 SVs r_5 should be close to 1, indicating that the low-rank property holds well when the ratio is closer to 1. Additionally, we checked the correlation between the observed distance matrix (D) and the approximation matrix ($\widehat{D}$) based on the top 5 SVs. Both Spearman and Pearson correlations were calculated. A higher correlation indicates a better low-rank property, as the approximation matrix closely represents the original observed distance matrix.

4.2.5 Allele-allele correlation in single-cell

Dip-C, a high-throughput single-cell Hi-C technique, provided data for 17 cells in the human GM12878 cell type and approximately 400 cells in mouse, including retina and main olfactory epithelium (MOE) cell types. By leveraging unique single nucleotide polymorphisms (SNPs) in paternal and maternal alleles, the reads from Dip-C were mapped to specific alleles, enabling the construction of allele-specific interaction frequency matrices. Let $A^{♂}$ and $B^{♀}$ represent the paternal allele of chromosome A and the maternal allele of chromosome B, respectively. $IF_{A^{♂}B^{♀}}$ represents the interaction frequency matrix between $A^{♂}$ and $B^{♀}$. For example:

1) $IF_{A^{\sigma}A^{\sigma}}$ is the intra-chromosomal interaction frequency matrix within the paternal allele of chromosome A.
2) $IF_{A^{\sigma}A^{\varphi}}$ is the inter-chromosomal interaction frequency matrix between the paternal and maternal alleles of chromosome A.
3) $IF_{A^{\sigma}B^{\varphi}}$ is the inter-chromosomal interaction frequency matrix between the paternal allele of chromosome A and the maternal allele of chromosome B.

Due to the high missing rate in single cells, inter-chromosomal interaction frequencies are almost unavailable. Therefore, we only used the single-cell intra-chromosomal interaction frequency matrices, $IF_{A^{\sigma}A^{\sigma}}$ and $IF_{A^{\varphi}A^{\varphi}}$, from Dip-C at 1Mb resolution. To control the confounding factors, including genomic distance, we normalized the read depths band-wisely using BandNorm[195]. For simplicity, the normalized $IF_{A^{\sigma}A^{\sigma}}$ and $IF_{A^{\varphi}A^{\varphi}}$ will be referred as $IF_{A^{\sigma}}$ and $IF_{A^{\varphi}}$, respectively. To distinguish the interaction frequency matrices in different cells, let $A_i{}^{\sigma}$ represent the paternal allele of chromosome A in cell i, and $IF_{A_i{}^{\sigma}}$ represent the intra-chromosomal interaction frequency matrix of $A_i{}^{\sigma}$.

We calculated different correlations for a quantitive comparison:

1) Allele level correlation in the same cell (allele-level correlation):
   For chromosome A in cell i (A_i), we computed the Spearman correlation between $IF_{A_i{}^{\sigma}}$ and $IF_{A_i{}^{\varphi}}$, which represents the correlation between paternal and maternal alleles of the same chromosome within a single cell.
2) Cell level correlation in the same allele (cell-level correlation):
   a. Paternal allele: For the paternal allele of chromosome A (A^{σ}), we calculated the Spearman correlation between $IF_{A_i{}^{\sigma}}$ and $IF_{A_j{}^{\sigma}}$, which represents the correlation of the paternal allele of the same chromosome between different cells i and j.
   b. Maternal allele: For the maternal allele of chromosome A, A^{φ}, we calculated the Spearman correlation between $IF_{A_i{}^{\varphi}}$ and $IF_{A_j{}^{\varphi}}$, which represents the correlation of the maternal allele of the same chromosome between different cells i and j.
3) Overall correlation in different alleles and cells (overall correlation):

For chromosome A, we computed the Spearman correlation between $IF_{A_i^{\sigma}}$ and $IF_{A_j^{\varphi}}$, which represents the correlation of the same chromosome between different alleles across different cells i and j.

Allele-level correlations were calculated for different cells, while cell-level correlations and overall correlations were calculated for different cell pairs. Boxplots were used to visualize the correlation distributions across different chromosomes j.

4.2.6 Embedding of single-cell alleles and distance calculation

We applied BandNorm to generate embeddings for each allele of a specific chromosome in single cells from mouse Dip-C data without imputation. For each chromosome, we collected the normalized interaction frequency matrices at 1Mb resolution of each allele across individual cells and performed principal component analysis (PCA). The loadings on the top 50 principal components (PCs) were used as the embedding of the corresponding allele in a single cell. To visualize the embeddings in a two-dimensional (2D) space, we employed two strategies: (1) using the loadings on the top 2 PCs directly, and (2) using Uniform Manifold Approximation and Projection (UMAP) to project the 50-dimensional embedding vectors onto a 2D space.

4.2.7 Multi-level variability quantification and comparison

We compared different levels of chromatin variability based on single-cell allele embeddings using sampling strategies, including tissue, age, cell, and allele levels. To compare the tissue-level variability and age-level variability, we used Dip-C mouse cells from MOE and retina tissues. All cells were from male mice, including retina cells at postnatal day 7 (P7) and P28, and MOE cells at P28. For each chromosome, we randomly sampled one allele from a retina cell at P7 (${L_{re}}^{7}$), one allele from a retina cell at P28 (${L_{re}}^{28}$), and one allele from an MOE cell at P28 (${L_{moe}}^{28}$). To remove confounding factors arising from different alleles, all three sampled alleles were either paternal or maternal, i.e., ${L_{re}}^{7}$, ${L_{re}}^{28}$, ${L_{moe}}^{28}$ were either all paternal or all maternal. Euclidean distances were calculated based on the embeddings to determine the age-level distance, $dist({L_{re}}^{7}, {L_{re}}^{28})$, and the tissue-level distance, $dist({L_{moe}}^{28}, {L_{re}}^{28})$. The distances were normalized by setting the age-level distance to 1 ($dist_{age} = 1$) and calculating the

normalized tissue-level distance as$dist_{tissue} = dist({L_{re}}^{7}, {L_{re}}^{28}) \,/\, dist({L_{moe}}^{28}, {L_{re}}^{28})$. The sampling process was repeated 1,000 times, generating 1,000 $dist_{tissue}$ values for each chromosome. Boxplots were used to visualize the distribution of $dist_{tissue}$ across different chromosomes.

Similarly, to compare the age-level variability and cell-level variability, we used Dip-C mouse cells from retina tissues. All cells were from male mice, including retina cells at P7, P28, and P56. For each chromosome, we randomly sampled two alleles from different cells at P7 ($L^7{}_1$, $L^7{}_2$), one allele at P28 (L^{28}), and one allele at P56 (L^{56}). All sampled alleles were either paternal or maternal. We calculated the Euclidean distances based on the embeddings: cell-level distance $dist(L^7{}_1, L^7{}_2)$, 28-day age-level distance $dist(L^{28}, L^7{}_1)$, and 49-day age-level distance $dist(L^{56}, L^7{}_1)$. The distances were normalized by setting the cell-level distance to 1 ($dist_{cell} = 1$) and calculating the normalized age-level distance as ${dist_{age}}^{28} = dist(L^{28}, L^7{}_1) \,/\, dist(L^7{}_1, L^7{}_2)$ and ${dist_{age}}^{49} = dist(L^{56}, L^7{}_1) \,/\, dist(L^7{}_1, L^7{}_2)$. The sampling process was repeated 1,000 times, generating 1,000 ${dist_{age}}^{28}$ and ${dist_{age}}^{49}$ values for each chromosome.

In addition, to compare the cell-level variability and allele-level variability, we only used Dip-C female mouse cells at P5 from MOE tissues. For each chromosome, we randomly sampled two cells, with P_1 and M_1 representing the paternal and maternal alleles from the first cell, and P_2 and M_2 represent the paternal and maternal alleles from the second cell. Euclidean distances were calculated based on the embeddings to determine the allele-level distance, $dist(M_1, P_1)$, the cell-level distance, $dist(M1, M2)$, and the overall distance, $dist(M_1, P_2)$. The distances were normalized by setting the allele-level distance to 1 ($dist_{allele} = 1$), calculating the normalized cell-level distance as $dist_{cell} = dist(M1, M2) \,/\, dist(M_1, P_1)$, and normalized overall distance $dist_{cell+allele} = dist(M1, P2) \,/\, dist(M_1, P_1)$. The sampling process was repeated 1,000 times, generating 1,000 $dist_{cell}$ and $dist_{cell+allele}$ values for each chromosome.

4.3 RESULTS

4.3.1 Cell-type specificity of intra-chromosomal chromatin across distinct scales and resolutions

Cell-type-specific intra-chromosomal chromatin organization has been widely studied using Hi-C datasets, leading to the development of various techniques for data preprocessing, normalization, imputation, and structure reconstruction[26–28,185,186,188–192]. However, a comprehensive analysis of intra-chromosomal chromatin across different cell types, scales, and resolutions while controlling for confounding factors is still lacking. In this study, we aimed to address this gap by systematically investigating the intra-chromosomal chromatin organization in multiple cell types.

As a sequence-based experiment, Hi-C relies on read depth to capture chromosome conformation, which limits data quality and resolution[14]. Limited read depth results in higher rates of missing data at higher resolutions, with the rates of missing data being strongly related to genomic distance. For example, analysis of bulk Hi-C data from the GM12878 cell type at 10kb resolution revealed that the available rate (1 - missing rate) substantially decreases with increasing genomic distance (**Figure 4.2 A**). The available rate is approximately 90% within 1Mb, 60% within 2Mb, 40% within 3Mb, 30% within 5Mb, and 10% within 10Mb. To control for the effect of genomic distance, we split the Hi-C matrix into different genomic distance bands: (0,1Mb], (1Mb,2Mb], (2Mb,5Mb], (5Mb,10Mb], (10Mb,20Mb], and (10Mb,50Mb]. The contact frequency distribution for available entries in each band was examined (**Figure C. 1**). We observed that when the genomic distance exceeds 10Mb, the entries become almost binary. Therefore, we focused our analysis on four genomic distance bands: (0,1Mb], (1Mb,2Mb], (2Mb,5Mb], and (5Mb,10Mb].

To investigate intra-chromosomal chromatin variability across cell types, we utilized eight bulk Hi-C datasets, including two replicates of GM12878 and one replicate each of K562, KBM7, HUVEC, IMR90, HMEC, and NHEK. All Hi-C data were obtained from Rao *et al*, with similar read depths to minimize experimental biases. We calculated the Spearman correlation between the primary replicate of GM12878 and the other six cell types to quantify the variability between cell types. Additionally, we calculated the

Spearman correlation between the two replicates of GM12878 to determine the variability within the same cell type. These calculations were performed at various resolutions ranging from 1Mb to 10kb to assess the multi-scale variability in four different genomic distance bands. Significantly higher correlations within the same cell type compared to across cell types indicate that chromatin organization is stable within the same cell type (**Figure 4.2 B**). At the short genomic distance band (0,1Mb], the correlation within GM12878 is higher than 0.9 under various resolutions from 500kb to 25kb, and higher than 0.85 at 10kb resolution, demonstrating stable chromatin organization at short genomic distances across various resolutions. However, as the genomic distance increases, the correlation decreases, suggesting that stable chromatin organization only holds at lower resolutions for longer genomic distances (**Figure 4.2 B**). These findings highlight the importance of considering both genomic distance and resolution when studying intra-chromosomal chromatin variability across cell types. The stability of chromatin organization within the same cell type and the decreasing stability with increasing genomic distance provide valuable insights into the multi-scale nature of chromatin organization and its relationship to cell type identity.

To further check the influence of read depth on data quality and determine reliable resolutions, we established cell-type-specific correlation boundaries based on across-cell-type correlations at different genomic distance bands and resolutions. The correlation within the same cell type should not be lower than these boundaries under the same genomic distance and resolution settings, as a lower correlation would indicate that the chromatin organization is more dynamic than across cell types, suggesting unreliability.

For the primary replicate of GM12878, we downsampled the mapped reads from 80% to 1% to generate Hi-C matrices with fewer reads and assessed the decrease in data quality based on the correlation between the downsampled and original matrices. At low resolutions, such as 1Mb and 500kb, even with downsampling rates as low as 1%, the correlation remained higher than the cell-type-specific correlation boundaries, demonstrating extremely stable chromatin organization across different genomic distance bands (**Figure 4.2 C, Figure C. 2**). Higher resolutions required higher read depths. For instance, at the genomic distance band (0,1Mb], the minimal downsampling rates for

resolutions of 100kb, 50kb, 25kb, and 10kb were 2%, 5%, 10%, and 20%, respectively (**Figure 4.2 C**).

Furthermore, to maintain consistent correlations at the same resolution across different genomic distance bands, the read depth varied. For example, to achieve consistent correlations at 50kb or 100kb resolution, the downsampling rates could be 1% for (0,1Mb], 10% for (1Mb,2Mb], 30% for (2Mb, 5Mb], and 50% for (5Mb, 10Mb], indicating that longer genomic distance bands require more read depth to maintain consistent correlations at the same resolution (**Figure C. 2**). Similarly, to maintain a consistent correlation at the same read depth, the reliable resolutions varied across genomic distance bands. For instance, when the downsampling rate was 10%, 20%, or 30%, the reliable resolutions were 10kb for (0,1Mb], 50kb for (1Mb,2Mb], and 100kb for (2Mb, 10Mb]. When the downsampling rate decreased to 1%, the reliable resolutions were 10kb for (0,1Mb], 100kb for (1Mb,2Mb], 100kb for (2Mb, 5Mb], and 250kb for (5Mb, 10Mb] (**Figure C. 2**).

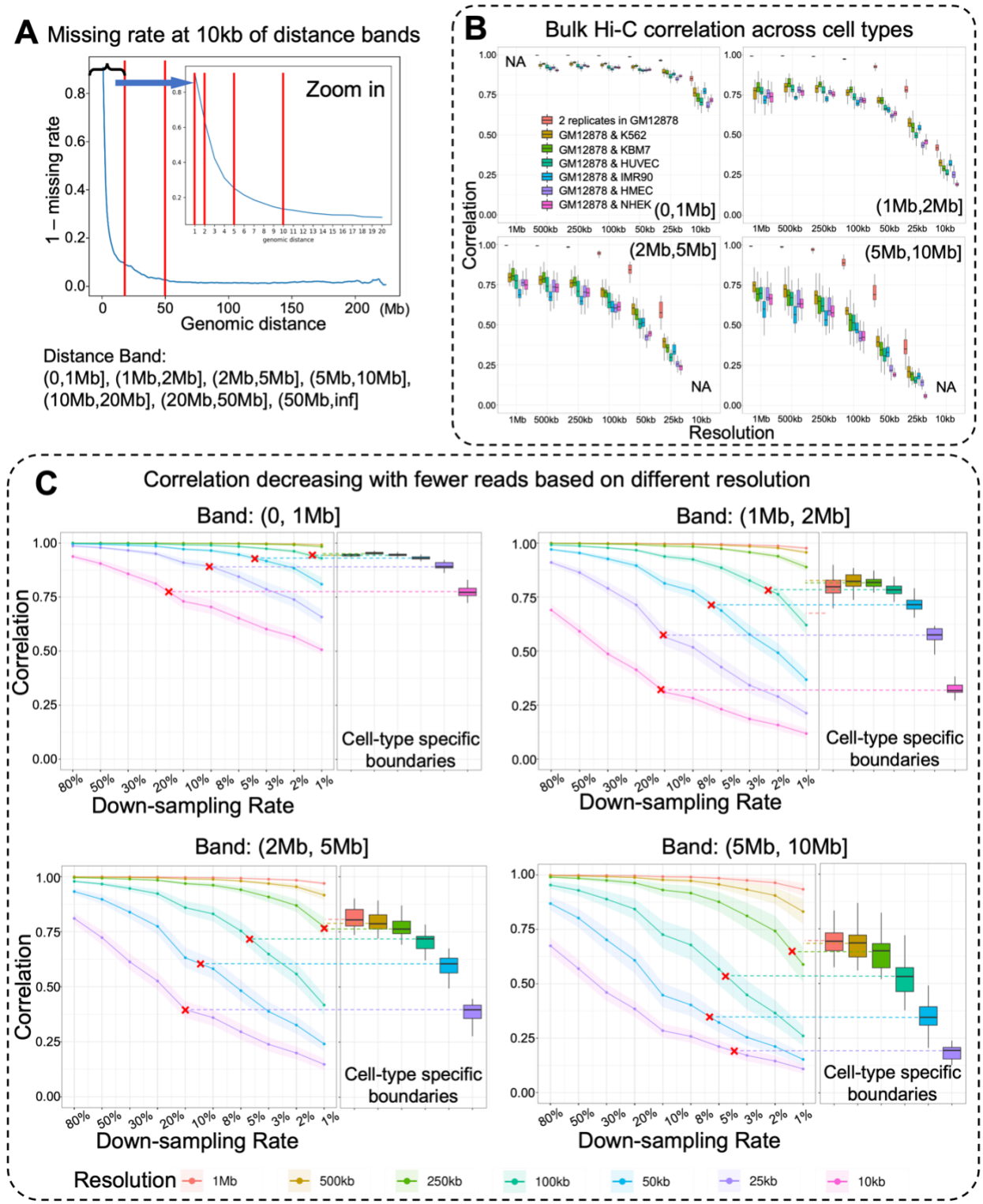

Figure 4.2. Cell-type specificity of intra-chromosomal chromatin across distinct scales and resolutions. (**A**) Missing rate based on different genomic distances based on Hi-C at 10kb resolution from GM12878 cell type. Hi-C map was split based on several genomic

Figure 4.2 (cont'd)

distance bands including (0,1Mb], (1Mb,2Mb], (2Mb,5Mb], (5Mb,10Mb], (10Mb,20Mb], (20Mb,50Mb], (50Mb,inf]. (**B**) Bulk Hi-C correlation between different cell types based on different genomic distance bands and resolution. The boxplots show the Spearman correlation in single chromosome between GM12878 with another cell type including K562, KBM7, HUVEC, IMR90, HMEC, NHEK, and another replicate of GM12878. (**C**) Correlation decreased with fewer read depths based on different genomic distance bands and resolutions. The GM12878 Hi-C was downsampled based on different rates from 80% to 1%, and the Spearman correlation was calculated between the downsampled one and original one. Each subfigure shows the results based on one genomic band. The line chart shows the correlations in single chromosome changes based on different downsampling rates. Different lines with various colors show the correlation changing trend at specific resolution. The right panel of each subfigure is the boxplot showing the cell-type specific boundary as a reference to determine the reliable correlation threshold of the downsampled Hi-C.

The observations from our analysis provide valuable biological insights into the variability of intra-chromosomal chromatin organization. We found that intra-chromosomal chromatin is more stable within the same cell type compared to the dynamics observed across cell types. This stability is particularly evident in the multi-scale nature of chromatin organization, where the large-scale skeleton is stable at low resolution, while the stable short-scale detailed organization can be captured at higher resolutions. Our findings also highlight the complex interplay between read depth, resolution, and genomic distance in determining the reliability of chromatin organization data. Read depth has a more significant influence on the reliability of longer-range chromatin interactions at higher resolutions. This analysis provides guidance for selecting appropriate resolutions and read depths when studying chromatin organization across different genomic scales. The results suggest that for any Hi-C experiment, the reliable resolution for longer genomic distances is lower compared to shorter genomic distances, due to limitations in read depth. This trade-off between resolution and genomic distance is an important consideration when designing and interpreting Hi-C experiments. To overcome this limitation, methods like FLAMINGO[28] employ a hierarchical approach to reconstruct the skeleton of an entire chromosome at low resolution and capture within-domain structures at high resolution.

In conclusion, our study reveals the stability of intra-chromosomal chromatin organization within cell types and the importance of considering the complex relationships

between read depth, resolution, and genomic distance when studying chromatin organization. These findings have implications for the design and interpretation of Hi-C experiments and the development of computational methods for multi-scale chromatin organization analysis.

4.3.2 Heterogeneity of inter-chromosomal chromatin contacts and spatial distances

Moving on to cell-type-specific inter-chromosomal chromatin, we investigated the chromatin organization between single chromosomes. Inter-chromosomal spatial distances are much longer than intra-chromosomal spatial distances[14,196,197], which results in 3C-like techniques capturing limited reads and causing extremely high missing rates for detailed inter-chromosomal organization information[196]. Moreover, due to the digestion and ligation procedures, Hi-C is biased towards intra-chromosomal interactions compared to inter-chromosomal interactions, as it lacks the ability to handle long spatial distances. SPRITE, which does not rely on digestion and ligation procedures, exhibits no bias between intra-chromosomal and inter-chromosomal interactions[194]. This property makes SPRITE a valuable tool for removing the bias in Hi-C data and enabling a fair comparison between intra-chromosomal and inter-chromosomal interaction frequencies (**Figure 4.3 A**). We plotted scatter plots showing interaction frequencies based on Hi-C and SPRITE for intra-chromosomal and inter-chromosomal interactions separately. At 1Mb resolution, inter-chromosomal interaction frequencies showed no significant correlation between Hi-C and SPRITE due to the high missing rate caused by limited read depth (**Figure C. 3**). However, at a lower resolution of 10Mb and after removing entries close to the centromere, Hi-C and SPRITE exhibited a clear regression relationship with a correlation of 0.956 and 0.776 for intra-chromosomal and inter-chromosomal interactions, respectively (**Figure 4.3 A, Figure C. 3**). We fitted separate regression models for intra-chromosomal and inter-chromosomal interactions to remove the bias in Hi-C data based on SPRITE at 10Mb resolution.

To validate the transformation applied to Hi-C data using SPRITE, we leveraged the exponential relationship between bulk-level intra-chromosomal interaction frequencies of Hi-C and real spatial distances, as established in a previous study[29]. This relationship is widely used to convert interaction frequencies to Euclidean distances in

various computational tools[27,28]. After removing the bias in inter-chromosomal interaction frequencies of Hi-C, we applied the same exponential transformation suggested by the previous study to convert the corrected inter-chromosomal interaction frequencies to spatial distances. To assess whether the converted spatial distances constitute a valid distance metric, we employed the triangle inequality. By sampling triplet entries in the inter-chromosomal distance matrix, we observed significant improvements in the fraction of cases where the triangle inequality holds, compared to the matrix based on the original Hi-C interaction frequencies (**Figure 4.3 B**). This finding suggests that our transformation of interaction frequencies based on SPRITE effectively removed the bias in Hi-C data and rendered the exponential transformation, originally developed for intra-chromosomal distances, applicable to inter-chromosomal distances as well.

Using the distance matrices converted from the corrected Hi-C interaction frequencies, we investigated whether a consensus 3D structure can represent them. Since the Euclidean distance matrix is calculated from 3D coordinates, it is expected to be low-rank, with a rank not exceeding 5[28]. We compared the inter-chromosomal distance matrix at 10Mb resolution with the intra-chromosomal distance matrix of chr1 at 1Mb resolution, as they have similar dimensions. Singular Value Decomposition (SVD) was applied to both matrices separately, and the singular values were ordered by their absolute values. The top 5 singular values could only explain 55.4% of the total variance in the inter-chromosomal distance matrix, compared to 70.2% in the intra-chromosomal distance matrix (**Figure 4.3 C**). The explanation ratio of the inter-chromosomal distance matrix was significantly and consistently lower than that of the intra-chromosomal distance matrix across different numbers of top singular values (**Figure C. 3**).

Additionally, we constructed a low-rank approximation of each distance matrix using the top 5 singular values and their corresponding singular vectors and compared the approximated distance matrices with the original ones. The approximated intra-chromosomal distance matrix was highly similar to the original matrix, with a Spearman correlation of 0.985 and a Pearson correlation of 0.977 (**Figure 4.3 C**), indicating that the bulk-level intra-chromosomal Hi-C distance matrix can be well represented by a consensus 3D structure of a single chromosome. In contrast, the approximated inter-chromosomal distance matrix differed from the original matrix, with a Spearman

correlation of 0.835 and a Pearson correlation of 0.661 (**Figure 4.3 C**). We further examined the correlation between the original matrices and their approximations based on different numbers of top singular values. The results clearly showed that the intra-chromosomal distance matrix has a rank no larger than 5, while the inter-chromosomal distance matrix has a rank much higher than 5 (**Figure C. 4**). This finding suggests that although the inter-chromosomal distance matrix represents an average across cells at the bulk level, it cannot be adequately represented by a single 3D structure.

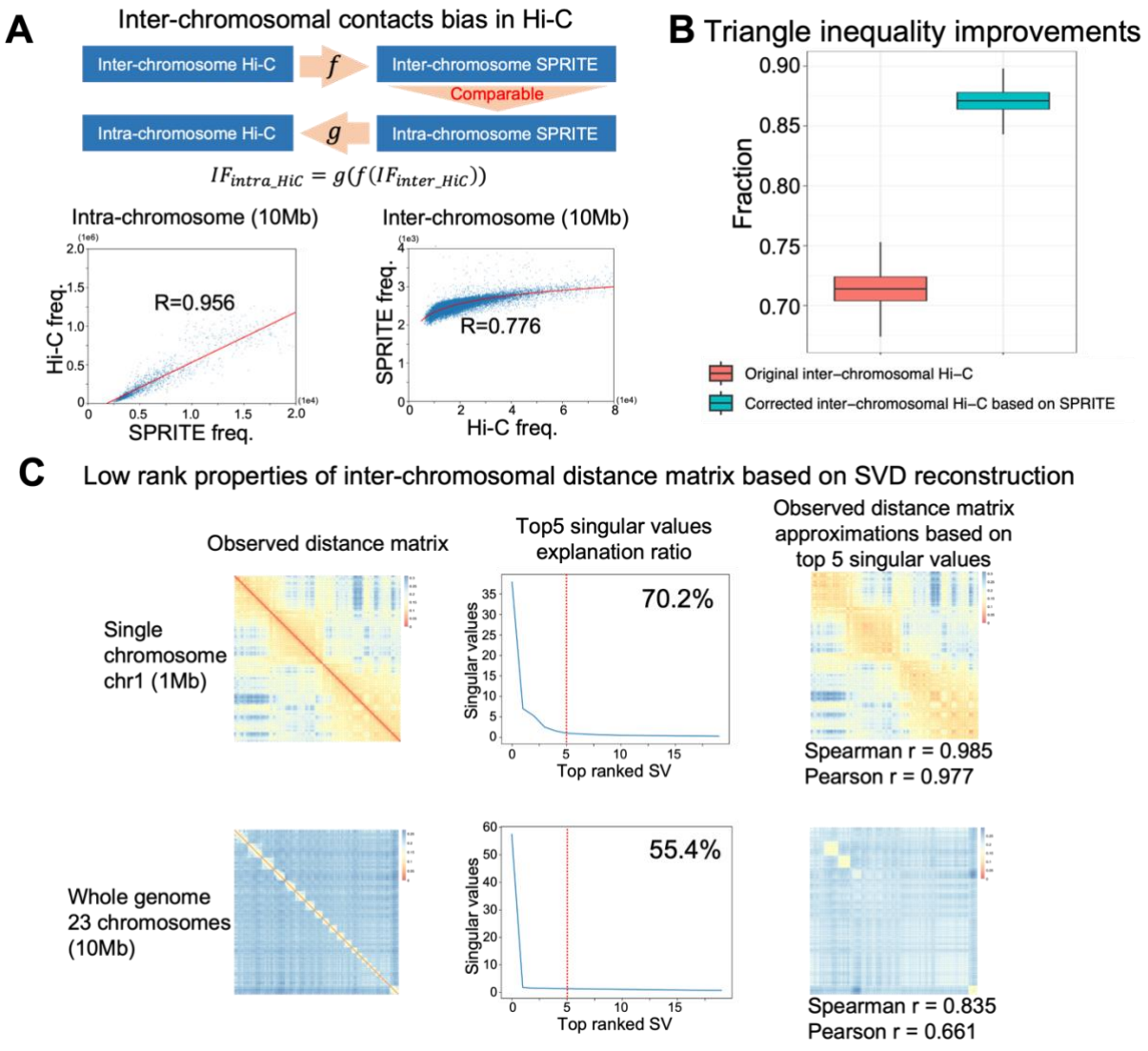

Figure 4.3. Heterogeneity of inter-chromosomal chromatin contacts in bulk tissue. (**A**) Removal of experimental biases in Hi-C data using SPRITE. The top panel shows schematic figures illustrating the correction of inter-chromosomal Hi-C contacts using

Figure 4.3 (cont'd)

SPRITE by fitting two regression models. The bottom panel displays scatter plots comparing Hi-C and SPRITE contact frequencies for intra-chromosomal and inter-chromosomal contacts separately, with the fitted models overlaid. (**B**) Comparison of the fraction of cases where the triangle inequality holds in distances transformed from original Hi-C data and corrected Hi-C data. The boxplots show the distribution of the fraction of cases across multiple samples, with higher fractions indicating better agreement with the triangle inequality and more reliable distance transformations. (**C**) Low-rank properties of intra-chromosomal and inter-chromosomal distance matrices based on singular value decomposition (SVD) approximation. Two scenarios are presented: the distance matrix derived from intra-chromosomal contacts of chromosome 1 at 1Mb resolution (top panel) and the distance matrix derived from inter-chromosomal contacts of the whole genome at 10Mb resolution (bottom panel). The line plots show the cumulative explanation ratio as a function of the number of top singular values used in the approximation. The heatmaps display the original distance matrices and their corresponding approximations based on the top 5 singular values and singular vectors, with higher correlations between the original and approximated matrices indicating better low-rank approximations.

Our results highlight the fundamental differences in the variability of intra-chromosomal and inter-chromosomal chromatin organization across cells within the same cell type. While intra-chromosomal chromatin can be well represented by a consensus 3D structure, inter-chromosomal chromatin exhibits a higher degree of complexity and variability that cannot be captured by a single structure due to the highly dynamic nature of interactions across cells. This finding suggests that the cell-to-cell variability in inter-chromosomal interactions is a crucial factor to consider when studying chromatin organization at this scale. Furthermore, our analysis indicates that inter-chromosomal interaction frequencies derived from Hi-C data are biased and noisy at high resolution. Unbiased experiments, such as SPRITE, can serve as a reference to remove these biases and provide a more accurate representation of inter-chromosomal interactions. This observation highlights the importance of integrating multiple experimental techniques to overcome the limitations of individual methods and obtain a more comprehensive understanding of chromatin organization. The findings of this study have significant implications for the development of computational methods aimed at modeling 3D chromatin structures based on bulk-tissue Hi-C data. While it is feasible to build models that reconstruct the consensus structure of single chromosomes within the same cell type, our results suggest that using a single consensus 3D structure to

represent inter-chromosomal chromatin organization based on bulk Hi-C data may be inadequate. The higher-rank nature of the inter-chromosomal distance matrix indicates the presence of substantial cell-to-cell variability, emphasizing the need for more advanced methods that can capture the complexity of inter-chromosomal interactions instead of relying on single consensus structures.

In conclusion, our study reveals the inherent differences in the variability of intra-chromosomal and inter-chromosomal chromatin organization across cells within the same cell type. These findings underscore the importance of considering cell-to-cell variability when investigating inter-chromosomal interactions and highlight the need for integrating multiple experimental techniques to remove biases and obtain a more accurate representation of chromatin organization. Moreover, our results have significant implications for the development of computational methods, emphasizing the need for advanced approaches that can capture the complexity and variability of inter-chromosomal interactions beyond single consensus structures. These insights contribute to a deeper understanding of chromatin organization and provide valuable guidance for future research in this field.

4.3.3 Diploid-specific variability of chromatin coupled with diverse sources of biological factors

Bulk Hi-C provides a cell-type-specific average representation of chromatin interactions from millions of diploid cells, each containing 46 chromosomes (23 maternal and 23 paternal alleles)[198]. Even within the same cell type or tissue, differences in cell age and allelic variation can result in different chromatin organizations. The rapid development of single-cell Hi-C (scHi-C) techniques has enabled the analysis of chromatin interactions at the single-cell level, providing the potential to investigate cell-to-cell intra-chromosomal chromatin variability[30–40]. However, compared to bulk Hi-C, scHi-C suffers from more missing data due to the extremely limited read depths, leading to lower resolution and increased challenges for analysis tools. Dip-C, an advanced scHi-C technique, generates a higher number of contacts with minimal false positives[34]. By detecting unique single-nucleotide polymorphisms (SNPs) based on paternal and maternal genomes, Dip-C distinguishes reads between the two haplotypes of each chromosome, allowing for the construction of diploid scHi-C maps. Additionally, Dip-C employs an algorithm to impute

the two chromosome haplotypes using reads without unique SNPs, assuming that the two homologs typically contact different partners[34]. This makes Dip-C a valuable resource for analyzing both cell-to-cell and allele-to-allele variability in chromatin organization. To quantify variability using the imputed diploid scHi-C maps from 17 GM12878 cells generated by Dip-C at 1Mb resolution[34], we calculated Spearman correlations, with lower correlations indicating higher variability. We computed paternal-paternal and maternal-maternal correlations across cells to represent cell-specific variability, and paternal-maternal correlations within the same cell to represent allele-specific variability. Moreover, paternal-maternal correlations from different cells were calculated to represent overall variability, which is a combination of cell-specific and allele-specific variability. Our analysis revealed that paternal-maternal correlations within the same cell were consistently lower than paternal-paternal and maternal-maternal correlations across cells (**Figure 4.4 A**). This finding indicates that allele-specific variability in chromatin organization is higher than cell-specific variability, suggesting that differences between homologous chromosomes within a cell are more pronounced than differences between the same chromosome across cells of the same type based on the imputed diploid Hi-C maps.

However, the conclusion drawn from the imputed diploid Hi-C data may be influenced by the biased assumption of the imputation algorithm. The imputed diploid Hi-C is based on the assumption that the two homologs typically contact different partners, which assigns reads with unknown haplotypes to different alleles by maximizing the difference between paternal and maternal alleles within a single cell. To address this potential bias, we re-examined the mapped reads within the homologous chromosomes without imputation (**Figure 4.4 B**). Our analysis revealed that the majority of reads (87.61%) do not harbor any unique SNPs, allowing us to determine the chromosome of origin but not the specific allele. Only 11.79% of reads are one-phased, meaning that one fragment of the read harbors unique SNPs that can determine the allele of origin for that fragment, while the allele of the other fragment remains unknown. The remaining 0.6% of reads are two-phased, indicating that both fragments of the read harbor unique SNPs, enabling us to determine the allele of origin for both fragments. Interestingly, among the two-phased reads, only 1.17% are inter-chromosomal contacts (**Figure 4.4 B**), which we

refer to as two-homolog reads. This finding suggests that the majority of contacts are intra-chromosomal. Therefore, although the specific allele of the other fragment in one-phased reads is unknown, we can reasonably impute them as originating from the same allele, considering them as intra-chromosomal contacts. Based on this rationale, we constructed single-cell diploid intra-chromosomal Hi-C maps using only one-phased reads without any biased imputation.

To account for the important confounding factor of genomic distance, we applied BandNorm[195] to normalize the single-cell diploid intra-chromosomal Hi-C maps and repeated the correlation analysis. After normalization, the previously observed significantly and consistently lower correlations for paternal-maternal comparisons within the same cell disappeared (**Figure 4.4 C**). Interestingly, the correlations for paternal-maternal comparisons within the same cell were slightly higher than the paternal-paternal and maternal-maternal correlations across cells, but the differences were not significant. These results demonstrate that the relationship between allele-specific variability and cell-specific variability is complex and requires further investigation.

To gain deeper insights, we utilized additional Dip-C data from mouse cells. We generated embeddings for each allele of the chromosomes and visualized them across cells using UMAP for each chromosome separately. In the resulting visualizations, each dot represents either the paternal or maternal allele of a specific chromosome in a single cell (**Figure 4.4 D**). The colors of the dots in different subfigures were based on different tissues, sexes, alleles, and cell ages, respectively. Our analysis revealed clearly distinct patterns between different tissues, sexes, and cell ages. Interestingly, the differences between alleles were not significant in chromosome 1 to chromosome 22. However, chromosome X showed significant differences, which were mainly attributed to the differences in sex. These observations suggest that the overall differences in chromatin organization across cells arise from multiple levels of variability, including tissue-level, cell-age-level, cell-level, and allele-level variability.

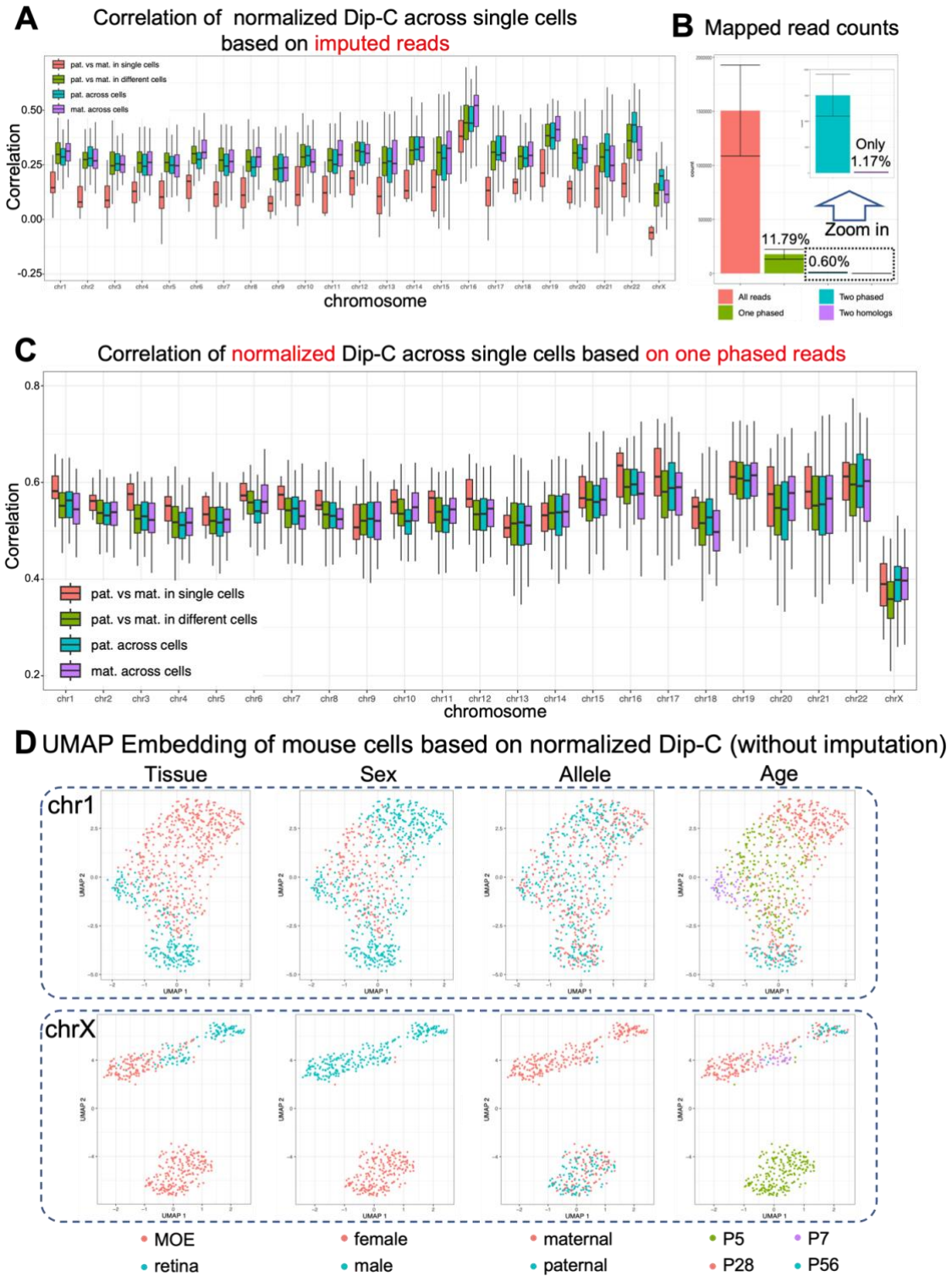

Figure 4.4. Diploid single-cell-specific variability of chromatin based on Dip-C. (**A**) Spearman correlation between two alleles under four scenarios in different chromosomes

Figure 4.4 (cont'd)

based on imputed Dip-C data. The boxplots show the distribution of correlations for each scenario: (1) paternal and maternal alleles within the same cell (red boxes), (2) paternal alleles between different cells (blue boxes), (3) maternal alleles between different cells (purple boxes), and (4) paternal allele and maternal allele in two different cells (green boxes). Higher correlations indicate greater similarity in chromatin organization between the compared alleles. (**B**) Counts of mapped reads from Dip-C experiments. The bar plot shows the number of reads in different categories, including all reads (red), one-phased reads (green), and two-phased reads (blue), and two-homologous reads (purple). (**C**) Spearman correlation between two alleles under four scenarios in different chromosomes based on one-phased reads from Dip-C data without imputation. The boxplots follow the same color scheme as in panel (A). (**D**) UMAP embeddings of mouse cells based on one-phased reads from normalized Dip-C data without imputation. The UMAPs are shown for different chromosomes separately, with each dot representing one allele of a specific chromosome. The colors of the dots represent different tissues, sexes, alleles, and cell ages, allowing for the visualization of cell-to-cell variability in chromatin organization across various biological factors.

These findings underscore the importance of considering various sources of variability when analyzing cell-to-cell chromatin organization. Making biased assumptions, such as assuming that the two homologs typically contact different partners, can lead to misinterpretations of the data. By utilizing normalized single-cell diploid intra-chromosomal Hi-C maps and employing unbiased analysis methods, we can gain a more accurate understanding of the complex interplay between different levels of variability in chromatin organization.

4.3.4 Ordered hierarchy of structural variability guiding 3D chromatin analyses under different contexts

To further investigate the relative contributions of tissue-level, cell-age-level, cell-level, and allele-level variability, we developed sampling methods based on the generated embeddings. By controlling for confounding factors, the sampling-based approach allows for a more rigorous and unbiased comparison of different levels of variability in chromatin organization.

First, to compare tissue-level and age-level chromatin variability, we focused on male cells from the Dip-C dataset, including main olfactory epithelium (MOE) cells at postnatal day 28 (P28), and retina cells at P28 and P7. For each chromosome, we randomly sampled one allele from a retina cell at P28, one allele from a retina cell at P7,

and one allele from an MOE cell at P28. To control for confounding factors arising from different alleles, all three sampled alleles were either paternal or maternal. We then calculated Euclidean distances based on the embeddings to determine the age-level distance (i.e., the distance between retina cells at P28 and P7) and the tissue-level distance (i.e., the distance between retina and MOE cells at P28) (**Figure 4.5 A**). The distances were normalized based on age-level distance to remove sampling bias. We repeated the sampling process 1,000 times for each chromosome and found that the normalized tissue-level distance was significantly larger than the age-level distance across all chromosomes (**Figure 4.5 B**). This result indicates that tissue-level variability in chromatin organization is higher than age-level variability across different cells.

Next, we compared age-level and cell-level chromatin variability using male retina cells from different cell ages, including P7, P28, and P56. Similarly, we randomly sampled one allele from cells at P28 and P56, and two alleles from cells at P7, while controlling for the allele origin (either paternal or maternal) to ensure a fair comparison. Euclidean distances were then calculated based on the embeddings to determine the age-level distances (i.e., the distance between P7 and P28, and the distance between P7 and P56) and the cell-level distance (i.e., the distance between two cells at P7) (**Figure 4.5 C**). To remove sampling bias, the distances were normalized based on the cell-level distance. We repeated the sampling process 1,000 times for each chromosome and found that the normalized age-level distance was significantly larger than the cell-level distance across all chromosomes (**Figure 4.5 D**). Interestingly, the magnitude of the difference in age-level distance increased with the age difference between the compared developmental stages (**Figure 4.5 D**). This result suggests that age-level variability in chromatin organization is more pronounced than cell-level variability across different cells, and that the extent of this difference is related to the magnitude of the age difference.

Finally, we investigated cell-level variability and allele-level variability using female MOE cells at P7 to control for tissue-level and age-level variabilities. For each chromosome, we randomly sampled both paternal and maternal alleles from two cells separately. Euclidean distances were then calculated based on the embeddings to determine the cell-level distance (i.e., the distance between maternal alleles from two cells), the allele-level distance (i.e., the distance between maternal and paternal alleles

from the same cell), and the overall distance (i.e., the distance between the maternal allele of one cell and the paternal allele from another cell) (**Figure 4.5 E**). To remove sampling bias, the distances were normalized based on the allele-level distance. We repeated the sampling process 1,000 times for each chromosome and found that the normalized cell-level distance was significantly larger than the allele-level distance for chromosomes 1 through 22 (**Figure 4.5 E**). This result suggests that cell-to-cell differences contribute more to the overall variability in chromatin organization than allele-specific differences within individual cells for these chromosomes. Interestingly, chromosome X exhibited a contrasting pattern, with the allele-level distance being significantly larger than the cell-level distance (**Figure 4.5 F**). This observation indicates that, for chromosome X, allele-specific differences in chromatin organization are more pronounced than cell-to-cell differences. The distinct behavior of chromosome X compared to the autosomes highlights the unique regulatory mechanisms and evolutionary pressures associated with this sex chromosome. The larger allele-level distance in chromosome X suggests that the maternal and paternal copies of this chromosome may be subject to different regulatory environments or epigenetic modifications, leading to more pronounced differences in their chromatin organization.

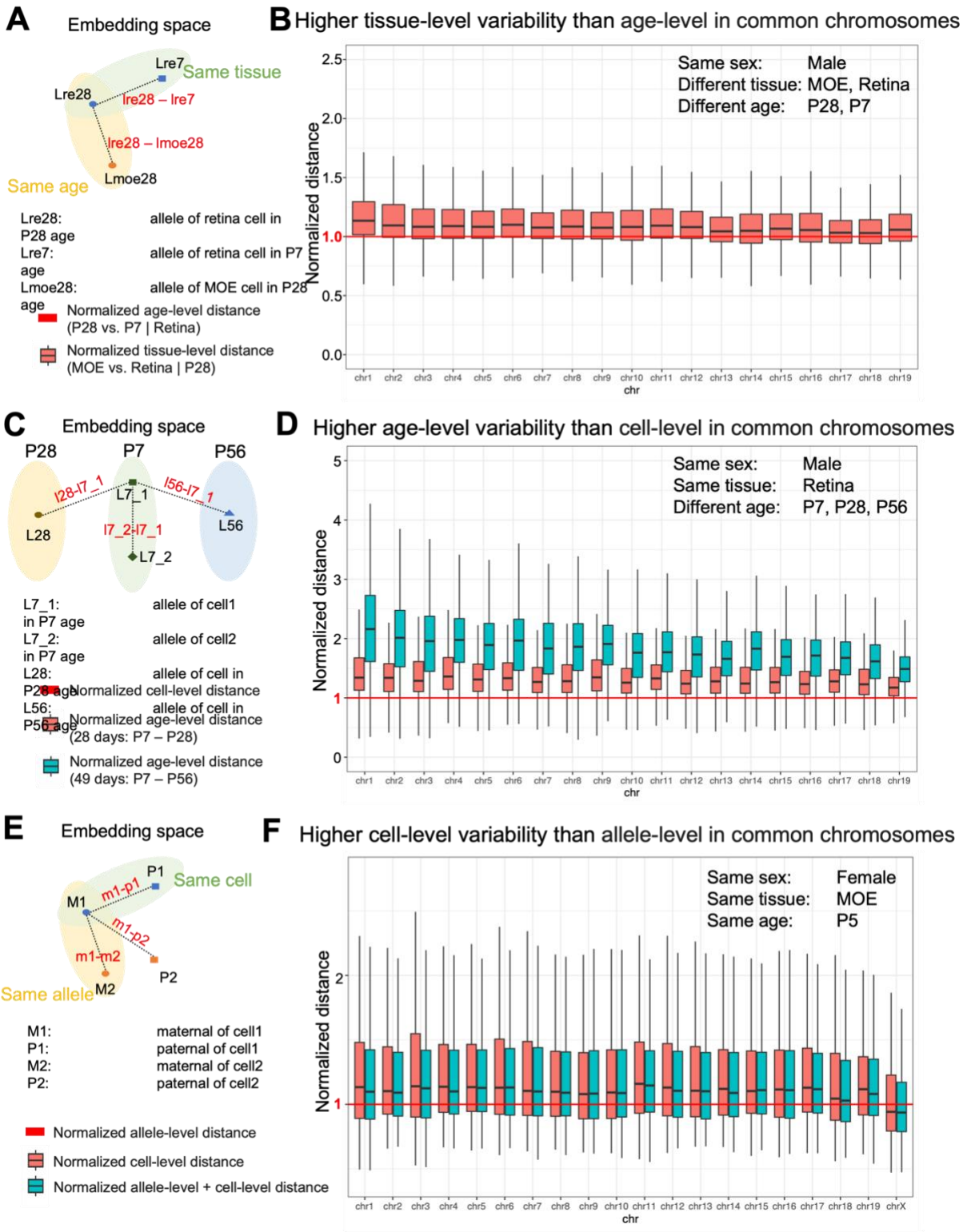

Figure 4.5. Ordered hierarchy of structural variability guiding 3D chromatin analyses under different contexts. (**A**-**B**) Comparison of tissue-level and cell-age-level variability.

Figure 4.5 (cont'd)

(**A**) Schematic figure illustrating the sampling strategy from embedding spaces. The tissue-level distance and age-level distance were calculated based on embeddings and normalized. (**B**) Boxplots showing the distribution of normalized tissue-level distances across different chromosomes. Higher values indicate greater tissue-level variability compared to age-level variability. (**C**-**D**) Comparison of cell-age-level and cell-level variability. (**C**) Schematic figure illustrating the sampling strategy from embedding spaces. The age-level distance and cell-level distance were calculated based on embeddings and normalized. (**D**) Boxplots showing the distribution of normalized age-level distances across different chromosomes, including age-level distances over 28 days (red) and 49 days (blue). Higher values indicate greater age-level variability compared to cell-level variability. (**E**-**F**) Comparison of cell-level and allele-level variability. (**E**) Schematic figure illustrating the sampling strategy from embedding spaces. The cell-level distance and allele-level distance were calculated based on embeddings and normalized. (**F**) Boxplots showing the distribution of normalized cell-level distances across different chromosomes. Higher values indicate greater cell-level variability compared to allele-level variability.

Our analysis of single-cell Hi-C data reveals the relative contributions of different sources of variability in chromatin organization. We found that tissue-level variability is more pronounced than age-level variability, indicating that cell type identity plays a crucial role in shaping chromatin structure. Additionally, age-level variability is more significant than cell-level variability, suggesting that developmental changes have a greater impact on chromatin organization than cell-to-cell differences within a specific age group. Furthermore, cell-level variability is generally more pronounced than allele-level variability for autosomes, while chromosome X exhibits a contrasting pattern with larger allele-level variability. These findings highlight the importance of considering the ordered hierarchy of cell type-specific factors, developmental dynamics, and allele-specific differences when studying chromatin structure and its relationship to gene regulation and cellular function. For autosomes, the variability order is as follows: tissue-level > age-level > individual cell-level > allele-level. In contrast, chromosome X exhibits a different order: tissue-level > age-level > allele-level > individual cell-level. These variability orders have important implications for modeling and analyzing chromatin structures at different levels of inquiry. When investigating chromatin organization at a specific level, variability sources lower in the hierarchy can be considered less influential and potentially ignored, while variability sources higher in the hierarchy should be carefully accounted for in the analysis.

For example, when studying chromatin structure at the tissue level, age-related changes, individual cell-to-cell differences, and allele-specific variations may have a smaller impact on the overall organization and could be treated as less significant factors. However, when investigating chromatin organization at the allelic level, it is crucial to consider the effects of tissue-specific factors, developmental dynamics, and cell-to-cell differences, as these sources of variability rank higher in the hierarchy and can substantially influence the observed chromatin structure. In conclusion, the insights of variability orders provide a valuable framework for future studies aimed at unraveling the complex interplay between chromatin organization, gene expression, and cellular identity.

CHAPTER 5
DISCUSSION AND FUTURE DIRECTIONS

Deciphering the effects of non-coding genetic variants is critical for understanding human diseases. With the knowledge gained from cell type-specific 3D chromatin organization, researchers have developed numerous machine learning methods to boost the prediction of causal genetic variants associated with corresponding phenotypes, including cancers, compared to traditional statistical methods. In this thesis, we developed APRIL to leverage long-range chromatin regulatory networks for discovering disease-associated genes, demonstrating the vast potential and interpretability of understanding underlying mechanisms through chromatin regulation networks. We also developed 3DVariantVision to provide a comprehensive understanding of genetic variants based on 3D chromatin information, including eQTL prediction, local and 3D effect discovery, and even insights into disease associations. Additionally, we provided a comprehensive analysis of the multi-level variability in multi-scale 3D chromatin organization, offering valuable insights into the complex interplay between cell type-specific factors, developmental dynamics, and allele-specific differences, which can guide future experimental designs and computational method development in the field of 3D chromatin organization.

Building upon these contributions, we propose the following future research directions: 1) Novel 3D chromatin structure reconstruction methods: Leveraging our comprehensive analysis of multi-level variability in multi-scale 3D chromatin organization, we aim to develop novel methods for reconstructing 3D chromatin structures that provide a more comprehensive chromatin network for long-range interactions. 2) Graph Neural Networks for disease-associated gene discovery: APRIL has demonstrated the importance of chromatin regulatory networks in discovering disease-associated genes. With the comprehensive chromatin network for long-range interactions reconstructed using our proposed methods, advanced techniques such as Graph Neural Networks can be applied to further investigate signals within the network, including disease-associated genes and GWAS hits. 3) Comprehensive modeling of genome-wide chromatin regulation networks: 3DVariantVision is a pioneering framework for comprehensive understanding of genetic variants based on 3D chromatin. However, it currently focuses on one-to-one

interactions (enhancer-promoter interactions) without considering the overall view of chromatin regulation. In the future, we plan to model the entire genome-wide chromatin regulation network to capture more complex and comprehensive underlying mechanisms. 4) Integration of Foundation Models for DNA sequences: With the rapid development of AI techniques, including Foundation Models for DNA sequences[76–79], we can further utilize the embeddings directly from these models to enhance the performance and interpretability of our entire framework.

By pursuing these future research directions, we aim to deepen our understanding of the complex relationships between genetic variants, 3D chromatin organization, and human diseases. The integration of advanced computational methods, such as Graph Neural Networks and Foundation Models, with the knowledge gained from our comprehensive analyses will pave the way for more accurate predictions and mechanistic insights, ultimately contributing to the development of precision medicine and improved clinical care.

BIBLIOGRAPHY

1. Klein, R. J. *et al.* Complement Factor H Polymorphism in Age-Related Macular Degeneration. *Science (1979)* **308**, 385–389 (2005).

2. Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. Five Years of GWAS Discovery. *The American Journal of Human Genetics* **90**, 7–24 (2012).

3. Hirschhorn, J. N. & Daly, M. J. Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet* **6**, 95–108 (2005).

4. Tam, V. *et al.* Benefits and limitations of genome-wide association studies. *Nat Rev Genet* **20**, 467–484 (2019).

5. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* **47**, 291–295 (2015).

6. Emmert-Streib, F., Dehmer, M. & Haibe-Kains, B. Gene regulatory networks and their applications: understanding biological and medical problems in terms of networks. *Front Cell Dev Biol* **2**, (2014).

7. Misteli, T. Beyond the Sequence: Cellular Organization of Genome Function. *Cell* **128**, 787–800 (2007).

8. Buermans, H. P. J. & den Dunnen, J. T. Next generation sequencing technology: Advances and applications. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease* **1842**, 1932–1941 (2014).

9. Luo, Y. *et al.* New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. *Nucleic Acids Res* **48**, D882–D889 (2020).

10. Zhang, F. & Lupski, J. R. Non-coding genetic variants in human disease. *Hum Mol Genet* **24**, R102–R110 (2015).

11. Bernstein, B. E. *et al.* The NIH Roadmap Epigenomics Mapping Consortium. *Nat Biotechnol* **28**, 1045–1048 (2010).

12. Xu, H., Zhang, S., Yi, X., Plewczynski, D. & Li, M. J. Exploring 3D chromatin contacts in gene regulation: The evolution of approaches for the identification of functional enhancer-promoter interaction. *Comput Struct Biotechnol J* **18**, 558–570 (2020).

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

14. Rao, S. S. P. *et al.* A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell* **159**, 1665–1680 (2014).

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

16. Wei, C.-L. *et al.* A Global Map of p53 Transcription-Factor Binding Sites in the Human Genome. *Cell* **124**, 207–219 (2006).

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

18. He, B., Chen, C., Teng, L. & Tan, K. Global view of enhancer–promoter interactome in human cells. *Proceedings of the National Academy of Sciences* **111**, (2014).

19. 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).

20. Singh, S., Yang, Y., Póczos, B. & Ma, J. Predicting enhancer-promoter interaction from genomic sequence with deep neural networks. *Quantitative Biology* **7**, 122–137 (2019).

21. Li, W., Wong, W. H. & Jiang, R. DeepTACT: predicting 3D chromatin contacts via bootstrapping deep learning. *Nucleic Acids Res* **47**, e60–e60 (2019).

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

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

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

25. Tjong, H. *et al.* Population-based 3D genome structure analysis reveals driving forces in spatial genome organization. *Proceedings of the National Academy of Sciences* **113**, (2016).

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

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

28. Wang, H., Yang, J., Zhang, Y., Qian, J. & Wang, J. Reconstruct high-resolution 3D genome structures for diverse cell-types using FLAMINGO. *Nat Commun* **13**, 2645 (2022).

29. Wang, S. *et al.* Spatial organization of chromatin domains and compartments in single chromosomes. *Science (1979)* **353**, 598–602 (2016).

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

31. Stevens, T. J. *et al.* 3D structures of individual mammalian genomes studied by single-cell Hi-C. *Nature* **544**, 59–64 (2017).

32. Nagano, T. *et al.* Cell-cycle dynamics of chromosomal organization at single-cell resolution. *Nature* **547**, 61–67 (2017).

33. Flyamer, I. M. *et al.* Single-nucleus Hi-C reveals unique chromatin reorganization at oocyte-to-zygote transition. *Nature* **544**, 110–114 (2017).

34. Tan, L., Xing, D., Chang, C.-H., Li, H. & Xie, X. S. Three-dimensional genome structures of single diploid human cells. *Science (1979)* **361**, 924–928 (2018).

35. Tan, L., Xing, D., Daley, N. & Xie, X. S. Three-dimensional genome structures of single sensory neurons in mouse visual and olfactory systems. *Nat Struct Mol Biol* **26**, 297–307 (2019).

36. Ramani, V. *et al.* Massively multiplex single-cell Hi-C. *Nat Methods* **14**, 263–266 (2017).

37. Kim, H.-J. *et al.* Capturing cell type-specific chromatin compartment patterns by applying topic modeling to single-cell Hi-C data. *PLoS Comput Biol* **16**, e1008173 (2020).

38. Lee, D.-S. *et al.* Simultaneous profiling of 3D genome structure and DNA methylation in single human cells. *Nat Methods* **16**, 999–1006 (2019).

39. Li, G. *et al.* Joint profiling of DNA methylation and chromatin architecture in single cells. *Nat Methods* **16**, 991–993 (2019).

40. Nguyen, H. Q. *et al.* 3D mapping and accelerated super-resolution imaging of the human genome using in situ sequencing. *Nat Methods* **17**, 822–832 (2020).

41. Zhang, R., Zhou, T. & Ma, J. Multiscale and integrative single-cell Hi-C analysis with Higashi. *Nat Biotechnol* **40**, 254–261 (2022).

42. Zeng, Z. & Bromberg, Y. Predicting Functional Effects of Synonymous Variants: A Systematic Review and Perspectives. *Front Genet* **10**, (2019).

43. Ritchie, G. R. S., Dunham, I., Zeggini, E. & Flicek, P. Functional annotation of noncoding sequence variants. *Nat Methods* **11**, 294–296 (2014).

44. Fu, Y. *et al.* FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. *Genome Biol* **15**, 480 (2014).

45. Lee, D., Karchin, R. & Beer, M. A. Discriminative prediction of mammalian enhancers from DNA sequence. *Genome Res* **21**, 2167–2180 (2011).

46. Ghandi, M., Lee, D., Mohammad-Noori, M. & Beer, M. A. Enhanced Regulatory Sequence Prediction Using Gapped k-mer Features. *PLoS Comput Biol* **10**, e1003711 (2014).

47. Lee, D. *et al.* A method to predict the impact of regulatory variants from DNA sequence. *Nat Genet* **47**, 955–961 (2015).

48. Ogata, J. D. *et al.* excluderanges: exclusion sets for T2T-CHM13, GRCm39, and other genome assemblies. *Bioinformatics* **39**, (2023).

49. Dozmorov, M. G. *et al.* CTCF: an R/bioconductor data package of human and mouse CTCF binding sites. *Bioinformatics Advances* **2**, (2022).

50. Carlson, A. L. *et al.* Infant gut microbiome composition is associated with non-social fear behavior in a pilot study. *Nat Commun* **12**, 3294 (2021).

51. Innocenti, F. *et al.* Next-generation sequencing (NGS) in metastatic colorectal cancer (mCRC): Novel mutated genes and their effect on response to therapy (Alliance). *Annals of Oncology* **30**, v198–v199 (2019).

52. Innocenti, F. *et al.* DNA Mutational Profiling in Patients With Colorectal Cancer Treated With Standard of Care Reveals Differences in Outcome and Racial Distribution of Mutations. *Journal of Clinical Oncology* **42**, 399–409 (2024).

53. Chen, J., Mu, W., Li, Y. & Li, D. On the Identifiability and Interpretability of Gaussian Process Models. in *Advances in Neural Information Processing Systems* (eds. Oh, A. et al.) vol. 36 70267–70278 (Curran Associates, Inc., 2023).

54. Mu, W. *et al.* Airpart: interpretable statistical models for analyzing allelic imbalance in single-cell datasets. *Bioinformatics* **38**, 2773–2780 (2022).

55. Lê Cao, K.-A. *et al.* Community-wide hackathons to identify central themes in single-cell multi-omics. *Genome Biol* **22**, 220 (2021).

56. Mu, W. *et al.* bootRanges: flexible generation of null sets of genomic ranges for hypothesis testing. *Bioinformatics* **39**, (2023).

57. Davis, E. S. *et al.* matchRanges: generating null hypothesis genomic ranges via covariate-matched sampling. *Bioinformatics* **39**, (2023).

58. Ma, H., Zeng, D. & Liu, Y. Learning Optimal Group-structured Individualized Treatment Rules with Many Treatments. *Journal of Machine Learning Research* **24**, 1–48 (2023).

59. Ma, H. *et al.* Disentangling sex-dependent effects of APOE on diverse trajectories of cognitive decline in Alzheimer's disease. *Neuroimage* **292**, 120609 (2024).

60. Ma, H., Liu, Y. & Wu, G. Elucidating Multi-Stage Progression of Neuro-degeneration Process in Alzheimer's Disease. *Alzheimer's & Dementia* **18**, (2022).

61. Ma, H., Zeng, D. & Liu, Y. Learning Individualized Treatment Rules with Many Treatments: A Supervised Clustering Approach Using Adaptive Fusion. in *Advances in Neural Information Processing Systems* (eds. Koyejo, S. et al.) vol. 35 15956–15969 (Curran Associates, Inc., 2022).

62. Liu, R., Yuan, H., Johnson, K. A. & Krishnan, A. CONE: COntext-specific Network Embedding via Contextualized Graph Attention. *bioRxiv* (2023) doi:10.1101/2023.10.21.563390.

63. Liu, R. & Krishnan, A. Open Biomedical Network Benchmark: A Python Toolkit for Benchmarking Datasets with Biomedical Networks. in *Proceedings of the 18th Machine Learning in Computational Biology meeting* (eds. Knowles, D. A. & Mostafavi, S.) vol. 240 23–59 (PMLR, 2024).

64. Ding, J. *et al.* DANCE: a deep learning library and benchmark platform for single-cell analysis. *Genome Biol* **25**, 72 (2024).

65. Liu, R. *et al.* Taxonomy of Benchmarks in Graph Representation Learning. in *Proceedings of the First Learning on Graphs Conference* (eds. Rieck, B. & Pascanu, R.) vol. 198 6:1–6:25 (PMLR, 2022).

66. Liu, R., Hirn, M. & Krishnan, A. Accurately modeling biased random walks on weighted networks using *node2vec*+. *Bioinformatics* **39**, (2023).

67. Liu, R. & Krishnan, A. PecanPy: a fast, efficient and parallelized Python implementation of *node2vec*. *Bioinformatics* **37**, 3377–3379 (2021).

68. Liu, R., Mancuso, C. A., Yannakopoulos, A., Johnson, K. A. & Krishnan, A. Supervised learning is an accurate method for network-based gene classification. *Bioinformatics* **36**, 3457–3465 (2020).

69. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning–based sequence model. *Nat Methods* **12**, 931–934 (2015).

70. Kelley, D. R., Snoek, J. & Rinn, J. L. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res* **26**, 990–999 (2016).

71. Quang, D. & Xie, X. DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Res* **44**, e107–e107 (2016).

72. Kelley, D. R. *et al.* Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Res* **28**, 739–750 (2018).

73. Zhou, J. *et al.* Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nat Genet* **50**, 1171–1179 (2018).

74. Vaswani, A. *et al.* Attention is All you Need. in *Advances in Neural Information Processing Systems* (eds. Guyon, I. et al.) vol. 30 (Curran Associates, Inc., 2017).

75. Avsec, Ž. *et al.* Effective gene expression prediction from sequence by integrating long-range interactions. *Nat Methods* **18**, 1196–1203 (2021).

76. Ji, Y., Zhou, Z., Liu, H. & Davuluri, R. V. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics* **37**, 2112–2120 (2021).

77. Zhou, Z. *et al.* DNABERT-2: Efficient Foundation Model and Benchmark For Multi-Species Genome. (2023).

78. Zhou, Z. *et al.* DNABERT-S: Learning Species-Aware DNA Embedding with Genome Foundation Models. (2024).

79. Dalla-Torre, H. *et al.* The Nucleotide Transformer: Building and Evaluating Robust Foundation Models for Human Genomics. *bioRxiv* 2023.01.11.523679 (2023) doi:10.1101/2023.01.11.523679.

80. Hwang, B., Lee, J. H. & Bang, D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp Mol Med* **50**, 1–14 (2018).

81. van der Wijst, M. G. P. *et al.* Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. *Nat Genet* **50**, 493–497 (2018).

82. Granja, J. M. *et al.* Single-cell multiomic analysis identifies regulatory programs in mixed-phenotype acute leukemia. *Nat Biotechnol* **37**, 1458–1465 (2019).

83. Fang, R. *et al.* Comprehensive analysis of single cell ATAC-seq data with SnapATAC. *Nat Commun* **12**, 1337 (2021).

84. Korkmaz, G. *et al.* Functional genetic screens for enhancer elements in the human genome using CRISPR-Cas9. *Nat Biotechnol* **34**, 192–198 (2016).

85. Cong, L. *et al.* Multiplex Genome Engineering Using CRISPR/Cas Systems. *Science (1979)* **339**, 819–823 (2013).

86. Fulco, C. P. *et al.* Systematic mapping of functional enhancer–promoter connections with CRISPR interference. *Science (1979)* **354**, 769–773 (2016).

87. Pulido-Quetglas, C. *et al.* Scalable Design of Paired CRISPR Guide RNAs for Genomic Deletion. *PLoS Comput Biol* **13**, e1005341 (2017).

88. Chanock, S. High marks for GWAS. *Nat Genet* **41**, 765–766 (2009).

89. Billings, L. K. & Florez, J. C. The genetics of type 2 diabetes: what have we learned from GWAS? *Ann N Y Acad Sci* **1212**, 59–77 (2010).

90. Gandhi, S. & Wood, N. W. Genome-wide association studies: the key to unlocking neurodegeneration? *Nat Neurosci* **13**, 789–794 (2010).

91. Simmonds, M. J. GWAS in autoimmune thyroid disease: redefining our understanding of pathogenesis. *Nat Rev Endocrinol* **9**, 277–287 (2013).

92. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

93. Lamparter, D., Marbach, D., Rueedi, R., Kutalik, Z. & Bergmann, S. Fast and Rigorous Computation of Gene and Pathway Scores from SNP-Based Summary Statistics. *PLoS Comput Biol* **12**, e1004714 (2016).

94. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* **169**, 1177–1186 (2017).

95. Hudson, T. J. Wanted: regulatory SNPs. *Nat Genet* **33**, 439–440 (2003).

96. Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).

97. Civelek, M. & Lusis, A. J. Systems genetics approaches to understand complex traits. *Nat Rev Genet* **15**, 34–48 (2014).

98. Nadeau, J. H. & Dudley, A. M. Systems Genetics. *Science (1979)* **331**, 1015–1016 (2011).

99. Wiseman, F. K. *et al.* A genetic cause of Alzheimer disease: mechanistic insights from Down syndrome. *Nat Rev Neurosci* **16**, 564–574 (2015).

100. Gotoda, T. From Association to Function in the Post-GWAS Era. *J Atheroscler Thromb* **22**, 442–444 (2015).

101. Gallagher, M. D. & Chen-Plotkin, A. S. The Post-GWAS Era: From Association to Function. *The American Journal of Human Genetics* **102**, 717–730 (2018).

102. Lee, T. I. & Young, R. A. Transcriptional Regulation and Its Misregulation in Disease. *Cell* **152**, 1237–1251 (2013).

103. Calabrese, G. M. *et al.* Integrating GWAS and Co-expression Network Data Identifies Bone Mineral Density Genes SPTBN1 and MARK3 and an Osteoblast Functional Module. *Cell Syst* **4**, 46-59.e4 (2017).

104. Gao, L. *et al.* Identifying noncoding risk variants using disease-relevant gene regulatory networks. *Nat Commun* **9**, 702 (2018).

105. Lee, I., Blom, U. M., Wang, P. I., Shim, J. E. & Marcotte, E. M. Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Res* **21**, 1109–1121 (2011).

106. Lieberman-Aiden, E. *et al.* Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science (1979)* **326**, 289–293 (2009).

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

108. 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).

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

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

111. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).

112. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43–49 (2011).

113. Maurano, M. T. *et al.* Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science (1979)* **337**, 1190–1195 (2012).

114. Huang, Y.-T., VanderWeele, T. J. & Lin, X. Joint analysis of SNP and gene expression data in genetic association studies of complex diseases. *Ann Appl Stat* **8**, (2014).

115. Huang, Y., Liang, L., Moffatt, M. F., Cookson, W. O. C. M. & Lin, X. iGWAS: Integrative Genome-Wide Association Studies of Genetic and Genomic Data for Disease Susceptibility Using Mediation Analysis. *Genet Epidemiol* **39**, 347–356 (2015).

116. Tian, D., Zhang, R., Zhang, Y., Zhu, X. & Ma, J. MOCHI enables discovery of heterogeneous interactome modules in 3D nucleome. *Genome Res* **30**, 227–238 (2020).

117. Spitz, F. & Furlong, E. E. M. Transcription factors: from enhancer binding to developmental control. *Nat Rev Genet* **13**, 613–626 (2012).

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

119. Meyer, M. B., Benkusky, N. A. & Pike, J. W. Selective Distal Enhancer Control of the Mmp13 Gene Identified through Clustered Regularly Interspaced Short Palindromic Repeat (CRISPR) Genomic Deletions. *Journal of Biological Chemistry* **290**, 11093–11107 (2015).

120. Huang, J. *et al.* Dissecting super-enhancer hierarchy based on chromatin interactions. *Nat Commun* **9**, 943 (2018).

121. Farh, K. K.-H. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**, 337–343 (2015).

122. Gjoneska, E. *et al.* Conserved epigenomic signals in mice and humans reveal immune basis of Alzheimer's disease. *Nature* **518**, 365–369 (2015).

123. Mells, G. F. & Hirschfield, G. M. Making the most of new genetic risk factors – genetic and epigenetic fine mapping of causal autoimmune disease variants. *Clin Res Hepatol Gastroenterol* **39**, 408–411 (2015).

124. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).

125. 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).

126. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res* **47**, D766–D773 (2019).

127. Kheradpour, P. & Kellis, M. Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Res* **42**, 2976–2987 (2014).

128. Pinero, J. *et al.* DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. *Database* **2015**, bav028–bav028 (2015).

129. Mattingly, C. J., Colby, G. T., Forrest, J. N. & Boyer, J. L. The Comparative Toxicogenomics Database (CTD). *Environ Health Perspect* **111**, 793–795 (2003).

130. Gutiérrez-Sacristán, A. *et al.* PsyGeNET: a knowledge platform on psychiatric disorders and their genes. *Bioinformatics* **31**, 3075–3077 (2015).

131. Rehm, H. L. *et al.* ClinGen — The Clinical Genome Resource. *New England Journal of Medicine* **372**, 2235–2242 (2015).

132. Tamborero, D. *et al.* Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations. *Genome Med* **10**, 25 (2018).

133. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).

134. Kanz, C. The EMBL Nucleotide Sequence Database. *Nucleic Acids Res* **33**, D29–D33 (2004).

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

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

137. Lupiáñez, D. G. *et al.* Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions. *Cell* **161**, 1012–1025 (2015).

138. Vandin, F., Upfal, E. & Raphael, B. J. Algorithms for Detecting Significantly Mutated Pathways in Cancer. *Journal of Computational Biology* **18**, 507–522 (2011).

139. Stark, C. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res* **34**, D535–D539 (2006).

140. Das, J., Felty, Q., Poppiti, R., Jackson, R. & Roy, D. Nuclear Respiratory Factor 1 Acting as an Oncoprotein Drives Estrogen-Induced Breast Carcinogenesis. *Cells* **7**, 234 (2018).

141. Nowyhed, H. N. *et al.* ATP Binding Cassette Transporter ABCA7 Regulates NKT Cell Development and Function by Controlling CD1d Expression and Lipid Raft Content. *Sci Rep* **7**, 40273 (2017).

142. Aikawa, T. *et al.* ABCA7 haplodeficiency disturbs microglial immune responses in the mouse brain. *Proceedings of the National Academy of Sciences* **116**, 23790–23796 (2019).

143. Jones, K. A., Kadonaga, J. T., Luciw, P. A. & Tjian, R. Activation of the AIDS Retrovirus Promoter by the Cellular Transcription Factor, Sp1. *Science (1979)* **232**, 755–759 (1986).

144. Fernandez-Zapico, M. E. *et al.* A functional family-wide screening of SP/KLF proteins identifies a subset of suppressors of *KRAS* -mediated cell growth. *Biochemical Journal* **435**, 529–537 (2011).

145. Dupuis-Maurin, V. *et al.* Overexpression of the Transcription Factor Sp1 Activates the OAS-RNAse L-RIG-I Pathway. *PLoS One* **10**, e0118551 (2015).

146. Ilsley, M. D. *et al.* Krüppel-like factors compete for promoters and enhancers to fine-tune transcription. *Nucleic Acids Res* **45**, 6572–6588 (2017).

147. Biffi, A. Genetic Variation and Neuroimaging Measures in Alzheimer Disease. *Arch Neurol* **67**, 677 (2010).

148. Gaiteri, C., Mostafavi, S., Honey, C. J., De Jager, P. L. & Bennett, D. A. Genetic variants in Alzheimer disease — molecular and brain network approaches. *Nat Rev Neurol* **12**, 413–427 (2016).

149. Ramos, P. S., Shedlock, A. M. & Langefeld, C. D. Genetics of autoimmune diseases: insights from population genetics. *J Hum Genet* **60**, 657–664 (2015).

150. Zheng, S. L. *et al.* Cumulative Association of Five Genetic Variants with Prostate Cancer. *New England Journal of Medicine* **358**, 910–919 (2008).

151. Wacholder, S. *et al.* Performance of Common Genetic Variants in Breast-Cancer Risk Models. *New England Journal of Medicine* **362**, 986–993 (2010).

152. Zhang, B., Beeghly-Fadiel, A., Long, J. & Zheng, W. Genetic variants associated with breast-cancer risk: comprehensive research synopsis, meta-analysis, and epidemiological evidence. *Lancet Oncol* **12**, 477–488 (2011).

153. Nicolae, D. L. *et al.* Trait-Associated SNPs Are More Likely to Be eQTLs: Annotation to Enhance Discovery from GWAS. *PLoS Genet* **6**, e1000888 (2010).

154. Smith, A. K. *et al.* Methylation quantitative trait loci (meQTLs) are consistently detected across ancestry, developmental stage, and tissue type. *BMC Genomics* **15**, 145 (2014).

155. McVicker, G. *et al.* Identification of Genetic Variants That Affect Histone Modifications in Human Cells. *Science (1979)* **342**, 747–749 (2013).

156. Nica, A. C. & Dermitzakis, E. T. Expression quantitative trait loci: present and future. *Philosophical Transactions of the Royal Society B: Biological Sciences* **368**, 20120362 (2013).

157. Wainberg, M. *et al.* Opportunities and challenges for transcriptome-wide association studies. *Nat Genet* **51**, 592–599 (2019).

158. Avsec, Ž. *et al.* Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nat Genet* **53**, 354–366 (2021).

159. Võsa, U. *et al.* Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nat Genet* **53**, 1300–1310 (2021).

160. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**, 580–585 (2013).

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

162. Wang, H., Huang, B. & Wang, J. Predict long-range enhancer regulation based on protein–protein interactions between transcription factors. *Nucleic Acids Res* **49**, 10347–10368 (2021).

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

164. Wen, X., Luca, F. & Pique-Regi, R. Cross-Population Joint Analysis of eQTLs: Fine Mapping and Functional Annotation. *PLoS Genet* **11**, e1005176 (2015).

165. Gasperini, M. *et al.* A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens. *Cell* **176**, 377-390.e19 (2019).

166. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A Simple New Approach to Variable Selection in Regression, with Application to Genetic Fine Mapping. *J R Stat Soc Series B Stat Methodol* **82**, 1273–1300 (2020).

167. Wen, X., Pique-Regi, R. & Luca, F. Integrating molecular QTL data into genome-wide genetic association analysis: Probabilistic assessment of enrichment and colocalization. *PLoS Genet* **13**, e1006646 (2017).

168. Hormozdiari, F., Kostem, E., Kang, E. Y., Pasaniuc, B. & Eskin, E. Identifying Causal Variants at Loci with Multiple Signals of Association. *Genetics* **198**, 497–508 (2014).

169. Brown, A. A. *et al.* Predicting causal variants affecting expression by using whole-genome sequencing and RNA-seq from multiple human tissues. *Nat Genet* **49**, 1747–1751 (2017).

170. Kircher, M. *et al.* Saturation mutagenesis of twenty disease-associated regulatory elements at single base-pair resolution. *Nat Commun* **10**, 3583 (2019).

171. Teslovich, T. M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707–713 (2010).

172. Lotem, J. *et al.* Runx3-mediated Transcriptional Program in Cytotoxic Lymphocytes. *PLoS One* **8**, e80467 (2013).

173. Vilagos, B. *et al.* Essential role of EBF1 in the generation and function of distinct mature B cell types. *Journal of Experimental Medicine* **209**, 775–792 (2012).

174. Yang, R. *et al.* Human T-bet Governs Innate and Innate-like Adaptive IFN-γ Immunity against Mycobacteria. *Cell* **183**, 1826-1847.e31 (2020).

175. Bickmore, W. A. The Spatial Organization of the Human Genome. *Annu Rev Genomics Hum Genet* **14**, 67–84 (2013).

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

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

178. Xiong, K., Zhang, R. & Ma, J. scGHOST: identifying single-cell 3D genome subcompartments. *Nat Methods* **21**, 814–822 (2024).

179. Dekker, J. *et al.* Spatial and temporal organization of the genome: Current state and future aims of the 4D nucleome project. *Mol Cell* **83**, 2624–2640 (2023).

180. Girelli, G. *et al.* GPSeq reveals the radial organization of chromatin in the cell nucleus. *Nat Biotechnol* **38**, 1184–1193 (2020).

181. Till, J. E. & McCulloch, E. A. Hemopoietic stem cell differentiation. *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer* **605**, 431–459 (1980).

182. Snijder, B. & Pelkmans, L. Origins of regulated cell-to-cell variability. *Nat Rev Mol Cell Biol* **12**, 119–125 (2011).

183. Sharp, A. J. *et al.* DNA methylation profiles of human active and inactive X chromosomes. *Genome Res* **21**, 1592–1600 (2011).

184. Brown, C. J., Carrel, L. & Willard, H. F. Expression of Genes from the Human Active and Inactive X Chromosomes. *The American Journal of Human Genetics* **60**, 1333–1343 (1997).

185. Knight, P. A. & Ruiz, D. A fast algorithm for matrix balancing. *IMA Journal of Numerical Analysis* **33**, 1029–1047 (2013).

186. Imakaev, M. *et al.* Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat Methods* **9**, 999–1003 (2012).

187. Xiong, K. & Ma, J. Revealing Hi-C subcompartments by imputing inter-chromosomal chromatin interactions. *Nat Commun* **10**, 5069 (2019).

188. Kaul, A., Bhattacharyya, S. & Ay, F. Identifying statistically significant chromatin contacts from Hi-C data with FitHiC2. *Nat Protoc* **15**, 991–1012 (2020).

189. Zhang, Y., Liu, W., Lin, Y., Ng, Y. K. & Li, S. Large-scale 3D chromatin reconstruction from chromosomal contacts. *BMC Genomics* **20**, 186 (2019).

190. Li, F.-Z. *et al.* Chromatin 3D structure reconstruction with consideration of adjacency relationship among genomic loci. *BMC Bioinformatics* **21**, 272 (2020).

191. Hirata, Y., Oda, A., Ohta, K. & Aihara, K. Three-dimensional reconstruction of single-cell chromosome structure using recurrence plots. *Sci Rep* **6**, 34982 (2016).

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

193. Boninsegna, L. *et al.* Integrative genome modeling platform reveals essentiality of rare contact events in 3D genome organizations. *Nat Methods* **19**, 938–949 (2022).

194. Quinodoz, S. A. *et al.* Higher-Order Inter-chromosomal Hubs Shape 3D Genome Organization in the Nucleus. *Cell* **174**, 744-757.e24 (2018).

195. Zheng, Y., Shen, S. & Keleş, S. Normalization and de-noising of single-cell Hi-C data with BandNorm and scVI-3D. *Genome Biol* **23**, 222 (2022).

196. Maass, P. G., Barutcu, A. R., Weiner, C. L. & Rinn, J. L. Inter-chromosomal Contact Properties in Live-Cell Imaging and in Hi-C. *Mol Cell* **69**, 1039-1045.e3 (2018).

197. Maass, P. G., Barutcu, A. R. & Rinn, J. L. Interchromosomal interactions: A genomic love story of kissing chromosomes. *Journal of Cell Biology* **218**, 27–38 (2019).

198. Segal, M. R. Can 3D diploid genome reconstruction from unphased Hi-C data be salvaged? *NAR Genom Bioinform* **4**, (2022).

APPENDIX A

SUPPLEMENTARY FIGURES FOR CHAPTER 2

Immune-associated phenotypes	
Index	Phenotypes
1	Leukemia
2	Lymphocytic
3	Chronic
4	B-Cell
5	Crohn Disease
6	Inflammatory Bowel Diseases
7	Leukocyte
8	Blood Cells
9	Colitis
10	Ulcerative

Figure A. 1. Summary of Immune-associated phenotypes to filter GWAS SNPs and disease-associated genes.

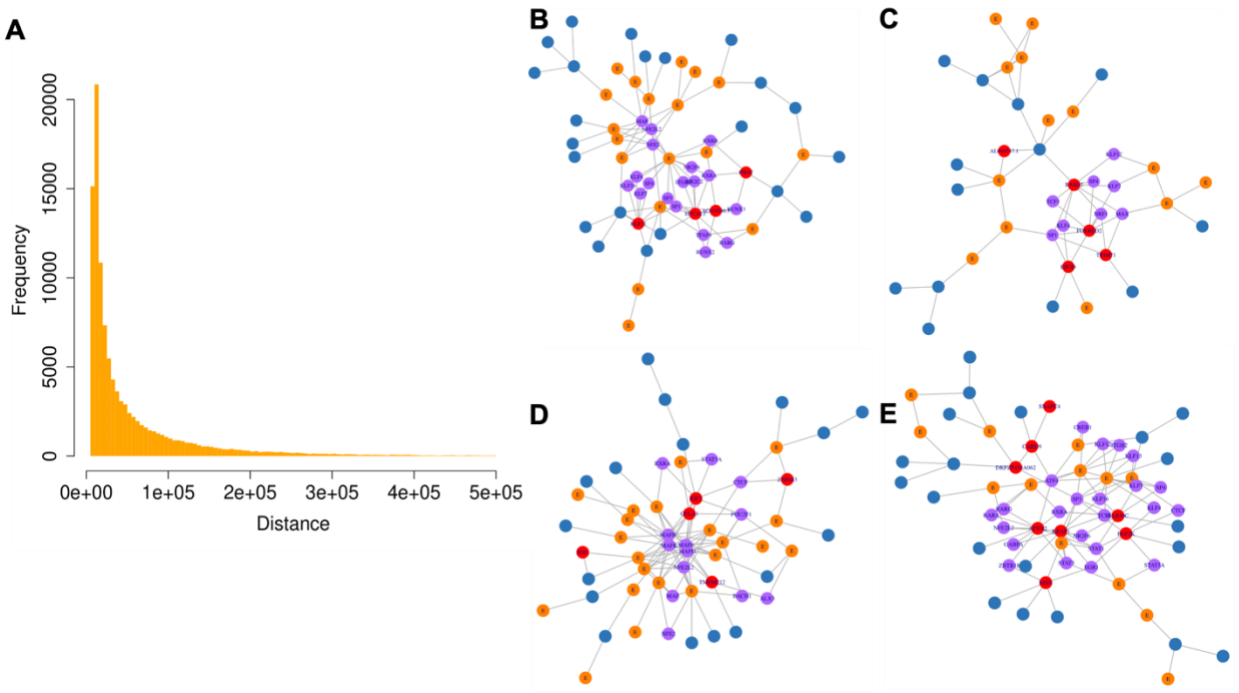

Figure A. 2. Constructed regulatory sub-networks using K562 ChIA-PET dataset. (A) Distance distribution of long-range chromatin interactions in K562. (B-E) Examples of regulatory sub-networks based on K562 ChIA-PET data. Nodes are annotated as expressed promoters (red), active enhancers (orange), other elements (blue), and TF nodes (purple).

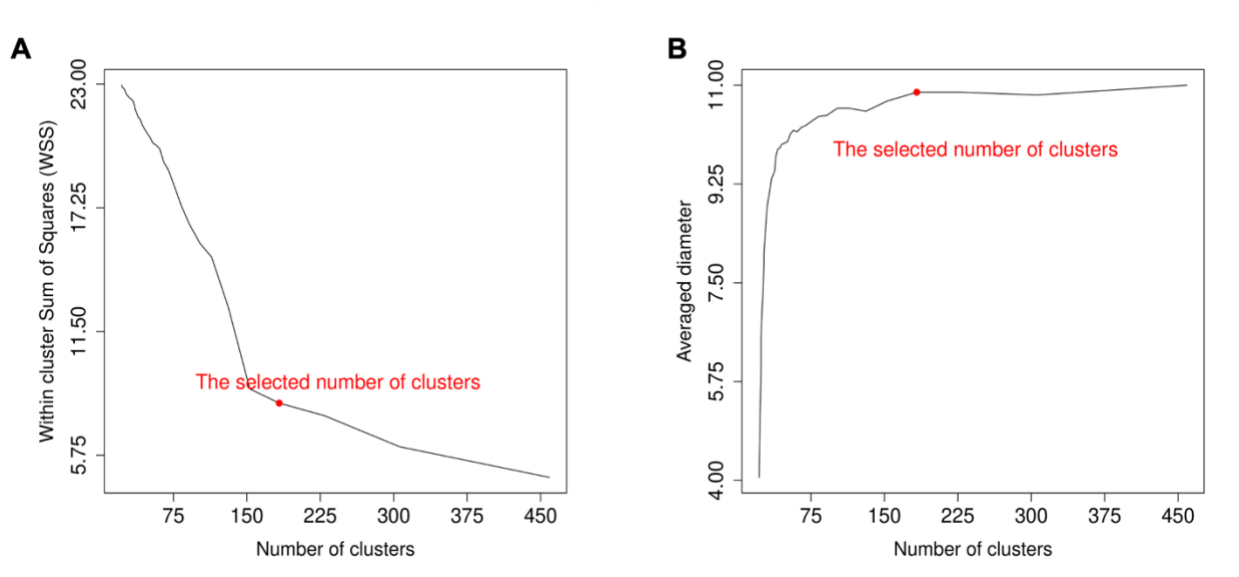

Figure A. 3. Determination of the number of clusters of 3D chromatin modules using (**A**) elbow methods and (**B**) averaged degree of the expanded regulatory sub-networks. The tested number of clusters of 3D chromatin module range from 0 to 450. The hierarchical clustering tree is cut with different numbers of clusters. (**A**) Within cluster Sum of Squares (WSS) is calculated for each cluster assignment to find an optimal number of lusters. (**B**) Expanded regulatory sub-networks are constructed based on different clustering assignments, and the averaged diameter are calculated across all regulatory sub-networks to indicate the selection on the number of clusters. The selected number of clusters is highlighted as red dot.

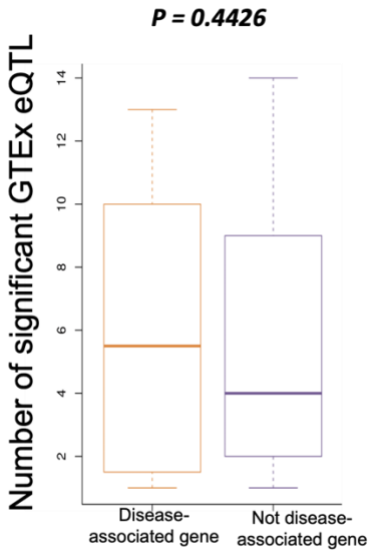

Figure A. 4. eQTLs in neighboring nodes of disease-associated genes in regulatory sub-networks. Disease-associated genes are annotated by known disease-gene associations from DisGeNet. The significant GTEx eQTLs are overlaid with the sub-networks. Only eQTLs contained by the neighboring nodes of genes are considered. Box-plot shows the number of significant GTEx eQTLs contained by the neighboring nodes (Y axis) of each disease-associated gene and non-disease-associated gene (X axis). P-value is calculated based on one-sided Student's' t-test.

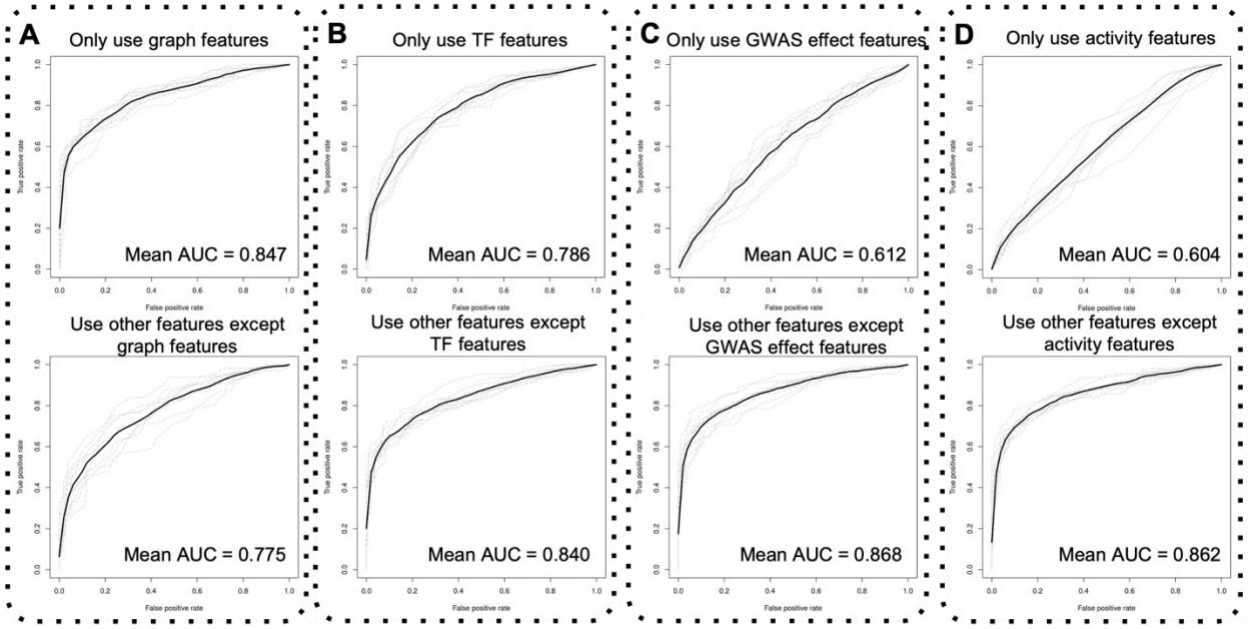

Figure A. 5. Performance of disease-associated gene prediction using APRIL regulatory sub-networks based on label propagation using different sets of features. Graph features and TF features show important role in prediction. (**A**) Compare the performance of using only graph features and without graph features; (**B**) Compare the performance of using only TF features and without TF features; (**C**) Compare the performance of using only GWAS effect features and without GWAS effect features; (**D**) Compare the performance of using only activity features and without activity features.

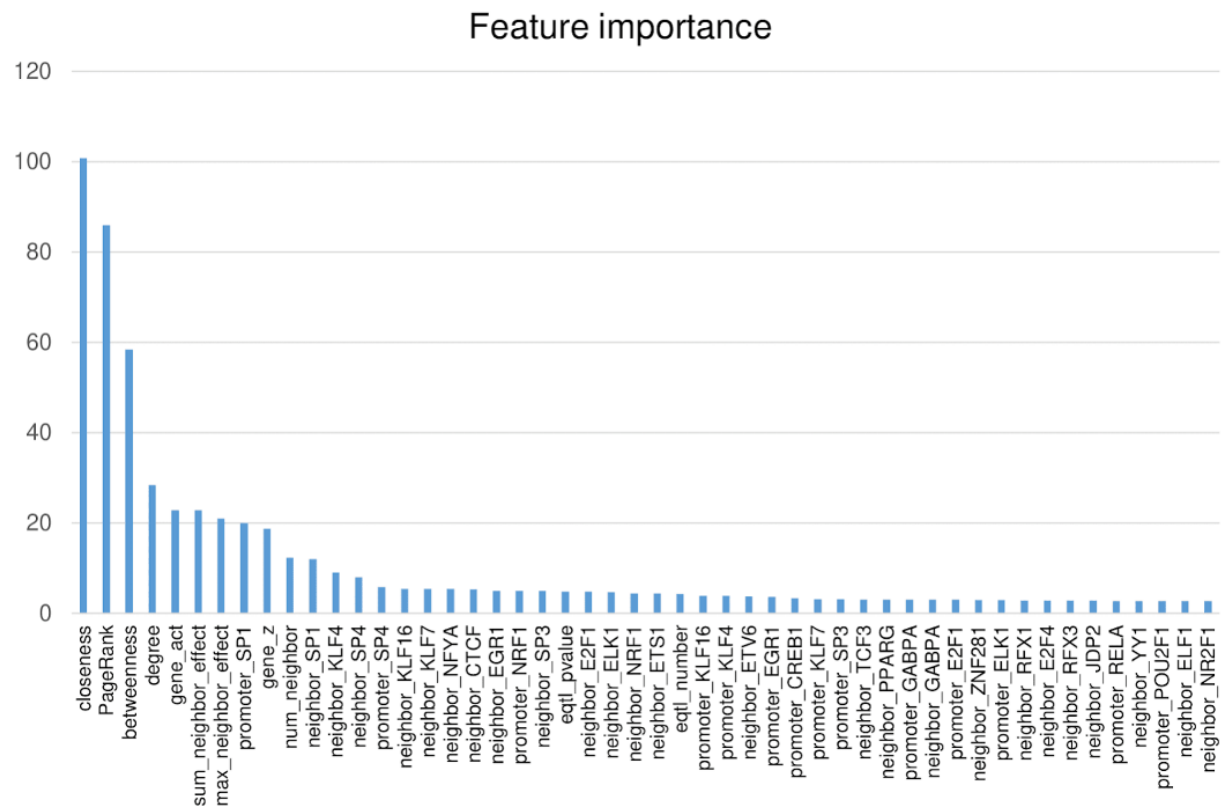

Figure A. 6. Bar-plot of feature importance for Top 50 features in the random forest model. Graph features have highest importance among all features. The GWAS effect sizes of neighboring nodes, gene activity, and specific neighboring TF nodes also have high feature importance.

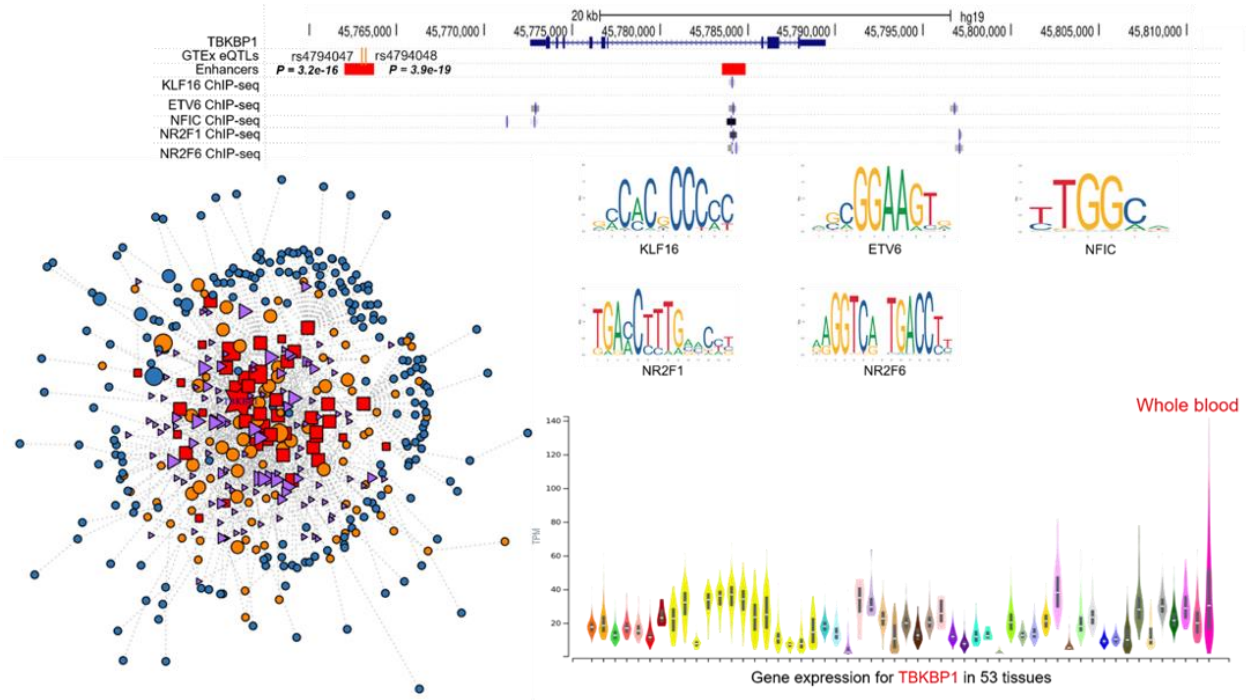

Figure A. 7. Example of newly discovered disease-associated gene TBKBP1. TBKBP1 has high expression in whole blood among 53 different tissues from GTEx. Besides, TBKBP1 has KLF16, ETV6, NFIC, NR2F1 and NR2F6 motifs validated by ChIP-seq data. Also, the neighboring enhancer linked to TBKBP1 in the APRIL sub-networks contains 2 significant GTEx eQTLs (*P=3.2e-16, P=3.9e-19*).

APPENDIX B

SUPPLEMENTARY FIGURES FOR CHAPTER 3

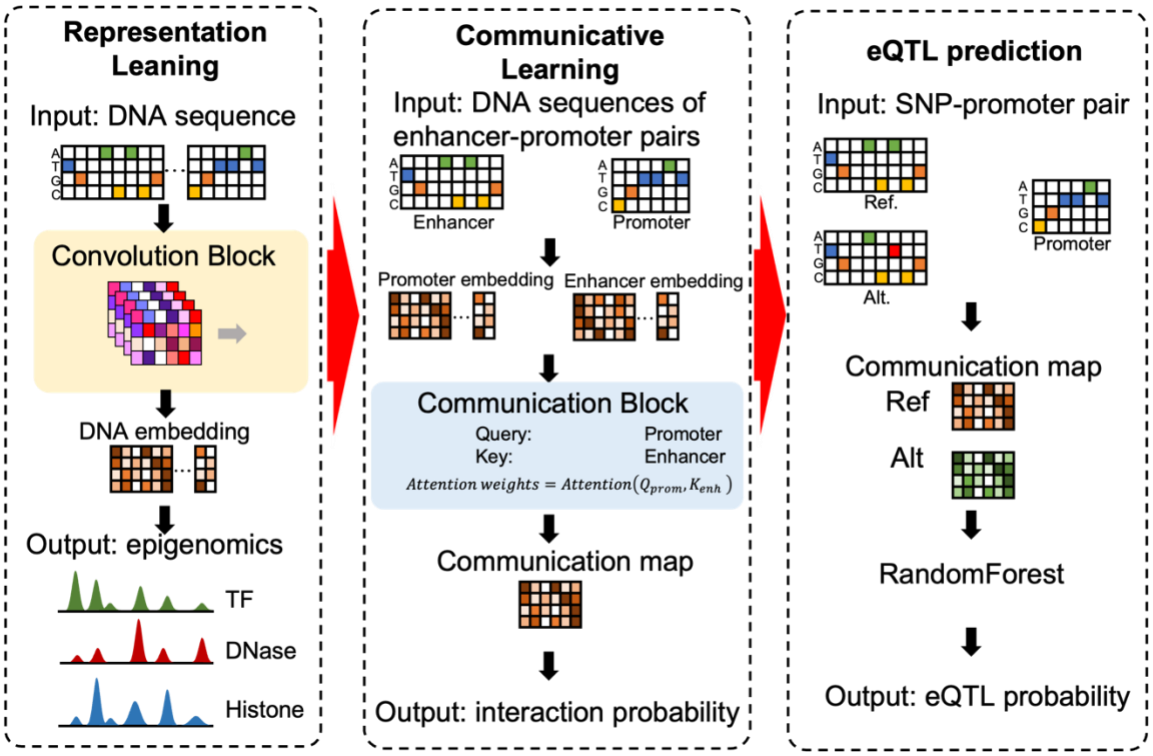

Figure B. 1. Overview of 3-stage 3DVariantVision. Stage 1 (left) uses representation learning to generate sequence embedding using CNN by predicting epigenomic peaks including TF bindings, histone modifications, DNase peaks. Stage 2 (middle) transfers the learned sequence embedding from stage 1 to generate communication maps (joint features) of enhancer-promoter pairs using communicative learning by predicting enhancer-promoter interactions. Stage 3 (right) transfers the learned communicative model to generated communication maps of the reference genome and alternative genome of SNPs. The random forest model is used to predict eQTLs based on the communication maps.

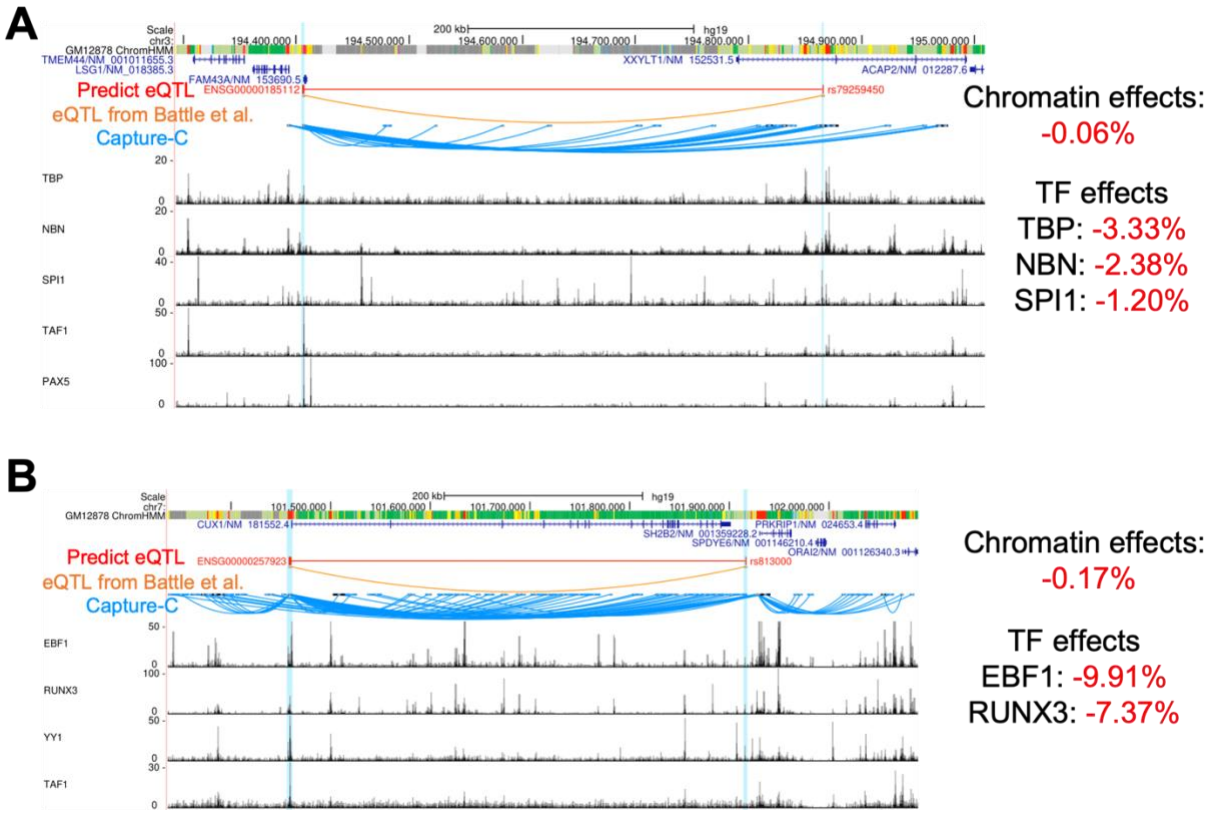

Figure B. 2. Examples of predicted eQTLs based on 3DVariantVision in whole blood tissue. (**A**) eQTL *rs79259450* is supported by Geuvadis with *p-value = 3.7x10-11*; (**B**) eQTL *rs813000* is supported by Battle with *p-value = 5.0x10-37*. Both predicted eQTLs are supported by Capture-C chromatin interactions. The right panel shows the predicted TF binding effects and chromatin interaction effects of the corresponding SNPs, which were calculated by the differences of scores between the alternative allele and the reference allele normalized by dividing the reference scores.

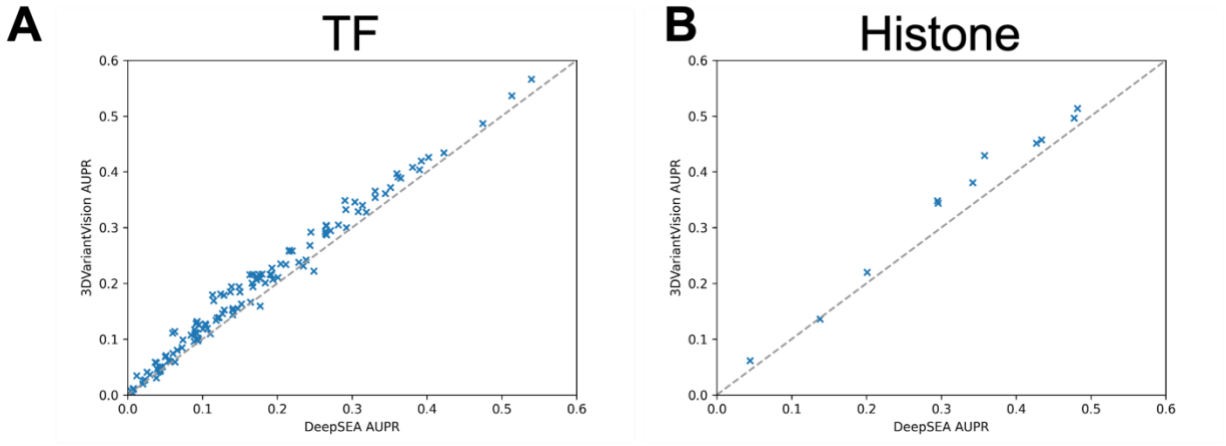

Figure B. 3. AUPR comparison of epigenomic peak predictions between 3DVariantVision and DeepSEA based on (**A**) TF binding sites, and (**B**) Histone modification peaks. Each point represents one epigenomic peak and x, y-axis represent the AUPR of DeepSEA and 3DVariantVision, respectively.

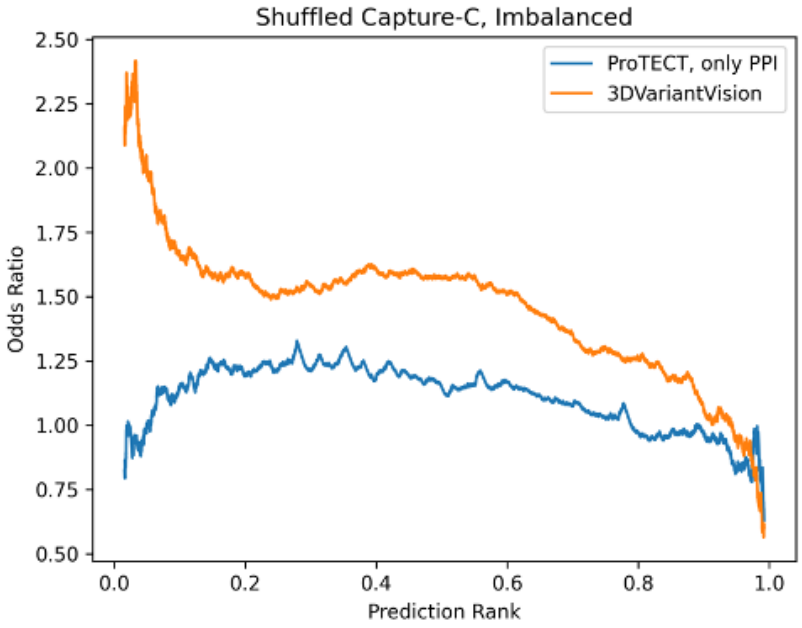

Figure B. 4. Odds ratio comparison of chromatin interaction predictions between 3DVariantVision and ProTECT using shuffled enhancer-promoter pairs based on Capture-C. The higher odds ratio of 3DVariantVision demonstrates the ability to capture the joint features between enhancer and promoter.

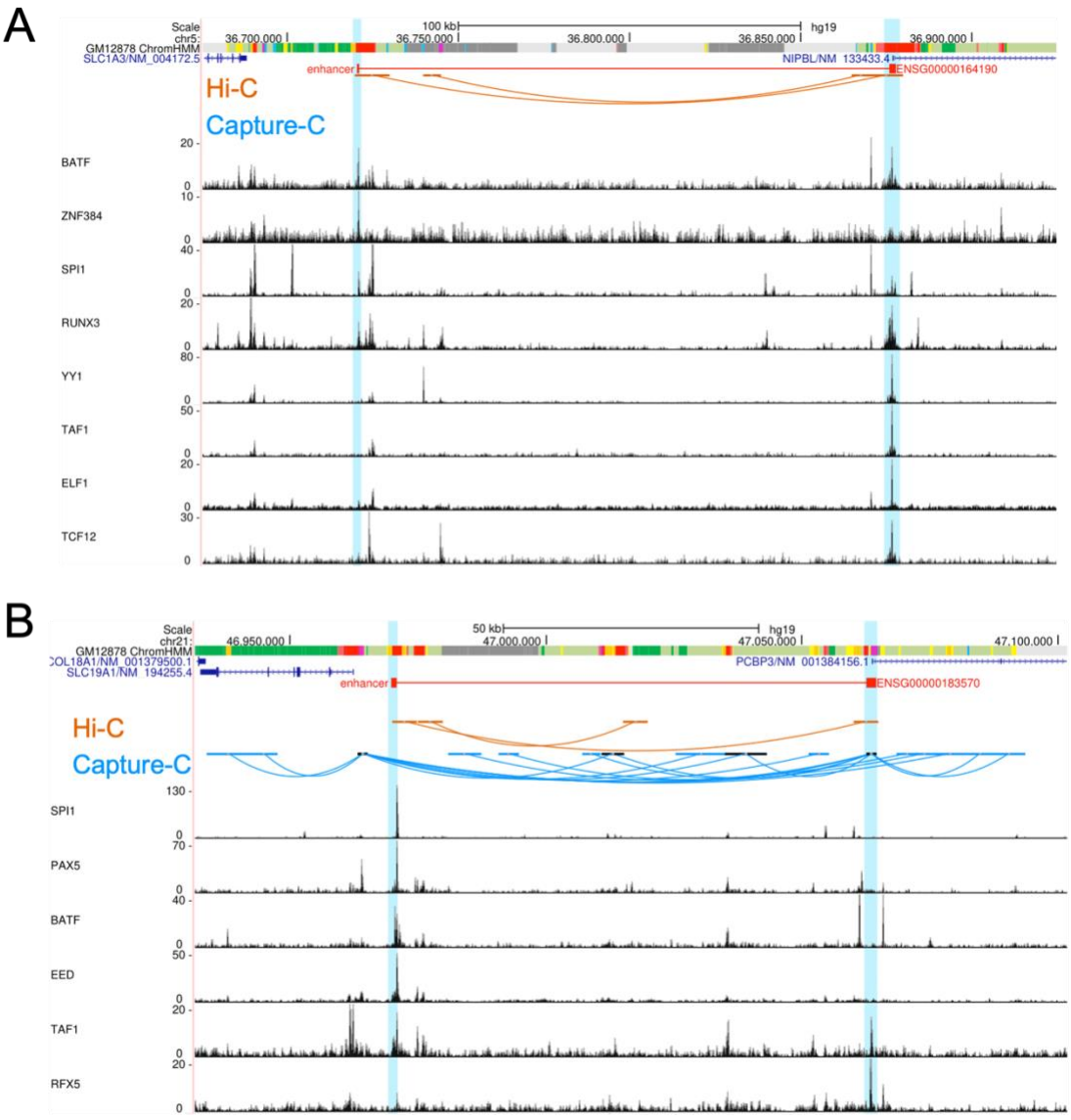

Figure B. 5. Examples of predicted long-range enhancer-promoter interactions based on 3DVariantVision in GM12878 cell line. (**A-B**) The newly discovered EPI missed by Capture-C is validated by Hi-C.

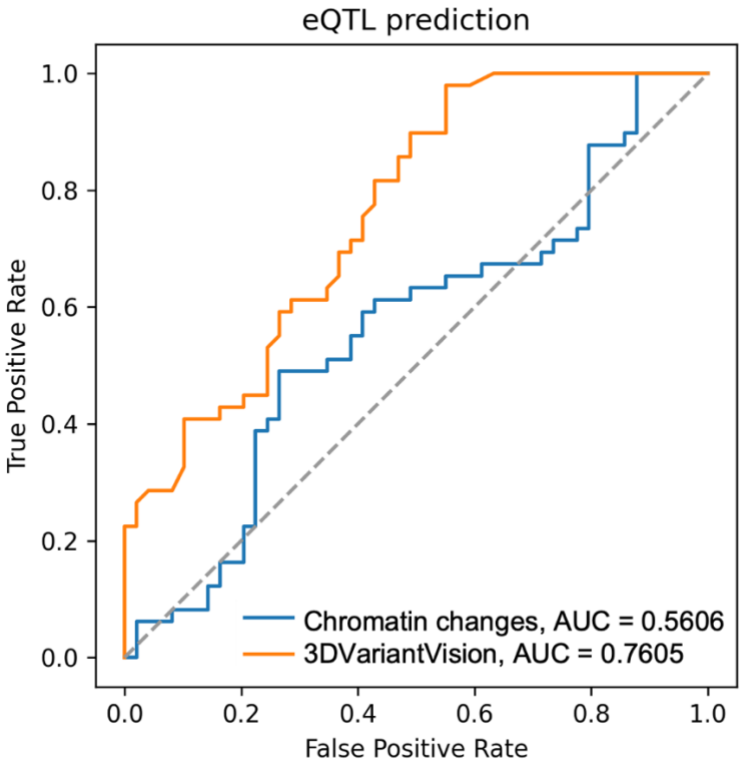

Figure B. 6. ROC of eQTL prediction based on chromatin interaction changes and 3DVariantVision on SuSIE fine-mapped eQTL dataset.

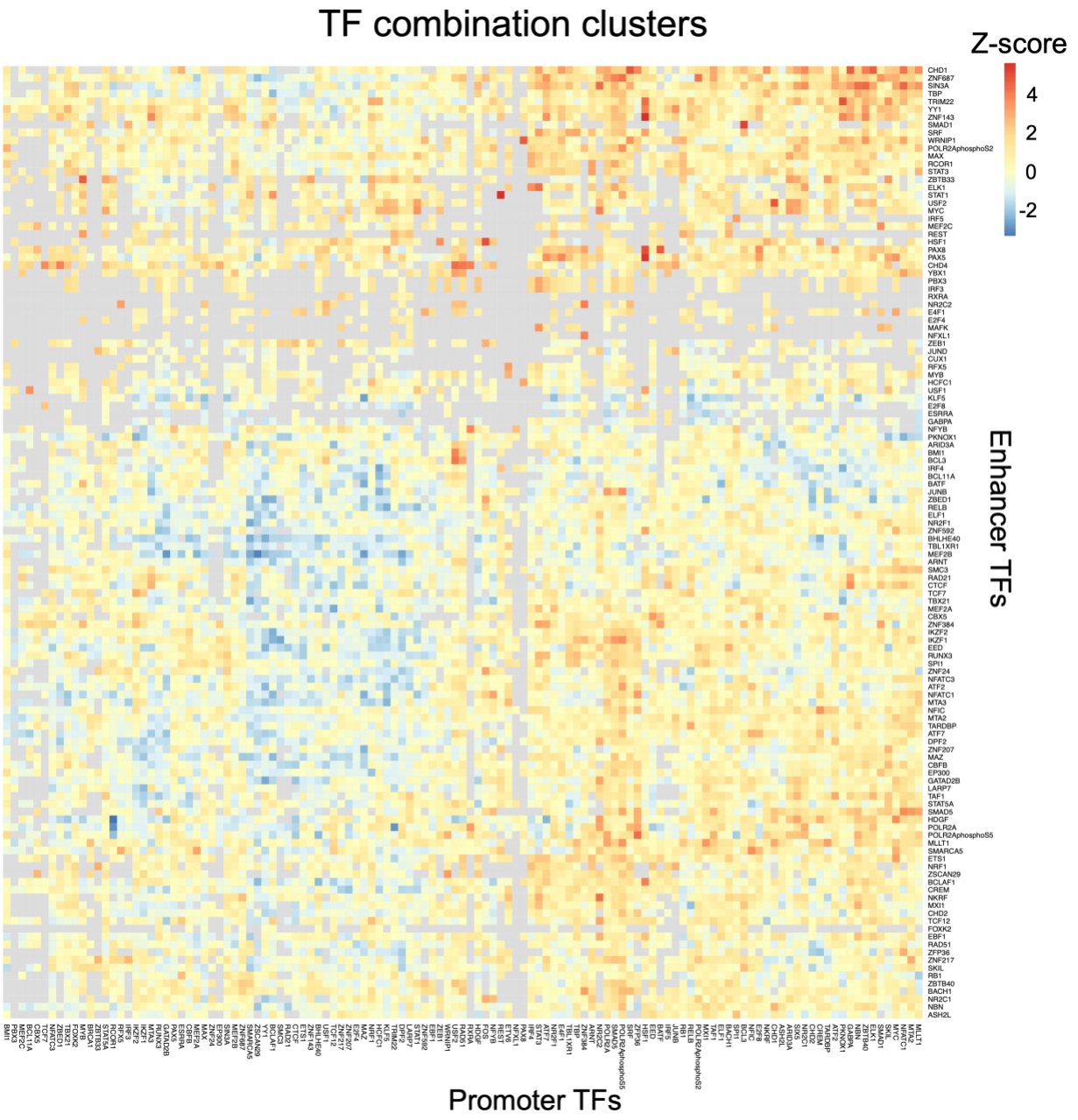

Figure B. 7. Enrichment of TF combinations prioritized by 3DVariantVision in enhancer-promoter interactions from Capture-C. The enrichment of each TF combination was quantified by the z-scores of the occurrences in top-weighted tiny region pairs of enhancer-promoter interactions against random tiny region pairs.

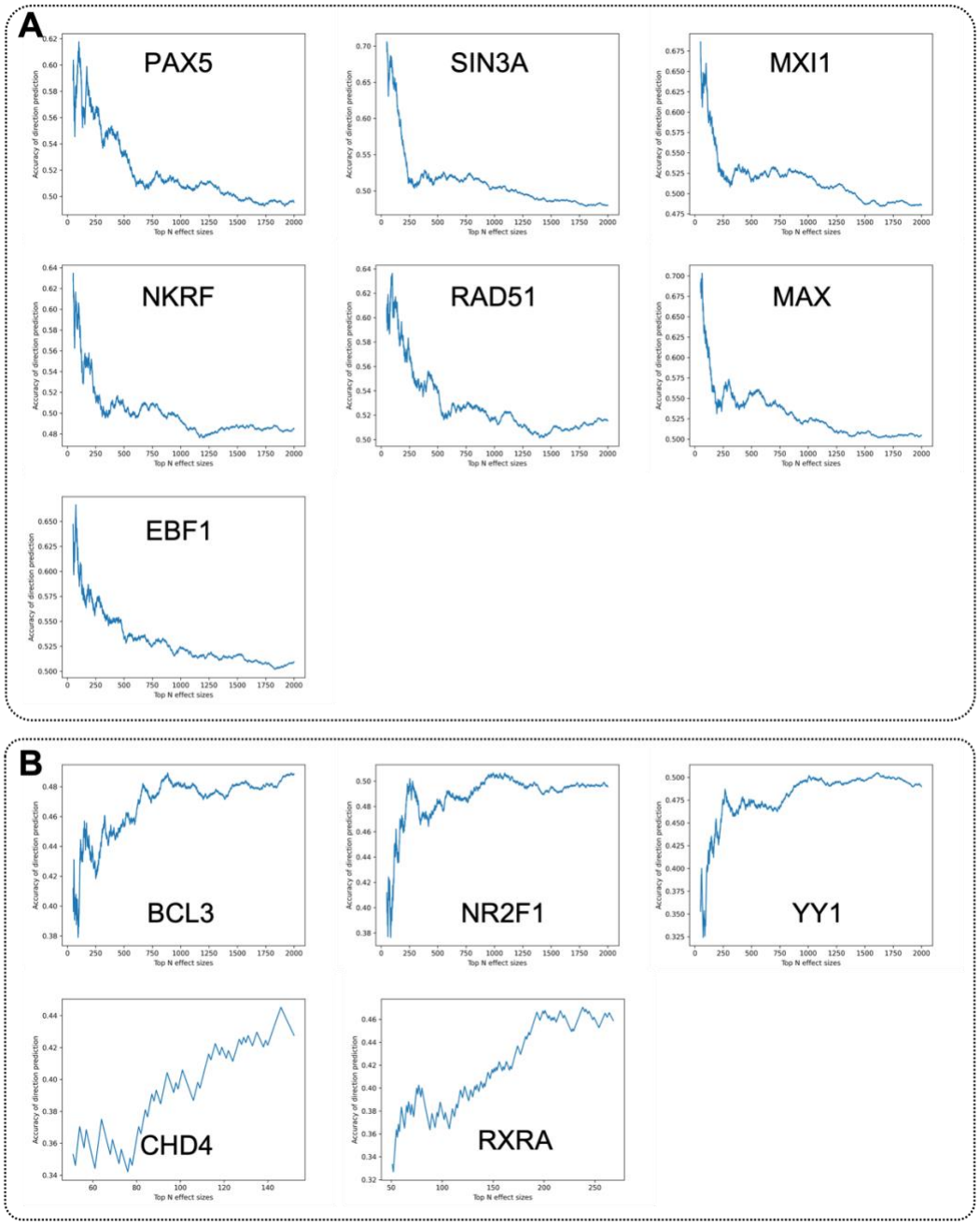

Figure B. 8. Effect direction consistency of eQTLs with different TF bindings. The effect direction consistency is calculated by the fraction of consistency between the effect size direction of top effect eQTLs and the direction of the corresponding chromatin changing from 3DVariantVision. (**A**) and (**B**) show several represented examples of positive and negative consistent TF binding eQTL groups, respectively.

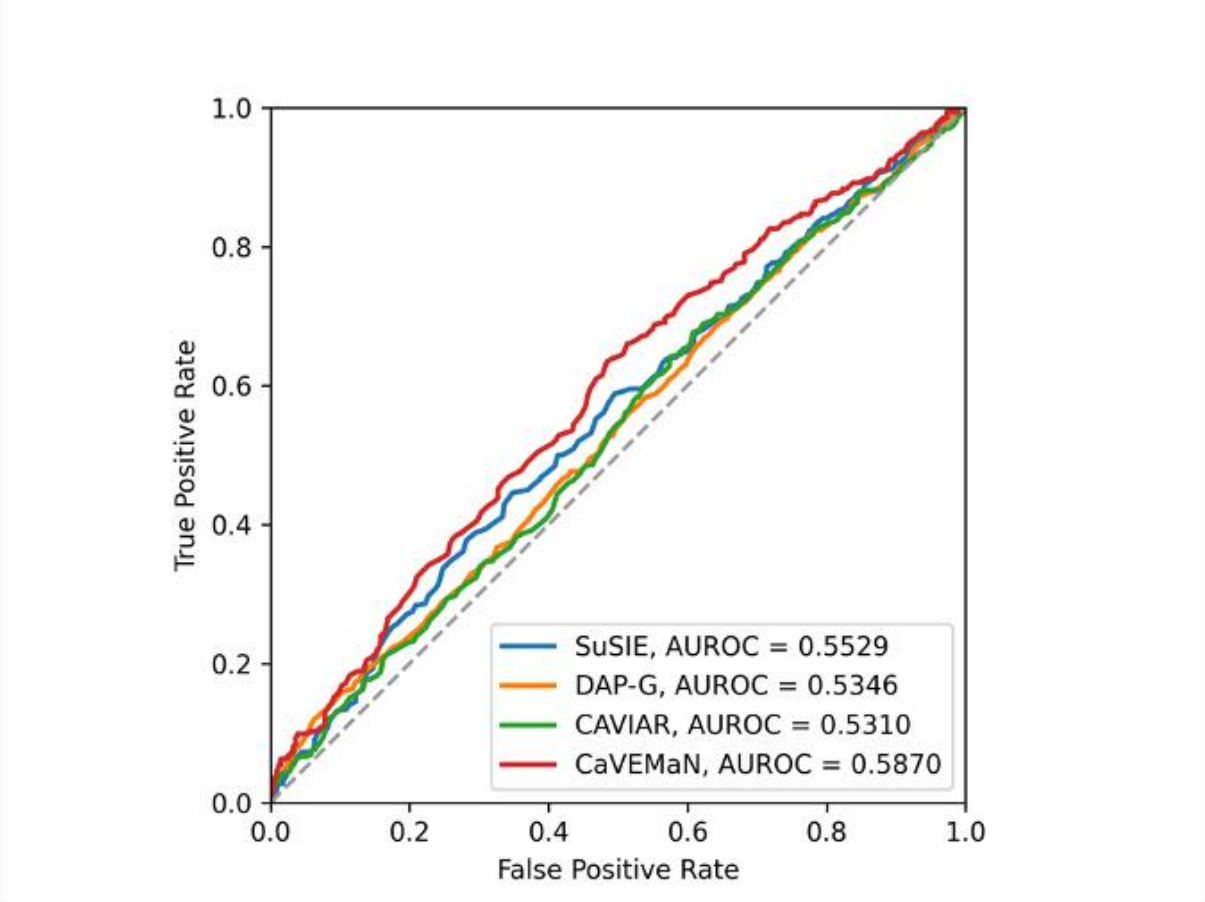

Figure B. 9. Performance of identifying target genes of eQTLs from background protein-coding genes. Four fine-mapped eQTL datasets were used to plot the ROC curves.

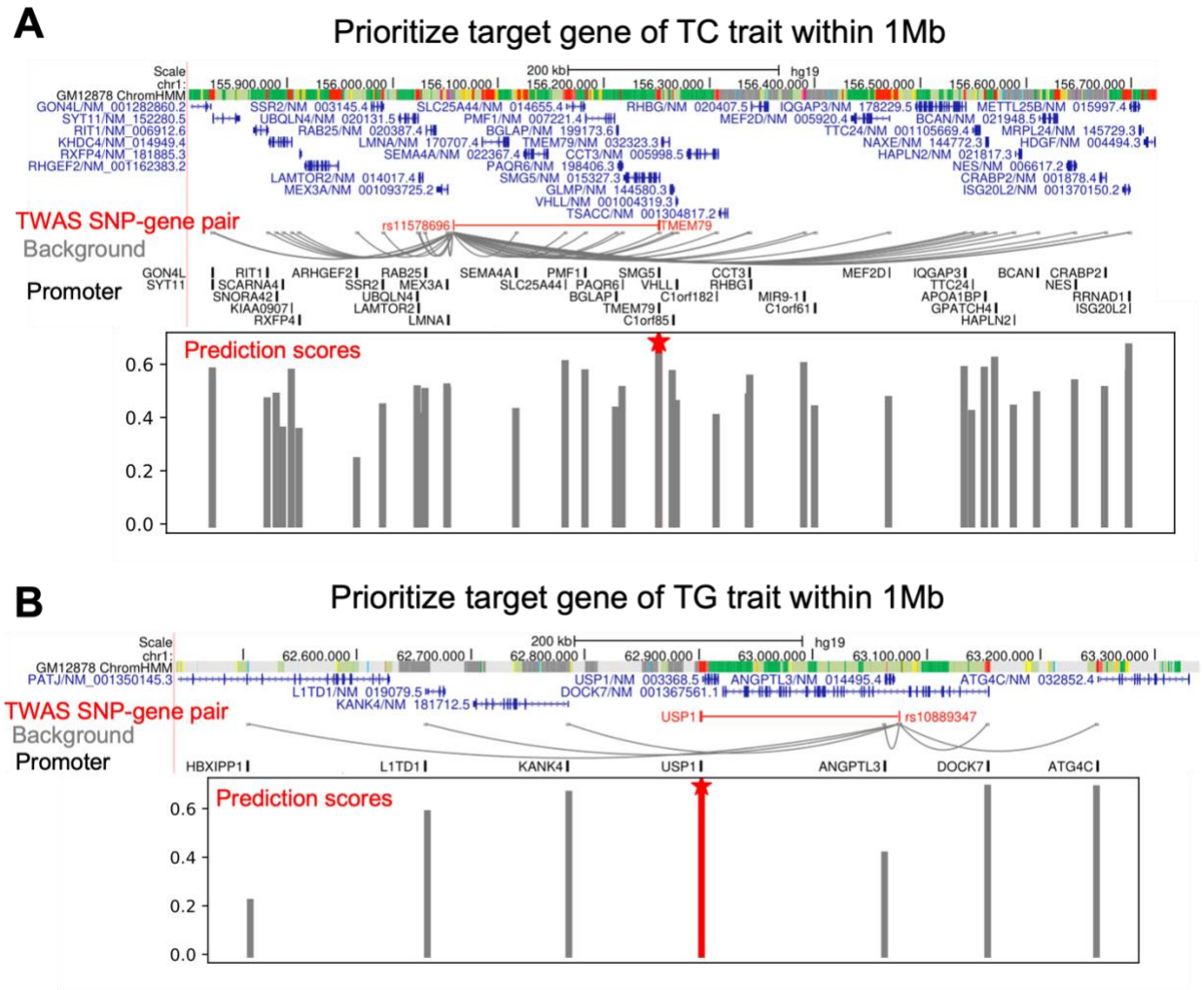

Figure B. 10. Examples of prioritizing target genes of TWAS SNPs based on 3DVariantVision. (**A**) and (**B**) is the examples of TC and TG trait associations, respectively.

APPENDIX C

SUPPLEMENTARY FIGURES FOR CHAPTER 4

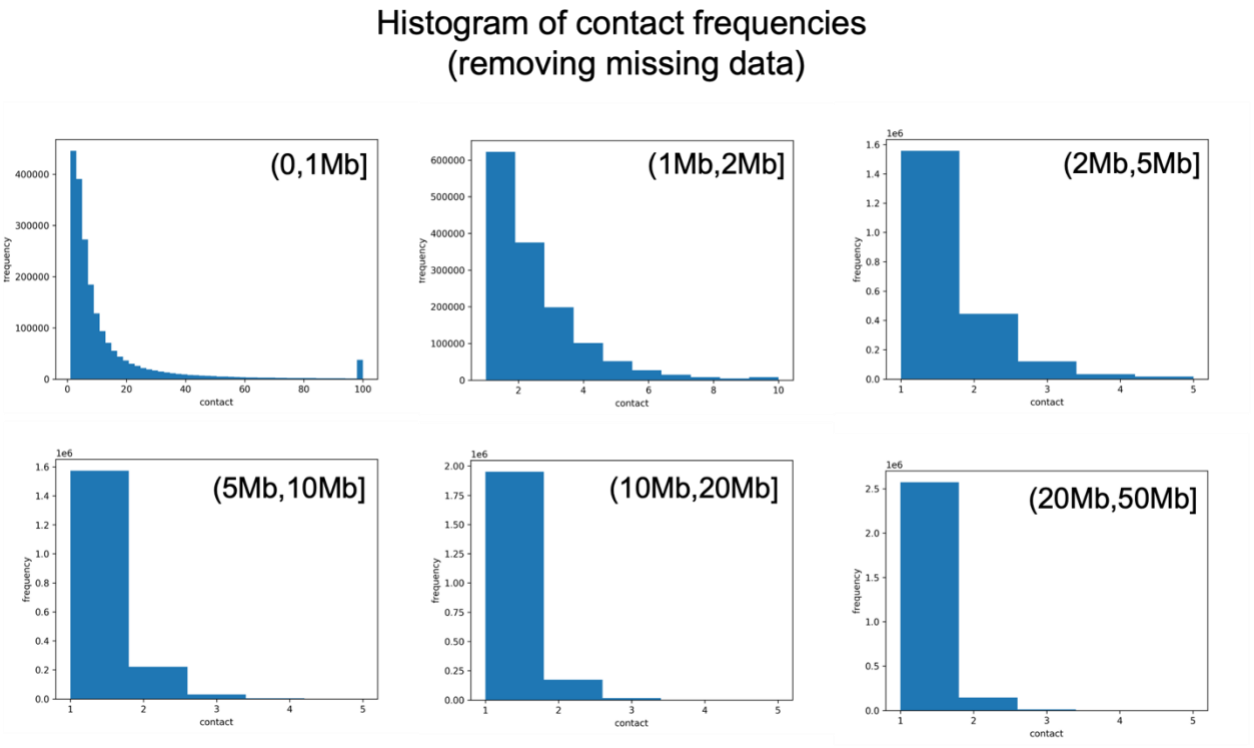

Figure C. 1. Hi-C contact frequencies at different genomic distance bands after removing missing data. The contacts tend to be binary values at long genomic distance bands (>5Mb).

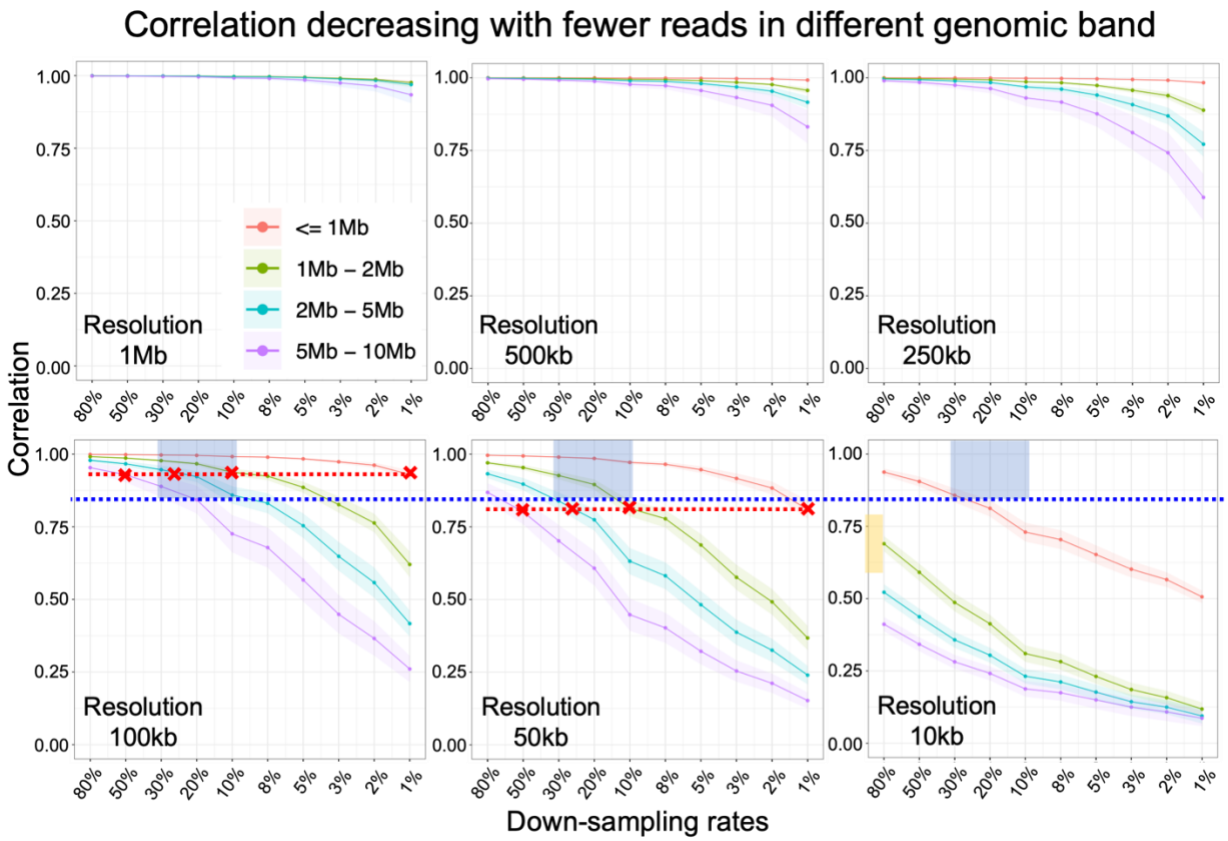

Figure C. 2. The effect of read depth on the reproducibility of Hi-C contact maps across different genomic distance bands and resolutions. Hi-C data from the GM12878 cell line was downsampled to different read depths, ranging from 80% to 1% of the original read depth. Spearman correlation coefficients were calculated between the downsampled and original Hi-C contact maps for each resolution and genomic distance band. Each subfigure shows the correlation curves for a specific resolution, with different colors representing different genomic distance bands. The blue dashed line indicates a correlation threshold of 0.85, and the blue shaded regions in each subfigure highlight the downsampling rate range between 30% and 10% where the correlation is above the threshold. The red dashed lines in the subfigures for 100kb and 50kb resolution demonstrate the different downsampling rates required to maintain the same correlation across different genomic distance bands.

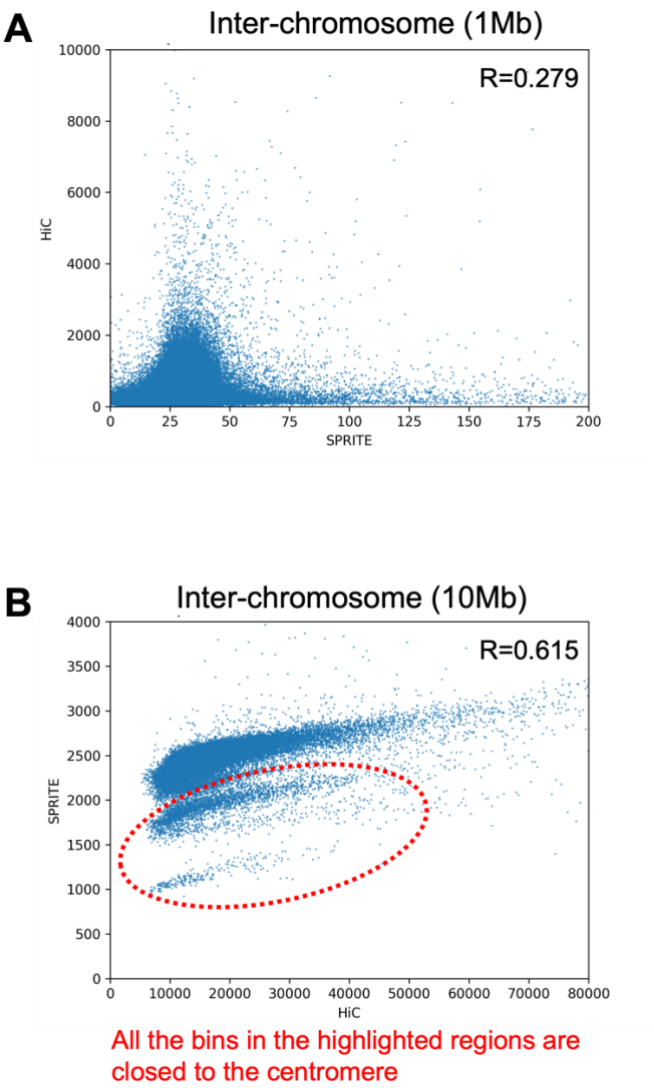

Figure C. 3. Scatter plot showing the relationship between inter-chromosomal interaction frequencies measured by Hi-C and SPRITE at 1Mb resolution (**A**) and 10Mb resolution (**B**). Each point represents a pair of genomic bins from different chromosomes. The highlighted regions in (**B**) correspond to genomic bins located close to the centromeres.

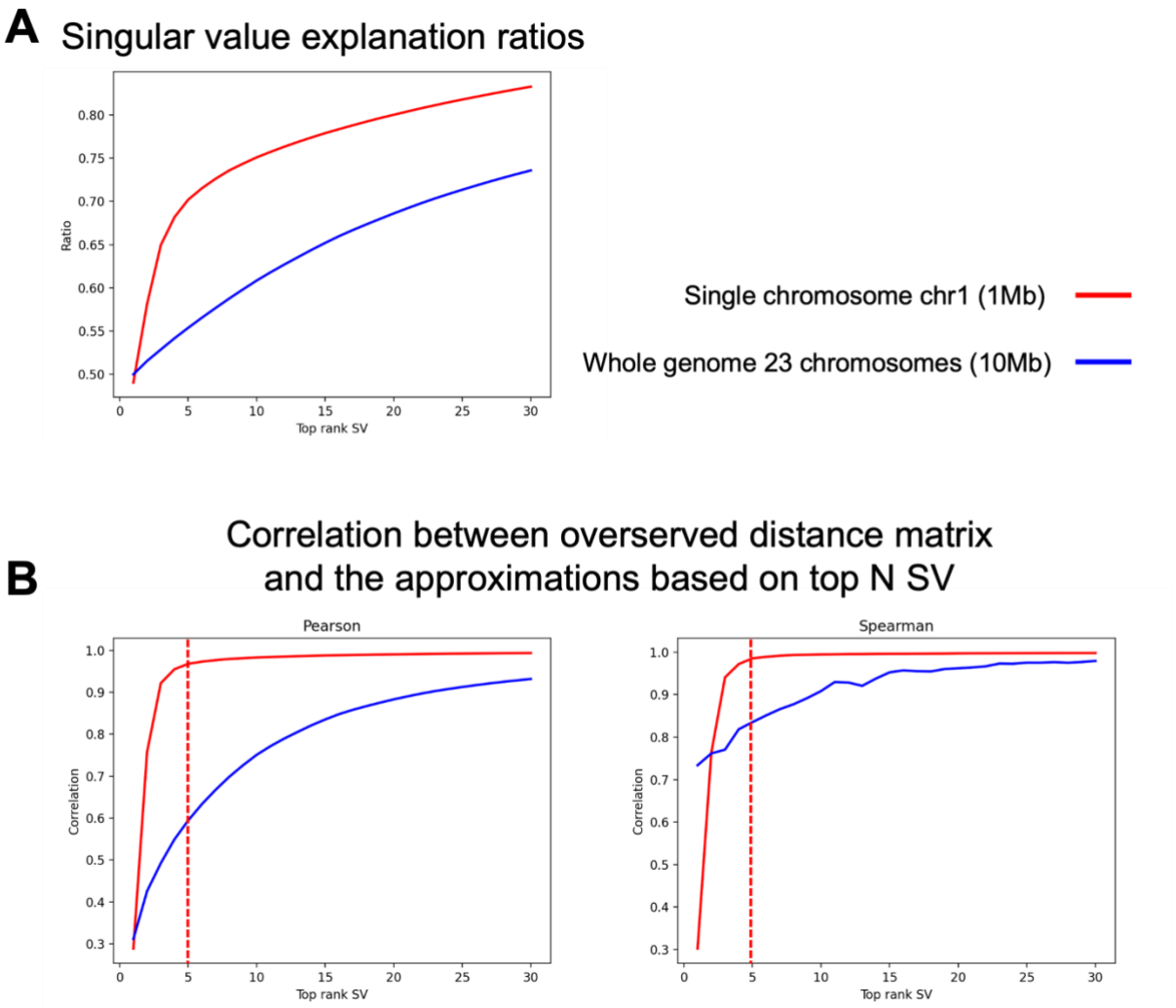

Figure C. 4. Approximation of intra-chromosomal and inter-chromosomal observed distance matrices using singular value decomposition (SVD). (**A**) Explanation ratios of the observed distance matrices for intra-chromosomal (red) and inter-chromosomal (blue) contacts using different numbers of top singular values. (**B**) Pearson correlation (left panel) and Spearman correlation (right panel) between the observed distance matrices and the approximated distance matrices using different numbers of top singular values and corresponding singular vectors for intra-chromosomal (red) and inter-chromosomal (blue) contacts.