Deciphering the Knowledge of Human Genome with Graphs

by

Fan Feng

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in the University of Michigan
2024

Doctoral Committee:

Professor Jie Liu, Chair
Professor Stephen CJ Parker
Professor Joshua Welch
Professor Jianzhi Zhang
Professor Xiaotian Zhang

Fan Feng
fanfeng@umich.edu
ORCID iD: 0000-0002-5990-312X

DEDICATION

I dedicate my dissertation to my family, friends, and all other people who helped me or cared about me during my Ph.D. studies.

ACKNOWLEDGEMENTS

More than half of my Ph.D. time at the University of Michigan was during the hard Covid time. Most of the courses, meetings, seminars, and conferences were online and a lot of my work was finished at home. Whether good or not, this experience was unique, and it even felt unreal when I started writing this dissertation - really, this is the end? Five years had already passed? Then I asked myself - did I learn a lot during my Ph.D. journey? Did I become a different person and a better researcher?

I would say yes. Although a single day is unlikely to change a person, five years is enough. When I started my Ph.D., I only knew I was interested in using algorithms to handle large amounts of biomedical data. At that time, I thought I was smart - at least above average. I was enthusiastic about and could always propose great solutions to research problems with given input and output. But during these five years, I realized science is not that easy. The "inputs" and "outputs" are not always provided, and some are not even well defined. I started to understand that an experienced researcher needs to notice gaps, ask questions, try to find solutions, and deliver the results to the community - every step is vital. It was a challenging time for me to go through these steps, and I have to show my appreciation for those who have supported me through this journey. Without their support, I would only be a programmer who knows some little tricks instead of a Ph.D. - the first step of an actual researcher.

First and foremost, I would like to thank my Ph.D. advisor, Prof. Jie Liu. His constant support, guidance, and encouragement have been valuable throughout my Ph.D. journey. He is serious and strict about my research, and also super considerate and helpful in all aspects. I still remember when I wrote my first paper, he spent a long time teaching me how to polish each paragraph and each sentence to make the results clear to the research community. He is strict on all the details, even including font size in figures and punctuation in the text, from which I learned how to set high standards for myself. I also enjoyed the discussion and meetings with Prof. Jie Liu. He always provides exceptional insights and often helps me turn preliminary ideas into well-formulated scientific questions. I have become a much better independent scientific mind and am grateful for the latitude and support you always provided me.

I would like to thank my lab mates, whose support has been a constant source of motivation. Thank you - Shuze Wang, Yuanhao Huang, Sean Moran, Zhenhao Zhang, Linghua Jiang, Xin Luo, Yicheng Tao, Yiqun Wang, Ricky Han, Tianjun Li, Feitong Tang, Yujuan Fu, Yijia Gao, Shuyuan Yang, and Dongyu Zhu. Our collaborative work and informal chats, whether conducted via Zoom or in person, provided the motivation for my work during the five years.

I would like to thank my committee members, each of whom played critical roles and provided valuable insights regarding my work - Prof. Stephen Parker, Prof. Joshua Welch, Prof. Jianzhi Zhang, and Prof. Xiaotian Zhang. You have constantly forced me to think more critically about my research, the motivations behind my work, and the questions being asked. Without your support, I would never have reached here.

I would like to thank my collaborators - Prof. Xiaotian Zhang from UT Health, Prof. Yali Dou and Dr. David Wang from USC, Prof. Anders Hansen and Dr. Clarice Hong from MIT, and Prof. Ben Hitz from Stanford. In all the collaborative work, I broadened my horizon to more diverse biomedical questions and better understood how science would change the world. I also learned a lot from each of you, which granted me the confidence to eagerly take on my journey ahead.

Lastly, I want to express my deepest gratitude to my family and friends. During the times when international travel was almost impossible, I could not meet most of them for more than two years. Although it was hard, your encouragement always played an integral role in my accomplishments. To my mom Zhimei Liu, my dad Yueke Feng, and my girlfriend (and fiancee-to-be in the near future) Jie Wang: Thank you for everything. I dedicate this PhD thesis to you.

TABLE OF CONTENTS

DEDICATION ii

ACKNOWLEDGEMENTS iii

LIST OF FIGURES viii

LIST OF TABLES xiii

LIST OF ACRONYMS xiv

ABSTRACT xvi

CHAPTER

1 Introduction 1

1.1 Central dogma and transcriptional regulation in the human genome 1
1.2 Human genome is not only a 1-D sequence 2
1.3 The complex network of the human genome from diverse data sources . . . 3

2 Revealing the 3D Structures of the Human Genome with Chromatin Conformation Capture (3C) Technology 5

2.1 Introduction of Chromatin Conformation Capture (3C) technology and chromatin contact maps 6
2.1.1 General concept of 3C technology 6
2.1.2 Hi-C technology 6
2.1.3 Micro-C technology 6
2.1.4 Hierarchical structural features revealed by chromatin contact maps 7
2.2 The relationships between transcriptional regulation and 3D chromatin organization 8
2.3 CTCF-mediated chromatin loops regulate fetal hemoglobin expression[55] . 9
2.3.1 Introduction: human β-globin and hemoglobinopathies 9
2.3.2 3'HS1 CTCF binding site in human β-globin locus regulates fetal hemoglobin expression 10
2.3.3 Summary and discussion 12
2.4 A Toolbox for analyzing single-cell Hi-C data: scHiCTools[79] 12
2.4.1 Overview of the method 13
2.4.2 Details of scHi-C similarity calculation 15

2.4.3 Benchmarking: scHiCTools calculates the similarity among single-cell Hi-C contact maps and produces satisfactory projections 17
2.4.4 Availability and future directions 20
2.5 An algorithm for identifying stripes from chromatin contact maps 20
2.5.1 Stripes in chromatin contact maps 20
2.5.2 Overview of quagga 21
2.5.3 Quagga detects stripes from publicly available chromatin contact maps 24
2.5.4 Quagga detects stripes related to CTCF-cohesin extrusion 25
2.5.5 Conclusion 27

3 Connecting High-resolution 3D Chromatin Organization with Epigenomics 28

3.1 Introduction 28
3.2 A deep learning model imputing high-resolution chromatin contact maps . 30
3.2.1 Detailed model architecture 30
3.3 Evaluation of the deep learning algorithm 33
3.3.1 Accurately predicting high-resolution chromatin contact maps . . . 33
3.3.2 Factors influencing CAESAR's performance 34
3.3.3 Recapitulating CRISPRi-validated enhancer activities 36
3.3.4 Recovering eQTL-gene interactions 36
3.4 Imputing high-resolution contact maps for more than 90 human samples . . 41
3.5 Identifying epigenomic features relevant to fine-scale 3D chromatin organization 41
3.6 Discussion and future directions 44

4 GenomicKB: A Knowledge Graph for the Human Genome 46

4.1 Background: why do we need a knowledge graph? 46
4.2 Building GenomicKB 48
4.2.1 Collecting data for GenomicKB 48
4.2.2 Schema, identity, and ontology in GenomicKB 49
4.3 GenomicKB supports graph-based relational queries 51
4.4 Applications of GenomicKB 51
4.4.1 GenomicKB simplifies cross-modality analysis as queries over the knowledge graph 51
4.4.2 GenomicKB encodes positional relations among different genomic entities 52
4.4.3 GenomicKB reconciles consensus or conflicting data sources of the same problem 54
4.5 Discussion and future directions 55

5 Conclusion 57

5.1 Dissertation summary: graph-based representations of genomic knowledge . 57
5.2 Perspective: data mining, data integration, and hypothesis-generating for human genomics 58
5.2.1 Data mining and data integration: accumulating the knowledge of human genome at a larger scale 58

5.2.2 Integrated analysis and hypothesis generation for the human genome 59

BIBLIOGRAPHY 61

LIST OF FIGURES

FIGURE

2.1 **3'HS1 modulates the hemoglobin gene expression in β-globin gene cluster. a.** Genome-wide Hi-C interaction map and regulatory landscape around β-globin gene cluster in human HUDEP2 cells. ATAC-seq and CTCF track of HUDEP2 cells is shown in the lower panel. Black cycle indicates the position of loops previously identified. Yellow dotted line indicates the three sub-TAD domains identified previously. HPFH1-7 deletion is illustrated and 3HS1 is marked in blue shade. **b.** The scheme of CTCF binding motif orientation engineering in HUDEP-2 cells. **c-e.** In situ Hi-C contact map around β-globin gene cluster in HUDEP-2 cells of wild type (c), 3HS1 deletion (d), and 3HS1 inversion (e). CTCF CUTRUN tracks of WT, 3HS1 deletion and 3HS1 inversion HUDEP-2 cells are shown on the top of corresponding Hi-C plots. All loops called in the HUDEP2 cells of three genotypes are marked with circles of different colors. **f.** The HiCCUPS quantification of loops interaction strength by q value in β-globin locus. Dotted line annotates q = 0.1. n.d.: not detected by HiCCUPS (q value $>$ 0.1). **g.** The composition of -like globin HUDEP-2 cells with 3HS1 deletion. qPCR measurement of -like globin HUDEP-2 in two clones (B6 and D3) of 3HS1 HUDEP-2 cells is shown. Mean $\pm$ SD is displayed, n = 3. **h.** Left panel: relative expression of HBE, HBG (probe measures both HBG1 and HBG2), and HBB in the 3HS1 deleted HUDEP-2 clone B6. Mean $\pm$ SD is displayed, n = 3. Right panel: relative expression of HBE, HBG (probe measures both HBG1 and HBG2), and HBB in the 3HS1 inverted HUDEP-2 clone A2. Mean $\pm$ SD is displayed, n = 3. **i.** The right panel shows the High-performance liquid chromatography (HPLC) for globin composition in Cas9-treated HUDEP-2 control and 3HS1 deletion clone B6. **j.** Flow cytometry plot of HbF in HUDEP-2 cell clones with 3HS1 deletion (B6 and D3), 3HS1 inversion (A2 and G3), and ΔHS5 clone. 11

2.2 **The workflow of scHiCTools.** The workflow of scHiCTools includes five steps: (1) reading input single-cell data in .txt, .hic, or .cool format, generating the summary plots of the cells, and screening cells based on their contact number and contact distance profile, (2) smoothing the scHi-C contact maps using linear convolution, random walk, or network enhancing, (3) calculating the pairwise similarity between cells using fastHiCRep, InnerProduct, or Selfish, (4) embedding or clustering the cells in a low-dimensional Euclidean space using dimension reduction methods, and (5) visualizing the two-dimensional or three-dimensional embedding in a scatter plot. 14

2.3 **Benchmarking experiment results. a.** The embedding of single cells in a cell cycle study [98]. **b.** Evaluating the three embedding methods with a cell-cycle phasing task by average ROCs. **c.** Smoothing methods do not perform well when all positions in Hi-C maps are randomly downsampled. The x-axis is the negative logarithm of sampling rates; y-axis is the average AUCs from ROC curves. **d.** Linear convolution improves the performance of embedding when the dropout rate is high. 18
2.4 Overview of Quagga. **a.** Outline of the workflow. Hi-C file via cooler file or hic file is processed into a horizontally or vertically loaded matrix and a naive peak finding algorithm is used. **b.** Stripe indices and width are calculated based on the vertically or horizontally averaged row or column sums. **c.** Significance testing is applied to called stripe peaks. Called region windows are used to determine the appropriate length of the stripe and whether or not the stripe is enriched over the local background. 22
2.5 **Hi-C stripes called by Quagga are closely related to CTCF/RAD21 extrusion. a-b.** For both 3' and 5' stripes, CTCF and RAD21 are enriched at GM12878 stripe anchors. **c.** The 3' stripe anchors are more enriched in positive-strand CTCF, and the 5' stripes are more enriched in negative-strand CTCF. **d.** Quagga identifies that the majority of Hi-C stripes will disappear after CTCF knockout, in which 95.2 % of the lost stripes are anchored at CTCF binding sites. **e.** Two example regions illustrate that Quagga identified the stripe loss after CTCF knockout. 26

3.1 **Overview of the model. a**, Model architecture. The model inputs are a Hi-C contact map and a number of epigenomic features including histone modifications, chromatin accessibility, and protein binding profiles. The lower-resolution Hi-C contact map is first interpolated into a 200 bp-resolution contact map, and then transformed into a graph $\mathcal{G}$ in which the nodes represent 200 bp genomic bins and the edges represent the interpolated contacts between the nodes. Positional encoding is unrelated to Hi-C or epigenomic data and only encodes node order in the genome. The epigenomic features and positional encoding are assigned to the corresponding nodes as node attributes. The inputs are fed into 1D convolutional and graph convolutional layers to generate hidden representations, which extract features from both nearby genomic regions along the 1D DNA sequence and spatially-contacting regions specified by $\mathcal{G}$. The output layers take input the hidden representations and predict the contact profile at each 200 bp bin as well as the chromatin contacts between bins. **b**, In an example region, the polycomb interactions are accurately predicted by CAESAR. In another example region, loops and stripes undetected by Hi-C are accurately predicted by CAESAR. 31

3.2 **Evaluating CAESAR's performance in multiple tasks.** **a**, The distance-stratified Pearson's correlation with the observed Micro-C contact map from CAESAR and two baselines, HiC-Reg and HiCPlus, in a cross-chromosome experiment. The black dotted lines in **a** and **b** are the correlation between the input Hi-C contact map and the observed Micro-C contact map. **b**, The distance-stratified Pearson's correlation with the observed Micro-C contact map from CAESAR in 1) a cross-chromosome experiment (train on hESC train set and test on hESC test set), 2) a cross-cell type experiment (train on HFF train set and test on hESC test set), and 3) a cross-species experiment (train on mESC train set and test on hESC test set). **c**, The Venn diagram of the loops called from 1) the input Hi-C contact map, 2) the CAESAR-imputed contact map, and 3) the observed Micro-C contact map. **d**, The pile-up visualization of the loops called from 1) the input Hi-C contact map, 2) the CAESAR-imputed contact map, and 3) the observed Micro-C contact map. **e**, The Venn diagram of the stripes called from 1) the input Hi-C contact map, 2) the CAESAR-imputed contact map, and 3) the observed Micro-C contact map. **f**, The pile-up visualization of the stripes called from 1) the input Hi-C contact map, 2) the CAESAR-imputed contact map, and 3) the observed Micro-C contact map. 35

3.3 **The relationships between CAESAR's performance with Hi-C quality, the number of epigenomic features, evolutionary conservation, A/B compartments, and early/late replication timing.** **a**, The epigenomic features in 13-epi, 7-epi, 6-epi, and 3-epi CAESAR models are listed in the table, which are chosen based on common availability. **b**, The distance-stratified Pearson's correlation with the observed Micro-C contact map from CAESAR in a cross-cell type experiment with different numbers of epigenomic features (i.e., 13, 7, 6, and 3). **c**, The distance-stratified Pearson's correlation with the observed Micro-C contact map from CAESAR in a cross-cell type experiment when 1) using the original Hi-C contact map with about 1 billion contacts, 2) randomly down-sampling the Hi-C contact map at different down-sampling rates (resulting in 100 million and 10 million chromatin contacts), and 3) using a surrogate Hi-C contact map with 1 billion contacts aggregated from HFF, GM12878, IMR-90, and K562 with equal proportions. **d**, The model performance in a specific region is quantified by the Spearman's correlation coefficient between the CAESAR-imputed and the Micro-C contact map. In cross-chromosome and cross-cell-type experiments, the model performance (i.e., Spearman's correlation coefficient) is significantly correlated with evolutionary conservation evaluated by sequence alignment scores (n[regions]=1,203, 960, and 240, one-sided t-test). In all the boxplots, the center line indicates median; the box limits are upper and lower quartiles; the whiskers are 1.5×interquartile range; the points are outliers. **e**, In cross-chromosome and cross-cell-type experiments, the correlation coefficient is significantly larger in A compartment than in B compartment (n[regions]=1,018 and 1,388, one-sided t-test). **f**, In cross-chromosome and cross-cell-type experiments, the correlation coefficient is significantly larger in early-replicating regions than in late-replicating regions (n[regions]=1,203, 960, and 240, one-sided t-test). 37

3.4 **The interactions between genes and their CRISPRi-validated enhancers in CAESAR-imputed contact maps.** **a**, The CAESAR-imputed contact map of K562 at *MYC* region (chr8: 127,600,000-127,850,000) demonstrates significant contacts between *MYC* and *PVT1*, which agree with with CRISPRi score peaks, but are not shown on the original input Hi-C contact map. The magnitude of the epigenomic features is the observed value divided by the genome-wide average. **b**, The CAESAR-imputed contact map of K562 at *GATA1* region (chrX: 48,725,000-48,825,000) demonstrates significant contacts between *GATA1* and *HDAC6*, which agree with with CRISPRi score peaks, but are not shown on the original input Hi-C contact map. 38

3.5 **The enrichment of eQTL-gene interactions in CAESAR-imputed contact maps.** **a**, The loop between gene *USB1*'s TSS and its pancreas-specific eQTL, which cannot be observed on the original Hi-C contact map, appears on the CAESAR-imputed contact map for pancreas. Although gene *TEPP*'s eQTL is lung-specific, the corresponding loop can be called from the CAESAR-imputed contact maps for both lung and pancreas. **b**, Pile-up analysis of the chromatin contacts between eQTLs and their corresponding gene TSS from twelve different human tissues and cell lines on CAESAR-imputed contact maps and Micro-C contact maps. The average contact values in the central 5×5 squares are marked on the plots, in which the bold fonts indicate that eQTLs and CAESAR-imputed contact maps are from the same tissue/cell line. 40

3.6 Attributing CAESAR outputs to epigenomic features via *integrated gradient*. Larger attribution magnitudes indicate more contribution to the model's prediction. **a**, The significant attribution of the particular stripe are from its anchor. Although all 6 epigenomic features have peaks at the anchor locus, the model predicts the stripe mostly from 1) ATAC-seq and CTCF peaks at the anchor, and 2) H3K4me1 modification surrounding the anchor. **b**, The significant attribution of the particular loop are from its two anchors. Although H3K27ac have peaks at the left anchor locus, its contribution is negative towards predicting the loop. The CTCF binding at the anchors and H3K4me1/H3K4me3 modifications next to the anchors have positive attribution in predicting the loop. 43

4.1 **An example subgraph of GenomicKB.** In this subgraph, three GWAS variants of type II diabetes are connected with entities including genes, tissues, and 3D chromatin structures. 47

4.2 **GenomicKB simplifies cross-modality analysis as queries over the knowledge graph.** If a user is interested in relations between T2D and genes, then instead of searching multiple databases including GWAS, ENCODE, and GO, a sub-graph query over GenomicKB returns all variants, genes, and gene ontologies that satisfy the query criteria. 53

4.3 **GenomicKB supports queries related to positional relations between genomic entities.** An example query of CTCF binding to loop anchors is illustrated. 54

4.4 **GenomicKB reconciles multiple data sources for the same problem, such as identifying enhancers and mapping enhancers to genes.** Query 1 demonstrates how GenomicKB evaluates the consensus enhancers between CCRE and EnhancerAtlas. Query 2 illustrates how enhancer-gene mapping from EnhancerAtlas is validated by eQTL-gene pairs in GenomicKB. 55

LIST OF TABLES

TABLE

3.1 CAESAR-imputed tissues and cell lines 42

4.1 Number of entities included in the genomic graph and their data sources 49
4.2 Number of relationships included in the genomic graph and their data sources . 50

LIST OF ACRONYMS

3C Chromosome conformation capture

4DN The 4D Nucleome Project

ABC model Activity-by-contact model

AUC Area under the curve

CBS CTCF binding site

CAESAR Chromosomal structure And EpigenomicS AnalyzeR

CCRE Candidate cis-regulatory elements

cHi-C Capture Hi-C

CNV Copy number variance

CRISPRi CRISPR interference

DGV Database of genomic variant

ESC Embryonic stem cell

ENCODE Encyclopedia of DNA elements

E-P Enhancer-Promoter

EPD Eukaryotic promoter database

eQTLs Expression quantitative trait loci

FP False positive

GC Graph convolution

GO Gene ontology

GTEx The Genotype-Tissue Expression Project

GWAS Genome-wide association study

HBB Adult β-globin

HBE Embryonic ϵ-globin

HBG Fetal γ-globin

HFF Human foreskin fibroblasts

HPFH Hereditary persistence of fetal hemoglobin

KG Knowledge graph

KO Knockout

MDS Multidimensional scaling

MNase Micrococcal nuclease

PCR Polymerase chain reaction

PHATE otential of heat-diffusion for affinity-based trajectory embedding

SCC Stratum-adjusted correlation coefficient

scHi-C Single-cell Hi-C

T2D Type II diabetes

TAD Topological associating domain

TF Transcriptional factor

TP True positive

t-SNE t-Distributed stochastic neighbor embedding

TSS Transcription start site

WT Wild type

ABSTRACT

Transcriptional regulation in human cells is a complex process that requires the collaboration of diverse genomic elements and chemicals. To understand the mechanisms, projects including the Encyclopedia of DNA Elements (ENCODE), Roadmap Epigenomics, and 4D Nucleome (4DN) have generated thousands of genomic and epigenomic datasets. These datasets annotated functional elements for the human genome (e.g., enhancers and promoters), summarized experimental results for epigenomic features (e.g., protein binding locations), and linked different modalities with statistical models (e.g., GWAS and eQTLs). From the available data, it has become apparent that the human genome should not be over-simplified as a 1-D linear sequence. Long-range dependencies on DNA sequences play a vital role in human transcriptional regulation. For example, enhancers, the primary units of gene expression regulation, often reside hundreds of kilobases away from their target genes. Enhancers engage in physical interactions with target genes across vast genomic distances to activate them. Therefore, interpreting the human genome requires a more advanced data structure capable of capturing long-distance and complicated relationships.

This dissertation discusses how to decipher the human genome as a graph. Graphs, composed of nodes (or vertices) and edges, provide a powerful framework for modeling relationships. Graphs have been proven effective in representing relationships in diverse real-world scenarios, such as social networks, transportation systems, and communication networks. In the subsequent chapters, the representations of the human genome as a graph will be introduced and explored.

Chapter 2 introduces the application of chromosome conformation capture (3C) technology, which unveils physical interactions among genomic regions. Analyzing the large-scale contact maps generated by 3C technology is instrumental in uncovering the long-range dependencies of genomic entities and understanding transcriptional regulation. Therefore, we developed computational tools including scHiCTools and Quagga to extract structural features from these maps.

In Chapter 3, we addressed the importance of high-resolution and high-quality chromatin contact maps. Therefore, we developed a computational model, CAESAR, to connect epigenomics and high-resolution chromatin structure. CAESAR successfully imputes an un-

precedented number of high-resolution human chromatin contact maps, which allows users to easily navigate these fine-scale chromatin structures and the corresponding regulatory mechanisms.

Beyond 3D interactions, numerous data consortia and databases unveil the characteristics of genomic entities and their relationships. Despite the invaluable insights provided by these consortia, the separately stored tabular data remain in a 1D sequential framework, posing inconveniences for genomic research and scientific discoveries. To address this challenge, we introduce the Genomic Knowledgebase (GenomicKB) in Chapter 4. GenomicKB is a knowledge graph that seamlessly integrates datasets and annotations related to the human genome into a knowledge graph. Through a graph-based interpretation of the human genome, we anticipate that genomic research will increasingly become data-driven. GenomicKB aims to provide high-quality and integrated data for large-scale machine learning methods, thereby facilitating scientific discoveries.

CHAPTER 1

Introduction

1.1 Central dogma and transcriptional regulation in the human genome

The central dogma, which is known as "DNA makes RNA, and RNA makes protein", despite some counter-examples like reverse transcription and RNA replication, is a well-accepted model for explaining the genetic information flow in biological systems. However, diverse cell and tissue types are generated from the same genome by varying the expression levels of different genes, and the differential expression of genes is controlled by non-coding sequences that regulate gene expression [23]. To understand the process, projects like Encyclopedia of DNA Elements (ENCODE)[17], Roadmap Epigenomics [13], and 4D Nucleome (4DN) [22] have generated thousands of epigenomic datasets in order to elucidate regulatory functions of different genomic regions.

The precise control of transcriptional regulation requires the collaboration of enhancers, promoters, transcriptional factors (TFs), and chromatin structures. The primary units of gene expression regulation are enhancers, which can be characterized by different approaches including CRISPRi functional validation, evolutionary conservation analysis, and epigenomic profiling [75, 105, 141, 86]. Enhancers physically interact with the target genes across vast genomic distances to activate them [32, 95, 106, 94, 136, 126].

Some preliminary computational models have been proposed for elucidating how enhancer properties and E-P interaction strength relate to gene expression. For example, the activity-by-contact (ABC) model [45] quantifies an element's effect on a gene as its enhancer activity multiplied by its 3D contact strength with the promoter under the assumption of a linear relationship between E-P interaction strength and gene expression. Nevertheless, other data suggest that expression is a non-linear function of E-P interaction strength [132, 140, 152, 113]. In these models, enhancer features are profiled by ENCODE and RoadMap Epigenomics (e.g., ATAC-seq and H3K27ac), and E-P interactions are profiled by Hi-C, Micro-C, and

intact Hi-C contact maps from 4DN and ENCODE projects.

1.2 Human genome is not only a 1-D sequence

In eukaryotic species, only 2% of genome is for coding proteins, and the remainder is riddled with *cis*-regulatory DNA elements such as promoters, enhancers, repressors, and insulators. During transcriptional regulation, in which massive regulatory elements target their corresponding genes, the existence of 3D "spatial regulomic elements" has been demonstrated in various studies. Specifically, enhancers, the primary units of gene expression regulation, often reside hundreds of kilobases away from their target genes. Enhancers engage in physical interactions with target genes across vast genomic distances to activate them. For example, long-range spatial enhancer-promoter contacts can control the expression of *Shh* gene [25, 138], polycomb-bound promoters around the *Hox* clusters can mediate gene repression in mouse embryonic stem cell (ESC) [121], and super-enhancers formed by spatial clustering of enhancers can work as a "regulatory factory" [63]. Other instances include long-distance co-activation of genes and co-accessibility of genomic regions.

From the available data, it has become apparent that the human genome should not be over-simplified as a 1-D linear sequence. Instead, long-range dependencies on DNA sequences play a vital role in human transcriptional regulation, and interpreting the human genome requires a more advanced data structure capable of capturing long-distance and complicated relationships.

Chromatin 3D structures should be also taken into consideration in understanding the regulatory process. For example, topological associating domains (TADs) and their boundaries control gene expression by assisting intra-domain enhancer–promoter links while inhibiting inter-domain contacts between regulatory elements to avoid gene mis-activation [130]. Research shows that changing the distance between mouse *Shh* gene and its enhancer zone within the TAD has little effect on gene expression, while disrupting the TAD results in the loss of *Shh* expression [129]. However, TAD boundaries are not strictly impassible, which is supported by results from promoter cHi-C [65] and expression quantitative trait loci (eQTLs). Polycomb regions are one example of which regulatory elements affect multiple TADs [121]. In addition, cohesin-mediated loops are usually considered as structural units of gene expression control [30, 66], but they might only play a minor role in transcriptional regulation since most active genes do not tend to locate near loop anchors [112] and anchor depletion only results in few significant changes of gene expression [111]. For TADs and loops, one assumption is that the absence of them will re-direct some enhancers or promoters to alternative targets [122]. However, three-dimensional structures at lower hierarchies

such as micro-TADs, short-range enhancer-promoter (E-P) and promoter-promoter (P-P) links are not well-studied yet. Moreover, although many studies have demonstrated that 3D chromatin interactions are established concomitantly with transcriptional regulation, people still lack the evidence to show whether 3D structures are the cause or the consequence of gene regulation.

1.3 The complex network of the human genome from diverse data sources

Since the completion of the Human Genome Project [64], ever-evolving biotechnologies have enabled us to characterize the human genome from different perspectives. Consequently, several landmarking consortia have made tremendous progress towards understanding the functions of human genome, such as the Encyclopedia of DNA Elements (ENCODE) [17], Roadmap [13], Genotype-Tissue Expression (GTEx) [51], 4D Nucleome (4DN)[24], and the Human BioMolecular Atlas Program (HubMAP) [19], among others. Each of these consortia has generated thousands of datasets, and provided different insights regarding human genome at an unprecedented scale and depth. Importantly, these consortia have provided credible evidence about the *connections* among genomic, epigenomic, and transcriptomic entities. For example,

- the functions of human genome are linked to different 3D chromatin organizational structures, such as chromatin loops, stripes, topologically associating domains (TADs), subTADs, microTADs, and compartments,
- a large number of genetic variants have been identified to be associated with gene expression (i.e., eQTLs), transcription factor binding, chromatin accessibility or histone modifications (i.e., chromQTLs), and 3D chromatin organization (i.e., 3dQTLs [49]) in a tissue-specific or cell type-specific manner, and
- spatially resolved gene expression data (e.g., Slide-seq [115]) and multiplex imaging data (e.g., CODEX [47]) demonstrate that the spatial positions of the cells in tissues strongly influence their functions.
- tissue/cell type-specifically expressed genes and super-enhancers are usually enriched in tissue/cell type-specific frequently interacting regions (FIREs),
- non-coding variants influence the binding affinity of transcriptional factors (TFs), which accounts for diverse human traits and diseases, and

- genes that located in the same TAD are more likely to co-express and activated by the same enhancer, but CTCF binding between them weakens their correlation,

However, these datasets and annotations are isolated in the sense that they are stored as tabular-structured data matrices at individual data portals. As a result, it is difficult to jointly analyze these datasets for scientific discoveries, such as understanding GWAS variants and eQTLs in the context of 3D chromatin organizations with a genomic region, understanding transcription regulation along a signaling pathway, and exploring cell-to-cell communications from spatially resolved gene expression data.

Therefore, we propose to decipher the human genome as a graph. Graphs, composed of nodes (i.e., vertices or entities) and edges (i.e., relationships), provide a powerful framework for modeling relationships. Nodes represent entities, while edges denote connections or relationships between these entities. Graphs have proven effective in representing relationships in diverse real-world scenarios, such as social networks, transportation systems, and communication networks. In the subsequent chapters, different representations of the human genome as a graph will be introduced and explored.

CHAPTER 2

Revealing the 3D Structures of the Human Genome with Chromatin Conformation Capture (3C) Technology

As introduced in the previous chapter, long-range dependencies of genomic entities such as E-P links are closely related to chromatin 3D structures. In the following chapter, we will introduce the experimental technology to capture chromatin 3D structures at a genome-wide scale. With the large-scale contact maps generated by 3C technology, analytical pipelines are instrumental in uncovering the long-range dependencies of genomic entities. Therefore, we developed computational tools including scHiCTools and Quagga to extract structural features from these maps. These methods aim to capture structural features in the 3D genome, revealing the genome-wide connection with transcriptional regulation.

Section 2.1 introduces the concept of 3C technologies and chromatin contact maps (e.g., Hi-C and Micro-C), and summarizes the hierarchical chromatin structures uncovered by the technologies. Section 2.2 connects the chromatin structures with transcriptional regulation in the nucleus. To exemplify the conclusion, section 2.3 introduces a typical example of CTCF-mediated 3D structures affecting gene expression, which is also closely related to genetic therapy.

Therefore, researchers are investigating 3D chromatin structures to comprehend the mechanisms of transcriptional regulation. During this process, analytical pipelines are vital in extracting 3D structural features from raw contact maps. I developed two computational tools for Hi-C/Micro-C contact map analysis, namely scHiCTools (section 2.4) and Quagga (section 2.5).

2.1 Introduction of Chromatin Conformation Capture (3C) technology and chromatin contact maps

2.1.1 General concept of 3C technology

Chromosome Conformation Capture (3C) is a pioneering molecular biology technique that utilizes formaldehyde cross-linking to fix chromatin interactions within the nucleus [21]. The cross-linked chromatin is then digested with a restriction enzyme, and the resulting fragments are ligated under dilute conditions, favoring intramolecular ligation events. The cross-links are subsequently reversed, and the DNA is purified for analysis. Polymerase chain reaction (PCR) or high-throughput sequencing can be employed to detect specific ligation products.

Results from 3C experiments reveal the proximity and interaction frequency of genomic loci. By examining the ligated DNA fragments, researchers can infer the spatial relationships between distant DNA sequences. This technique has been pivotal in uncovering the existence of chromatin loops, which bring enhancers and promoters into physical proximity, influencing gene expression patterns.

2.1.2 Hi-C technology

Hi-C builds upon the principles of 3C but extends the scope to a genome-wide scale [81]. After cross-linking, the chromatin is digested with a restriction enzyme, similar to 3C. However, in Hi-C, the digested fragments are end-ligated, irrespective of their linear proximity. This process captures all possible chromosomal interactions within the nucleus. Subsequent high-throughput sequencing generates a comprehensive map of the entire interactome.

Hi-C results provide detailed information on the spatial organization of the genome, revealing topologically associated domains (TADs), chromatin loops, and long-range interactions [112]. The details will be introduced in the following section. These findings have significantly advanced our understanding of how the 3D genome structure influences gene regulation and cell identity.

2.1.3 Micro-C technology

Recently, new approaches including DNase Hi-C[88] and Micro-C[62, 72] have begun to provide increasingly higher-resolution 3D chromatin organization. Micro-C represents a refinement of Hi-C, introducing the use of micrococcal nuclease (MNase) to digest chromatin [62, 72]. This enzyme preferentially cleaves linker DNA, yielding smaller fragments and enhancing the resolution of chromatin interaction maps. After digestion, the fragments are

end-ligated and subjected to high-throughput sequencing.

Micro-C results offer higher-resolution maps of chromosomal interactions, allowing for the identification of sub-TAD structures and finer details in the 3D genome organization. Researchers can discern subtle variations in chromatin architecture, gaining insights into how specific genomic elements, such as regulatory regions and insulators, contribute to the overall spatial organization of the genome. This increased resolution is particularly valuable for understanding the intricacies of gene regulation and the impact of chromatin structure on cellular processes.

2.1.4 Hierarchical structural features revealed by chromatin contact maps

Chromosome conformation capture (3C) techniques [21] are a set of molecular biology methods to analyze chromatin spatial organization. By proximity ligation, these methods quantitatively measure contact frequencies between genomic loci in 3D space to obtain chromatin contact maps, in which high-throughput sequencing technology can further increase the sequencing depth (referred as Hi-C [81]).

Hierarchical chromatin structures are revealed by 3C technologies. A/B compartments [81] and sub-compartments [112] at megabase level are discovered with Hi-C. According to contact profiles, the entire genome could be split into A and B compartments, in which genomic loci tend to interact preferentially with loci in the same compartment. Further analysis found that A compartment, which usually displaces the interior of the nucleus, is richer in genes, G-C base pairs and histone marks for active transcription; B compartment is opposite. Sub-compartments (A1/A2/B1/B2/B3/B4) [112] are only found in deep-sequenced Hi-C data with billions of contacts, which also demonstrate the preference of intra-sub-compartment interactions. Lower-level structure topological associating domains (TADs)[104], discovered from Hi-C contact maps, are self-interacting genomic regions with sub-megabase lengths, whose separation is controlled by binding proteins including CCCTC binding factor (CTCF) and cohesin [111, 139]. Compartment switches happen frequently during differentiation [28], while TADs are relatively conservative among different cell types and even species [29]. Another type of structures at hundreds of kilobase (kb) level, chromatin loops (a.k.a. insulated neighbourhoods) [112], are usually formed by interactions between two CTCF-bound sites and mediated by cohensin[43], which frequently link promoters with enhancers, providing spatial restrictions for gene regulations [30, 52]. Finer-scale structures at sub-TAD level (hundreds of bp to tens of kb) are newly discovered with Micro-C [62, 72]. For example, micro-TADs are self-interacting domains like TADs but only spread

tens of kilobases, whose boundary formation is more complex than CTCF/cohensin separation. The structure of polycomb repressive regions is also observed to be nested sets of inter-spaced contacts [62] instead of loops in previous low-resolution contact maps [33].

2.2 The relationships between transcriptional regulation and 3D chromatin organization

In eukaryotic species, only 2% of the human genome is for coding proteins, and the remainder is riddled with *cis*-regulatory DNA elements such as promoters, enhancers, repressors and insulators. During transcriptional regulation, in which massive regulatory elements target their corresponding genes, the existence of 3D "spatial regulomic elements" has been demonstrated in various studies. For example, long-range spatial enhancer-promoter contacts can activate gene transcription[25, 138]; polycomb-bound promoters around the *Hox* clusters can mediate gene repression [121]; "super-enhancers" formed by spatial clustering of enhancers can work as regulatory entities[63]. These 3D spatial regulatory elements are also related to 3D chromatin structures introduced in last subsection. For example, TADs and their boundaries control gene expression by assisting intra-domain enhancer–promoter links while inhibiting inter-domain contacts between regulatory elements to avoid gene mis-activation [130]. Research shows that changing the distance between mouse *Shh* gene and its enhancer zone within the TAD has little effect on gene expression, while disrupting the TAD results in the loss of *Shh* expression [129]. However, TAD boundaries are not strictly impassible, which is supported by results from promoter CHi-C [65] and expression quantitative trait loci (eQTLs). Polycomb regions are one example which regulatory elements effect across multiple TADs [121]. Cohesin-mediated loops are usually considered as structural units of gene expression control [30, 66], but they might only play a minor role in transcriptional regulation since most active genes do not tend to locate near loop anchors [112] and anchor depletion only results in few significant changes of gene expression [111]. For TADs and loops, one assumption is that the absence of them will re-direct some enhancers or promoters to alternative targets [122]. However, 3D structures at lower hierarchies such as micro-TADs, short-range enhancer-promoter (E-P) and promoter-promoter (P-P) links are not well-studied yet. Moreover, although many studies have demonstrated that 3D chromatin interactions are established concomitantly with transcriptional regulation, people still lack the evidence to show whether 3D structures are the cause or the consequence of gene regulation.

*This work was published on Elife, of which I am a co-first author. In this work, I did the analysis of Hi-C contact maps, including Hi-C resolution enhancing and loop calling.

2.3 CTCF-mediated chromatin loops regulate fetal hemoglobin expression[55]

In this section, we use an example to show that genetic editing and 3D genome changes can have therapeutic implications in treating diseases [55]. In this example, the deletion of a CTCF site alone induces fetal hemoglobin expression in both adult CD34+ hematopoietic stem and progenitor cells and HUDEP-2 erythroid progenitor cells. This induction is driven by the ectopic access of a previously postulated distal enhancer located in the OR52A1 gene downstream of the locus, which can also be insulated by the inversion of the 3'HS1 CTCF site.

2.3.1 Introduction: human β-globin and hemoglobinopathies

The human β-globin locus consists of five globin genes embedded in the olfactory receptor cluster. During early development, these globin genes undergo gene switching from embryonic ϵ-globin (HBE) to fetal γ-globin (HBG1/2) and finally to adult β-globin (HBB). Inherited mutations in the HBB gene lead to dysfunction of the adult β-globin protein, causing hemoglobinopathies [9]. The symptoms of these disorders, including sickle cell disease and β-thalassemia, can be alleviated by persistent expression of fetal hemoglobin (hereditary persistence of fetal hemoglobin [HPFH]) throughout adulthood, which compensates for the mutant adult β-globin [7, 54]. As such, multiple genome-editing strategies have been proposed to mimic HPFH as a treatment for hemoglobinopathies [10, 14, 116, 118, 117, 131]. Two types of HPFH have been identified based on patient genetics. First is the non-deletional HPFH caused by point mutations in the BCL11A binding site at the HBG1/2 promoters, and disruption of this transcriptional repressor binding leads to the activation of these genes [83, 93, 131]. Second is the deletional HPFH that consists of the excision of a large genomic region within the β-globin locus, frequently including HBB and HBD [143]. These deletions can vary in length, and it remains unclear as to how they lead to the expression of fetal globin in adulthood [143].

The human β-globin gene locus is flanked by five CTCF binding sites (CBSs), which form the anchors for six chromosomal loops. Two convergent CBSs, designated as 3'HS1 and HS5, are located at the borders of the globin gene cluster. These two CBSs are nested between a downstream CBS (referred to as 3'-OR52A5-CBS) and two closely spaced upstream CBSs (referred together as 5'-OR51B5-CBSs). The HPFH deletions frequently cover the 3'HS1 CBS (Figure 2.1a). Therefore, we hypothesized that 3'HS1 may play a role in regulating β-globin cluster gene expression. To explore this, we first deleted the 3'HS1 us-

ing CRISPR/Cas9 genome-editing technology in K562 myelogenous leukemia cells, which express high levels of hemoglobin (Figure 2.1). At the same time, we also deleted HS5 as a control in K562 cells. We observed that deletion of the HS5 CTCF site resulted in the upregulation of the 3' genes including HBB and HBG1/2. Interestingly, the disruption of 3'HS1 CBS led solely to the upregulation of HBG1/2 (Figure 2.1). These results show that altering the CTCF binding profile across the locus can significantly change the expression of the β-globin genes.

2.3.2 3'HS1 CTCF binding site in human β-globin locus regulates fetal hemoglobin expression

We performed Hi-C and capture Hi-C to examine the changes to 3D chromatin organization at the β-globin locus following alterations to the CBSs (Figure 1c–e). In situ Hi-C data was generated with high resolution at 5 kb. A total of 15,207–16,529 loops could be detected in the HUDEP-2 clones used for in situ Hi-C using Mustache. The CTCF bound around the β-globin locus form four chromosomal loops and separate the cluster into three distinct domains (Figure 2.1a and c). Of notice, we could detect the enhancer to target gene interaction between the LCR and the HBB gene. We also tested the copy number variance (CNV) in the three particular HUDEP-2 clones, we could verify all clones have chromosome number 49–50, XY, which is of the normal range in unmodified HUDEP-2 cells. Next, we tested if the chromosomal loops were altered by the 3'HS1 editing. We applied the HiCCUPS method to call the significant chromosomal loops in the β-globin locus, and four loops were identified with q value less than 0.1 (Figure 2.1c and f). We then use the q value of the called loops by HiCCUPS to quantify the strength of loop interactions between CBSs Of the convergent CTCF interactions, 3'HS1 to 5'-OR51B5-CBSs was not called as loop with q value over 0.25. One loop was called between the two forward CTCF CBSs – 3'HS1 and 3'-OR52A5 CBS (Figure 2.1d and f). In the 3'HS1 deletion clone, the loss of CTCF at 3'HS1 resulted in the total loss of loops between 3'HS1 and HS5 as well as loops between 3'HS1 and 5'-OR51B5-CBSs (not called as loop). Concomitantly, a strong increase in the interaction between HS5 and 3'-OR51A5-CBS was observed (Figure 2.1d and f). This reveals how the loss of a CTCF anchor drastically alters the 3D chromatin organization in the β-globin locus. The inversion of the 3'HS1 CTCF caused a significant increase in the interaction between 3'HS1 and 3'-OR52A5 CBS. Meanwhile, 3'HS1 upstream interactions with HS5 and 5'-OR51B5-CBSs were decreased (Figure 2.1e and f). This revealed that the inversion of 3'HS1 CTCF drove the formation of chromosomal loops between the convergent CBSs, which may lead to stronger insulation of regulatory elements.

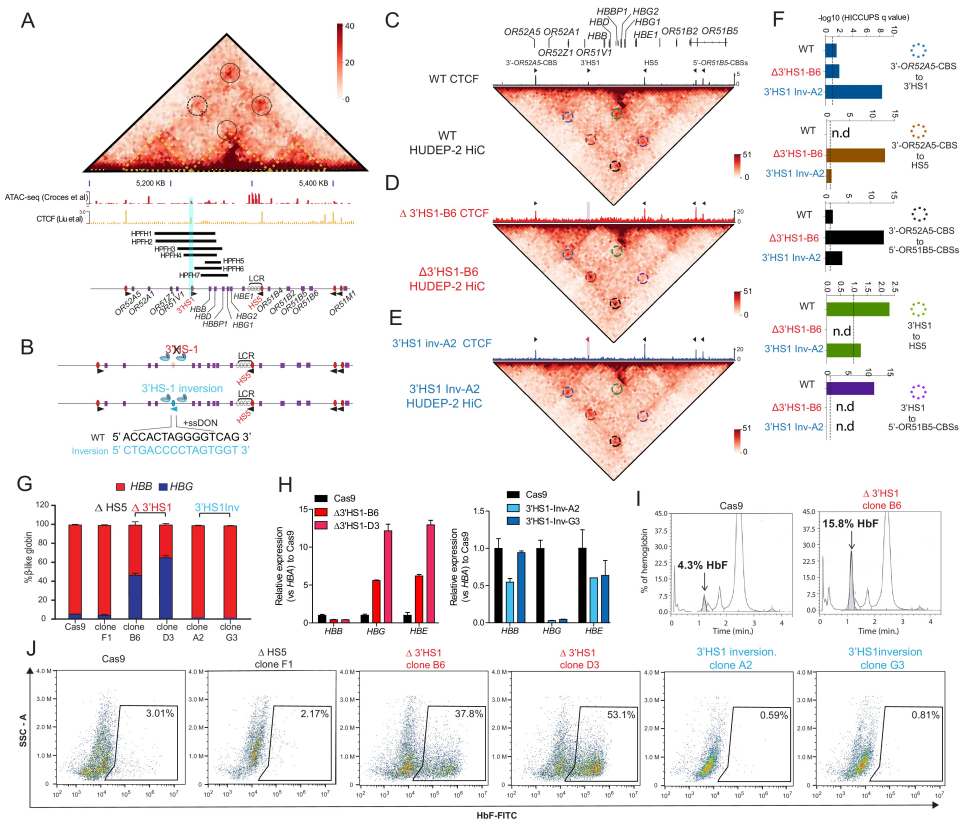

Figure 2.1: **3'HS1 modulates the hemoglobin gene expression in β-globin gene cluster.**

a. Genome-wide Hi-C interaction map and regulatory landscape around β-globin gene cluster in human HUDEP2 cells. ATAC-seq and CTCF track of HUDEP2 cells is shown in the lower panel. Black cycle indicates the position of loops previously identified. Yellow dotted line indicates the three sub-TAD domains identified previously. HPFH1-7 deletion is illustrated and 3HS1 is marked in blue shade. **b.** The scheme of CTCF binding motif orientation engineering in HUDEP-2 cells. **c-e.** In situ Hi-C contact map around β-globin gene cluster in HUDEP-2 cells of wild type (c), 3HS1 deletion (d), and 3HS1 inversion (e). CTCF CUTRUN tracks of WT, 3HS1 deletion and 3HS1 inversion HUDEP-2 cells are shown on the top of corresponding Hi-C plots. All loops called in the HUDEP2 cells of three genotypes are marked with circles of different colors. **f.** The HiCCUPS quantification of loops interaction strength by q value in β-globin locus. Dotted line annotates q = 0.1. n.d.: not detected by HiCCUPS (q value > 0.1). **g.** The composition of -like globin HUDEP-2 cells with 3HS1 deletion. qPCR measurement of -like globin HUDEP-2 in two clones (B6 and D3) of 3HS1 HUDEP-2 cells is shown. Mean ± SD is displayed, n = 3. **h.** Left panel: relative expression of HBE, HBG (probe measures both HBG1 and HBG2), and HBB in the 3HS1 deleted HUDEP-2 clone B6. Mean ± SD is displayed, n = 3. Right panel: relative expression of HBE, HBG (probe measures both HBG1 and HBG2), and HBB in the 3HS1 inverted HUDEP-2 clone A2. Mean ± SD is displayed, n = 3. **i.** The right panel shows the High-performance liquid chromatography (HPLC) for globin composition in Cas9-treated HUDEP-2 control and 3HS1 deletion clone B6. **j.** Flow cytometry plot of HbF in HUDEP-2 cell clones with 3HS1 deletion (B6 and D3), 3HS1 inversion (A2 and G3), and ΔHS5 clone.

Next, we evaluated the expression of the β-globin genes and found that the HBG1/2 and HBE genes upregulated 2.5- to 8-fold in the Δ3'HS1 clones (Figure 1G and H). In contrast, the inversion of 3'HS1 resulted in a >50% reduction of HBE and near-complete depletion of HBG1/2 (Figure 2.1h). With additional experiments, we show that the long-range interaction of a distal enhancer in the OR52A1 gene drives the expression of HBG1/2 (which is out of the scope of this dissertation and not introduced in detail).

2.3.3 Summary and discussion

Our study reveals how CTCF binding at this locus modulates the accessibility of the fetal HBG1/2 genes to a downstream enhancer. In the HPFH enhancer scenario, 3'HS1 limits the HPFH enhancer access to HBG1/2 by forming the sub-TAD with 5'HS. When the 3'HS1 CBS is deleted, the HPFH enhancer gains access to HBG1/2 without the hinder of 3'HS1–HS5 loop. When the 3'HS1 CBS motif is inverted, the HPFH enhancer is further restricted by the pairing of 3'-OR52A5-CBS to the inverted 3'HS1 CBS, which results in the strong chromosomal loop formation between the two CBSs. This insulation leads to the reduced HBG1/2 expression and upregulation of OR52A5.

2.4 A Toolbox for analyzing single-cell Hi-C data: scHiCTools[79]

Recent single-cell Hi-C sequencing (scHi-C) technologies profile three-dimensional (3D) chromatin contact maps in individual cells, allowing us to characterize chromatin organization dynamics and cell-to-cell heterogeneity [99, 38, 109]. However, the interpretation of scHi-C data exposes several inherent data analysis challenges. First, unlike RNA-seq data and ATAC-seq data which are vectors of m-dimensional measures, Hi-C data are essentially symmetric matrices of $m \times m$-dimensional pairwise measures, where the number of genomic loci m is usually more than tens of thousands, depending on the resolution of the contact maps. Second, scHi-C analysis suffers from high dimensionality, the sparsity of the contact maps, and sequencing noise. Typically in a scHi-C experiment, up to a few thousand single cells are profiled, whereas the number of contacts in each cell ranges from a few thousand to hundreds of thousands. Third, single cells in one experiment usually reside in a low-dimensional manifold, such as a circular cell cycle structure or a bifurcation differentiation structure. Thus, proper embedding of scHi-C data in a low-dimensional Euclidean space is vital in scHi-C data analysis.

*This work was published on PLOS Computational Biology, of which I am the second author. I did the mathematical derivation and implemented the first version of the Python code of scHiCTools.

In this work, we implemented a versatile scHiCTools which includes many common approaches in the entire workflow of analyzing single-cell Hi-C data [79]. In particular, we implemented three similarity measures, including a faster version of HiCRep, a new "InnerProduct" approach, and another efficient Hi-C similarity measure named Selfish [3]. Among the three methods implemented, InnerProduct provides the most efficient and satisfactory similarity measure. Benchmarking experiments demonstrate that the new InnerProduct approach runs thousands of times faster than the original HiCRep, and produces comparably accurate projection. To deal with the sparsity in scHi-C data, different smoothing approaches are implemented, including linear convolution, random walk, and network enhancing [135]. Among the three approaches, linear convolution appears to be most effective for smoothing contact maps in our experiments. In addition to the computational components, our toolbox supports different input file formats, diagnostic summary plots, and flexible projection plots Our open-source toolbox, scHiCTools, as the first toolbox of such kind, can be useful for analyzing scHi-C data.

2.4.1 Overview of the method

Our scHiCTools implements commonly used approaches to analyze single-cell Hi-C data. The key component of the toolbox is a number of dimension reduction approaches which takes a number of single cells' contact maps as input, and embeds the cells in a low-dimensional Euclidean space. The toolbox also provides a number of built-in auxiliary functions for flexible and interactive visualization. The entire workflow of scHiCTools, illustrated in Fig 1, includes five steps: (1) reading single-cell data in .txt, .hic, or .cool format, generating diagnostic summary plots, and screening cells by their contact number and contact distance profile, (2) smoothing scHi-C contact maps using linear convolution, random walk, or network enhancing, (3) calculating pairwise similarity between cells using fastHiCRep, InnerProduct, or Selfish, (4) embedding or clustering the cells in a low-dimensional space using dimension reduction methods, and (5) visualizing the two-dimensional or three-dimensional embedding in a scatter plot (Figure 2.2). Except for the two pairwise similarity calculation methods, fastHiCRep and InnerProduct, other methods are implemented as originally stated.

Three embedding approaches are implemented in scHiCTools. The first approach is a faster implementation of original HiCRep [142]. Original HiCRep calculates m stratum-adjusted correlation coefficients (SCCs) of the m strata near the diagonal of two contact maps, and then uses weighted sum to aggregate them into one score. It is equivalent to finding a feature vector for each contact map and then computing the inner product among the feature vectors (Supplementary Note 1). This simplification reduces HiCRep's computa-

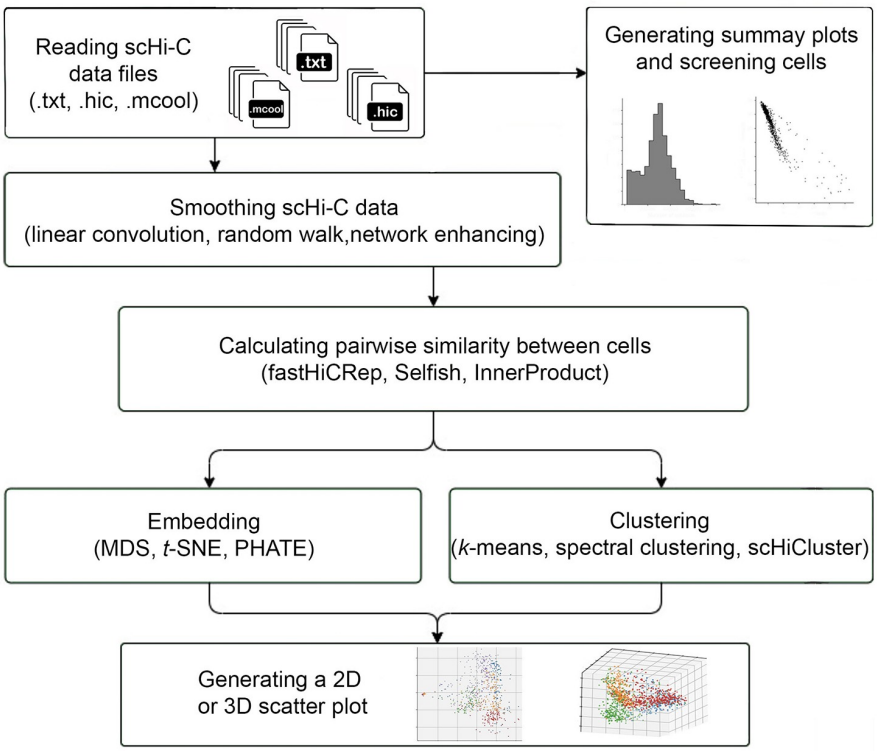

Figure 2.2: **The workflow of scHiCTools.**
The workflow of scHiCTools includes five steps: (1) reading input single-cell data in .txt, .hic, or .cool format, generating the summary plots of the cells, and screening cells based on their contact number and contact distance profile, (2) smoothing the scHi-C contact maps using linear convolution, random walk, or network enhancing, (3) calculating the pairwise similarity between cells using fastHiCRep, InnerProduct, or Selfish, (4) embedding or clustering the cells in a low-dimensional Euclidean space using dimension reduction methods, and (5) visualizing the two-dimensional or three-dimensional embedding in a scatter plot.

tion complexity from $O(n^2)$ to $O(n)$, and we name it **fastHiCRep**, which is implemented in our toolbox. Alternatively, we can further simplify fastHiCRep by directly setting the concatenated z-normalized strata as feature vectors (Supplementary Note 1). With the feature vectors, an inner product is then calculated to obtain the similarity matrix of a group of cells. We name this second approach **InnerProduct**. In the end, a dimension reduction method, Multidimensional Scaling (MDS), is used to get a lower-dimensional embedding of each cell. The third embedding approach **Selfish** [3] was recently proposed for bulk Hi-C comparative analysis. It first uses a sliding window to obtain a number of square regions along the diagonal of the contact map, and then counts overall contact numbers in each region. Then, it generates a one-hot "fingerprint matrix" for each contact map based on pairwise comparison of these reads. Gaussian kernels over the fingerprint matrices are calculated as similarities among the cells.

Our toolbox scHiCTools includes three smoothing approaches. **Linear convolution** is based on a 2D filters (a.k.a., convolution kernels) with equal values in every position, which can be viewed as smoothing over nearby bins in Hi-C contact maps. For example, original HiCRep uses a parameter h to describe a $(2h+1)\times(2h+1)$ kernel, i.e. $h = 1$ indicating a 3×3 kernel with each element equals $\frac{1}{9}$. Because this approach is similar to reducing resolution, it is believed to be effective when contact maps are sparse. **Random walk** is a stochastic process updating the elements of the input matrix W by $W' = W \cdot B$, in which $B_{ij} = \frac{W_{ij}}{\sum_i W_{ij}}$. In **network enhancing** [135], a special random walk is used to increase gaps between leading eigenvalues of a doubly stochastic contact matrix, which makes the partition of contact maps more prominent, enhancing the boundaries for topologically associated domains (TADs).

scHiCTools includes three different dimension reduction methods that use pairwise similarity matrices among the cells to embed them in a low-dimensional Euclidean space. The three dimension reduction methods are as follows. **MDS (Multidimensional scaling)** takes in a pairwise distance matrix evaluated in the original space, and embeds the data points in a lower-dimensional space which preserves the pairwise distance matrix. **t-SNE** embeds high-dimensional data in a low-dimensional space with an emphasis on preserving local neighborhood. **PHATE (Potential of Heat-diffusion for Affinity-based Trajectory Embedding)** is a dimension reduction approach that preserves both local and global similarity.

2.4.2 Details of scHi-C similarity calculation

The most important part of the approach is the calculation of scHi-C contact map similarities. We introduce the detailed algorithms in details in this subsection.

Let W denote the original scHi-C contact map of size $N \times N$, assuming the whole chromosome or the genome is split into N bins at a certain resolution. $W_{i,j}$ stands for the i-th row and j-th column of the pairwise contact (adjacency) matrix. Next, we show HiCRep score is a pairwise inner product of feature vectors from each contact map.

HiCRep [142] is based on stratum-adjusted correlation coefficient (SCC). The n-th stratum v_n (in this section, n indicates its distance from the matrix diagonal) captures all interactions with genomic distances $\in [n \times resolution, (n+1) \times resolution)$:

$$v_n = [v_{n,1}\ v_{n,2}\ ...\ v_{n,N-n}],\ v_{n,k} = W_{k,k+n}$$

In practice, we only use the first s strata since there are too few long-distance contacts. When calculating the reproducibility score between two contact maps (denoted as x and y), Pearson correlation coefficient r_n is obtained from v_n^x and v_n^y (the n-th stratum of matrix x and y), then the overall similarity is calculated from the weighed average of all strata, in which weight $\omega_n = \frac{N_n\sqrt{var(v_n^x)var(v_n^y)}}{\sum_{n=1}^{s} N_n\sqrt{var(v_n^x)var(v_n^y)}}$, N_n is the length of v_n^x and v_n^y, so $N_n = N - n$ if we do not delete any element. That is:

$$\text{SCC} = \sum_{n=1}^{s} \frac{r_n N_n \sqrt{var(v_n^x)var(v_n^y)}}{\sum_{n=1}^{s} N_n \sqrt{var(v_n^x)var(v_n^y)}}$$

However, if we denote the t-th element of v_n as $v_{n,t}$ and length of v_n as N_n, and define $var(v_n) = \frac{\sum_{t=1}^{N_n}(v_{n,t} - \overline{v_n})^2}{N_n}$, then:

$$r_n = \frac{\sum_{t=1}^{N_n}(v_{n,t}^x - \overline{v_n^x})(v_{n,t}^y - \overline{v_n^y})}{\sqrt{\sum_{t=1}^{N_n}(v_{n,t}^x - \overline{v_n^x})^2 \sum_{t=1}^{N_n}(v_{n,t}^y - \overline{v_n^y})^2}} = \frac{\sum_{t=1}^{N_n}(v_{n,t}^x - \overline{v_n^x})(v_{n,t}^y - \overline{v_n^y})}{N_n\sqrt{var(v_n^x)var(v_n^y)}}.$$

So overall similarity between contact maps x and y is actually a simple inner product of concatenated strata vectors divided by a constant related to variances and lengths of strata:

$$r_{xy} = \sum_{n=1}^{s} \frac{r_n N_n \sqrt{var(v_n^x)var(v_n^y)}}{\sum_{n=1}^{s} N_n \sqrt{var(v_n^x)var(v_n^y)}} = \frac{\sum_{n=1}^{s}\sum_{t=1}^{N_n}(v_{n,t}^x - \overline{v_n^x})(v_{n,t}^y - \overline{v_n^y})}{\sum_{n=1}^{s} N_n\sqrt{var(v_n^x)var(v_n^y)}}.$$

Here we can see the numerator is exactly the inner product of all s strata (subtracted by its mean) and the denominator is some normalization factor which can also be represented as an inner product. Therefore, we are able to implement a fast version of HiCRep, named **fastHiCRep**, by first presenting the contact maps as vectors and then calculating SCC as an inner product. Further more, we suspect that simply picking s strata and then normalizing their individually, the model can also achieve good performance. We named the new method

InnerProduct, which also starts calculation from strata v_n (n from 1 to s). Then, z-normalization is applied to each v_n to get a zero-mean and unit-variance vector v'_n. By concatenating all strata, we could obtain the feature vector for each contact map: $V_{map} = [v'_1\ v'_2\ ...\ v'_s]$. That is

$$v'_{n,t} = \frac{v_{n,t} - \overline{v_n}}{\sqrt{var(v_n)}}.$$

If we directly calculate the inner product of the two feature vectors V_x and V_y of map x and y, then we can find it's also a kernel defined by inner product

$$r_{xy} = \sum_{n=1}^{s} \sum_{t=1}^{N_n} \frac{(v^x_{n,t} - \overline{v^x_n})(v^y_{n,t} - \overline{v^y_n})}{\sqrt{var(v^x_n)var(v^y_n)}}.$$

By doing this for all chromosomes, we can obtain a kernel matrix for each chromosome. Taking average keeps the matrix positive definite, and thus gives us an overall kernel matrix of all cells in the dataset. However in practice, although taking median may make the matrix no longer positive definite, sometimes it has the potential to remove outliers and improve the result.

The intuition of **Selfish** is that although total contacts vary between different contact profiles, the relative contact strength between regions is supposed to be consistent if two contact maps are similar. When comparing two regions i and j, they set value S_{ij} in the fingerprint matrix as $I(reads\ in\ i > reads\ in\ j)$, so that all ratios of contacts between two regions are binarized to 0 or 1, which could possibly result in information loss. This approach is also proved to be not as good for scHi-C embedding.

2.4.3 Benchmarking: scHiCTools calculates the similarity among single-cell Hi-C contact maps and produces satisfactory projections

We benchmarked the projection performance and run time of these methods on a recent scHi-C dataset [98], exactly following the evaluation procedure in a recent work [82]. In [82], HiCRep + MDS was shown to be able to embed scHi-C data [99] into 2-D space to obtain a circular pattern, with different stages of cell cycle correctly projected (Fig. 1a). This provides us a way of mapping scHi-C data to pseudo-time throughout the cell cycle. Also, in the Nagano dataset [99], the cells are labeled with 4 different stages (i.e. G1, early-S, mid-S, late-S/G2) through the cell cycle. Thus, we can evaluate the algorithms by treating the embedding as a clustering task. The specific steps are

- Calculate the pairwise distance of all cells in the Nagano dataset;

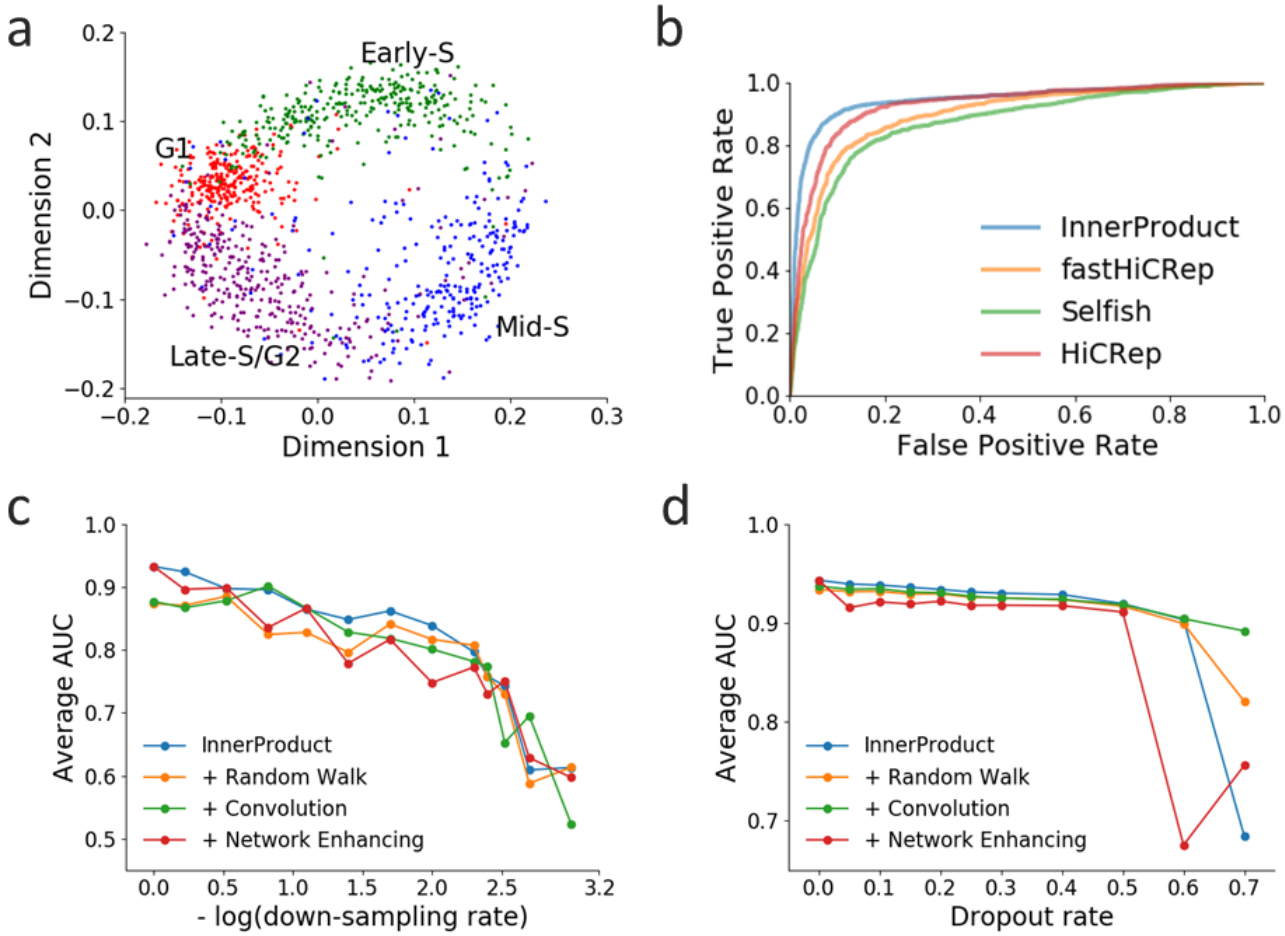

Figure 2.3: **Benchmarking experiment results. a.** The embedding of single cells in a cell cycle study [98]. **b.** Evaluating the three embedding methods with a cell-cycle phasing task by average ROCs. **c.** Smoothing methods do not perform well when all positions in Hi-C maps are randomly downsampled. The x-axis is the negative logarithm of sampling rates; y-axis is the average AUCs from ROC curves. **d.** Linear convolution improves the performance of embedding when the dropout rate is high.

- Use MDS to get a 2-D embedding and transform it into polar coordinates;
- Only keep the angular term and fit each cluster with a von Mises distribution;
- Pick the means of the four distributions as cluster centroids, obtain true positive (TP) and false positive (FP) rates then calculate the ROC curve and the area under the curve (AUC).
- A higher average AUC indicates a better embedding result.

We had the following observations.

InnerProduct produced satisfactory projection. InnerProduct produced satisfactory projection of the single cells (Fig. 2.3a), achieving an average area under the ROC curve (AUC) of 0.943, which was as good as original HiCRep reported in the recent work [82]. The AUCs from fastHiCRep and Selfish were relatively lower (Fig. 2.3b). Implemented fastHiCRep did not perform as well as the original HiCRep in this task, which might be due to their subtle difference.

All the three embedding methods are efficient. The run time of the three methods was compared in Supplementary Table 1. Overall, the three embedding methods were efficient. For embedding 800 cells, all three methods finished within minutes up to an hour. Given the fact that all of the three embedding approaches have $O(n)$ computation complexity, they can scale up very well for a large number of cells. FastHiCRep was slightly slower than InnerProduct, which was slower than Selfish under the default parameters. Note that the run time of these approaches depends on parameter settings, which is further discussed in Supplementary Note 4.

Linear convolution smoothing and random walk improves projection at high dropout rates. We applied two sparsification methods on the scHi-C dataset [98], and applied InnerProduct together with the three smoothing approaches, and evaluated the projection performance (see Supplementary Note 5 for additional details). The first sparsification method was used to randomly reduce 40% $\sim$ 99.9% of the contacts for all positions (reducing the contact number from $\sim$200,000 to $\sim$500 in each cell). The second one was used to discard contacts from 5% $\sim$ 60% genomic loci (to simulate dropouts in sequencing data). It was observed that under the second sparsification method, linear convolution and random walk showed some consistent improvement. Linear convolution increased projection accuracy more effectively at higher dropout rates. However, none of the three improved the projection performance when the first sparsification was used.

2.4.4 Availability and future directions

Our scHiCTools is implemented in Python. The source code is available and maintained at Github: https://github.com/liu-bioinfo-lab/scHiCTools. This package is also available on PyPI python package manager. The current code runs under Python 3.7 or newer versions. Other dependency includes numpy, scipy, matplotlib, pandas, simplejson, six, and h5py. For the interactive scatter plot function, you need to have plotly installed. In the future, we will keep updating the toolbox with new scHi-C analysis algorithms, including new embedding methods such as UMAP and new clustering methods such as hierarchical clustering.

2.5 An algorithm for identifying stripes from chromatin contact maps

Although we have introduced A/B compartments, TADs, and loops in previous sections, there are additional chromatin 3D structural features that are less understood, including stripes, micro-TADs, and polycomb complexes. In this section, we will introduce our research on stripes, including what are stripes, how to identify stripes from chromatin contact maps, and the biological factors related to stripes.

2.5.1 Stripes in chromatin contact maps

Chromatin conformation capture techniques, especially proximity ligation-based methods, have revealed the hierarchical structures of DNA folding including compartments, topologically associating domains (TADs), and chromatin loops [27, 80, 102, 110, 124]. Recently, with higher-resolution contact maps generated by *in situ* Hi-C, another chromosomal structural feature, architectural stripes [134]. On the contact frequency map, stripes appear as vertical or horizontal lines extending from the main diagonal. These stripes reflect interactions between a single locus (stripe anchor) and a continuum of genomic regions. Stripes in Hi-C contact maps usually span hundreds of kilobases, and are interpreted as the result of asymmetric extrusion of CTCF and cohesin, i.e., one cohesin subunit is captured by a proximal CTCF-binding site, while another one slides across the domain; biologically, stripe anchors represent major hubs of transcription and recombination [134]. With new techniques such as Micro-C, our understanding of fine-scale 3D chromatin organization has increased to nucleosome resolution [61, 73, 107], where far more stripe patterns were discovered. Different from Hi-C stripes, these patterns are much smaller (∼10–50 kb) and frequently link genes and promoters with regulatory elements in a CTCF/cohesin-independent manner. Moreover,

*This is an ongoing work and has been presented at the 4DN annual meeting in Boston in December 2023. I proposed the initial idea and implemented the data processing and stripe statistical evaluation parts of the package.

Micro-C stripes' significant correlation with Pol II binding, accessible chromatin, and active histone marks demonstrates another mechanism of stripe formation [61].

An integrated analysis of stripes with genomic and epigenomic features at a genome-wide scale shows vast potential in understanding the cooperation between regulatory elements in 3D space [148, 61]. However, unlike compartments, TADs, and loops, there are few well-established algorithms for automatically identifying stripes in the genome ("Zebra" used by previous papers does not consider the effect of loops, TADs, and sub-compartments and requires manual removal of some stripes). Other methods require biological data such as CTCF or NIPBL chromatin immunoprecipitation sequencing (ChIP-seq) tracks in order to infer stripes from the genome [134]. These data are not necessarily available and do not match the resolution of newer, high-resolution techniques such as Micro-C, and so the need for a newer stripe calling method that can operate on a minimal amount of data is necessary.

Therefore, we have developed a Python package, Quagga, to call stripes solely from Hi-C/Micro-C/HiChIP chromatin contact maps efficiently in an unbiased way. In Quagga we propose a new, lightweight approach to determining architectural stripes that considers the local signal bias in contact matrices. We aggregated a benchmark on multiple, real Hi-C, Micro-C datasets to measure its speed and performance, and investigated the potential biological impacts of found stripes.

2.5.2 Overview of quagga

Stripes are architectural features of the genome that represent asymmetric extrusions of CTCF and cohesion, thought to play regulatory roles in a cell's development or state. Therefore, integrated analysis of stripes with genomic and epigenomic features at a genome-wide scale shows vast potential in understanding the cooperation between regulatory elements in 3D space. However, unlike compartments, TADs, and loops, there are few well-established algorithms for automatically identifying stripes in the genome ("Zebra" used by previous papers does not consider the effect of loops, TADs, and sub-compartments and requires manual removal of some stripes). To this end, we developed Quagga, a tool for the detection and statistical verification of genomic architectural stripes from Hi-C or Micro-C chromatin contact maps. Quagga relies on robust image processing techniques and poisson sampling for enrichment. Quagga outperforms other stripe detection methods in speed and accuracy and is robust when working with Hi-C or Micro-C data. Our method provides a flexible, easy-to-use tool to help scientists explore the relationships between chromatin architectural stripes and important biological questions.

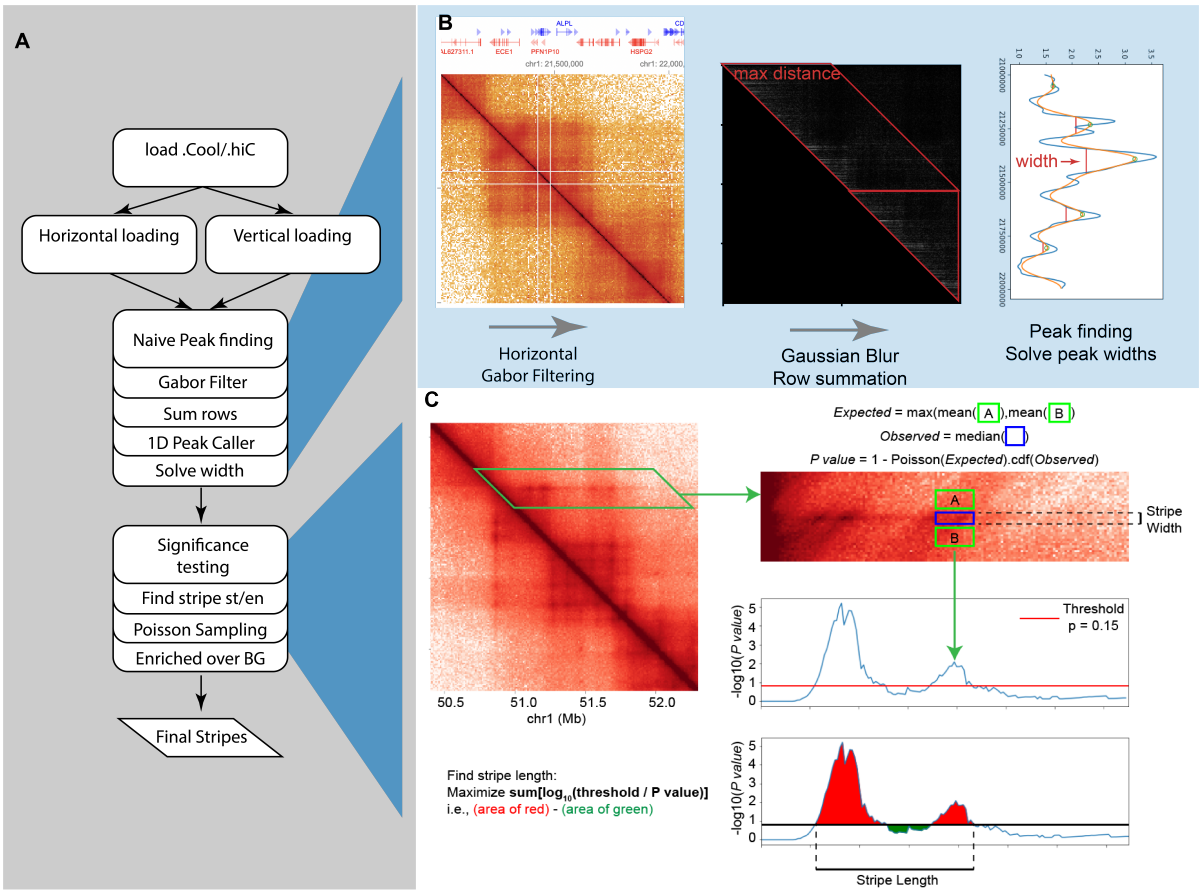

Figure 2.4: Overview of Quagga.
a. Outline of the workflow. Hi-C file via cooler file or hic file is processed into a horizontally or vertically loaded matrix and a naive peak finding algorithm is used. **b.** Stripe indices and width are calculated based on the vertically or horizontally averaged row or column sums. **c.** Significance testing is applied to called stripe peaks. Called region windows are used to determine the appropriate length of the stripe and whether or not the stripe is enriched over the local background.

Input files. The only required inputs for Quagga are a contact frequency map and the genomic assembly chromosomal lengths file. For contact frequency maps, Quagga takes any edgelist information that may be related to chromosome contact mapping. Quagga was developed and tested with Hi-C and Micro-C chromatin contact maps in hic or cooler formats. As an optional input, centromere regions may also be input to Quagga using supplying appropriate bed-files; these regions can be intentionally excluded from Quagga calls.

Preprocessing and Detection of stripe position and width. In order to determine stripe coordinate and stripe width, a rough pass over the contact map is made using the mean sum over rows or columns to form a spectrograph of contact frequencies, limited by a distance from the main diagonal of the contact frequency matrix. The user may determine to what distance from the main diagonal to use, as well as how many spaces off the main diagonal should be blanked, to diminish its dominating influence. Prior to computing row-sums or column-sums, we also apply Gabor filter to focus the contact frequency into the most essential components and denoise. Using the derived mean-frequency spectrograph, scipy's 1D Gaussian filter is applied to smooth the spectrogram, and a local peak finding algorithm is applied to find local peaks on the 1D array (Fig. 2.4A-B). Stripe width is determined by estimating the width of the peak from a relative distance to the total height of the peak call.

Stripe checking and statistical validation

For example, to identify horizontal stripes, we first calculate the observed contact value of each pixel on a candidate stripe and its neighbor regions. The observed value of the interaction between bins i and j $(i < j)$ is:

$$\text{Obs}_{i,j} = \text{median}_{j-w<k<j+w} M_{i,k}$$

in which M represents the normalized contact matrix and w is the window size. The use of the median, as opposed to the mean, mitigates the influence of significantly large values resulting from chromatin loops. The average of neighbor regions is then calculated as follows:

$$\text{Exp}_{(i,j)} = \max(\text{median}_{i-w<l<i,j-w<k<j+w} M_{l,k}, \text{median}_{i<l<i+w,j-w<k<j+w} M_{l,k})$$

in which the top and bottom neighbor regions are calculated separately. Taking a maximum of the two neighbor regions avoids TAD boundaries being called. To assess whether a pixel exhibits significantly higher contacts than its upper/lower neighbor regions, we employ Poisson statistics to derive a corresponding P value.

Due to sparsity or noise, the enrichment might not be significant for some pixels along the

stripe. Therefore, after obtaining all P values along the horizontal line, we allow breaking points when identifying stripes. The start and end positions are pinpointed as follows.

$$head, tail = \text{argmax}_{st,ed} \sum_{st}^{ed} \log \frac{thr}{P_i},$$

in which *head*, *tail* are two ends of the stripe, *thr* is the threshold and P_i is the P value of pixel i. This maximization is performed with the efficient dynamic programming approach.

In the end, the significance (p-value) of the stripe is determined through the calculation:

$$P_{stripe-i} = \exp(\text{mean}_{head \leq k \leq tail}(\log P_{i,k})).$$

2.5.3 Quagga detects stripes from publicly available chromatin contact maps

We ran Quagga on Hi-C contact maps for GM12878 and K562 from Rao et. al. [110], as well as the Hi-C and Micro-C contact maps for H1 and HFFc6 from Krietenstein et. al. [73]. We selected GM12878 and K562 due to their high sequencing depth and available orthogonal datasets such as SPRITE and HiChIP. H1 and HFFc6 are chosen to unbiasedly compare Hi-C and Micro-C stripes and evaluate the default parameters for calling Hi-C and Micro-C stripes. We applied the preset parameters: 5 kb resolution, distance range within 2 Mb, 200 kb minimum stripe length, σ=2, and p=0.15 to Hi-C contact maps, and p=0.25 to Micro-C contact maps. Quagga calls 142 stripes for GM12878 and 77 stripes for K562 (Hi-C), and 372 stripes for H1 and 444 stripes in HFFc6 (Micro-C). These called stripes appear to be authentic during our visual validation.

Quagga's time and memory usage are related to parameter settings. A finer resolution and a longer distance range increases the time and memory consumption. For GM12878 running 5 kb resolution on chromosome 1 is 708 seconds, chromosomes 1 and 5 are 1,246 seconds, and 1, 5, and 10 are 1,651 seconds. The inclusion of more chromosomes also results in a longer running time and larger memory, but the time-chromosome length relationship is concave instead of linear. This is because Quagga accelerates statistical tests by storing previous results and importing pre-computed results, and therefore most statistical tests are simplified when calculating the later chromosomes. Quagga takes advantage of multi-processing, and for multicore workstations or computer clusters, it can operate parallel over a user-specified number of CPU cores.

2.5.4 Quagga detects stripes related to CTCF-cohesin extrusion

A common assumption of the Hi-C stripe mechanism is one-sided extrusion, signifying a strong correlation between Hi-C stripes and the enrichment of CTCF/RAD21 proteins, as well as the orientation of CTCF binding sites. In this section, we examine whether Quagga effectively captures these crucial epigenomic characteristics of Hi-C stripes within the GM12878 dataset. In our study, we identified 3,667 stripes at a 10 Kb resolution. The binding sites of CTCF/RAD21 are identified from available ChIP-seq data, and the orientations of CTCF binding sites were annotated based on the underlying DNA sequence motifs.

Our analysis revealed a significant enrichment of CTCF and RAD21 at the anchor points of stripes identified by Quagga (Figure 2.5a). Notably, Quagga exhibited a higher level of CTCF/RAD21 enrichment when compared to the baseline method, Stripenn, implying its superior ability to detect the characteristic CTCF/RAD21-extrusion stripes in Hi-C data. By further categorizing the Quagga-identified stripes into two groups — 3' stripes spanning downstream and 5' stripes spanning upstream, we detect distinct patterns of CTCF orientation for the two groups of stripes. Specifically, the 3' stripes exhibited a significant enrichment of forward-strand (+) CTCF orientations and a depletion of backward-strand (-) CTCF orientations, with the reverse being true for the 5' stripes. This observation serves as additional evidence that Quagga effectively identifies the prototypical CTCF/RAD21-extrusion stripes in Hi-C data.

CTCF plays a pivotal role in regulating chromosomal structure. Previous studies have demonstrated that targeted degradation of CTCF leads to the dissociation of topologically associating domains (TADs) and loop structures on a genome-wide scale. Since CTCF also contributes to the maintenance of Hi-C stripes, we investigated whether CTCF knockout results in the depletion of stripes in Hi-C contact maps.

We obtained contact maps for mouse embryonic stem cells (mESC) in both wild-type (WT) and CTCF-knockout (KO) scenarios from Nora *et al.* [103]. Using Quagga, we identified stripes at a 20 Kb resolution. In the WT mESC contact map, we detected 362 stripes, while only 122 were found in the CTCF-KO contact map, despite the latter having a deeper sequencing depth. Notably, 95.2 % (296 out of 311) of the lost stripes in the CTCF-KO contact map had a CTCF-binding anchor. This finding confirms that Quagga effectively identifies stripe depletion in Hi-C contact maps resulting from CTCF knockout.

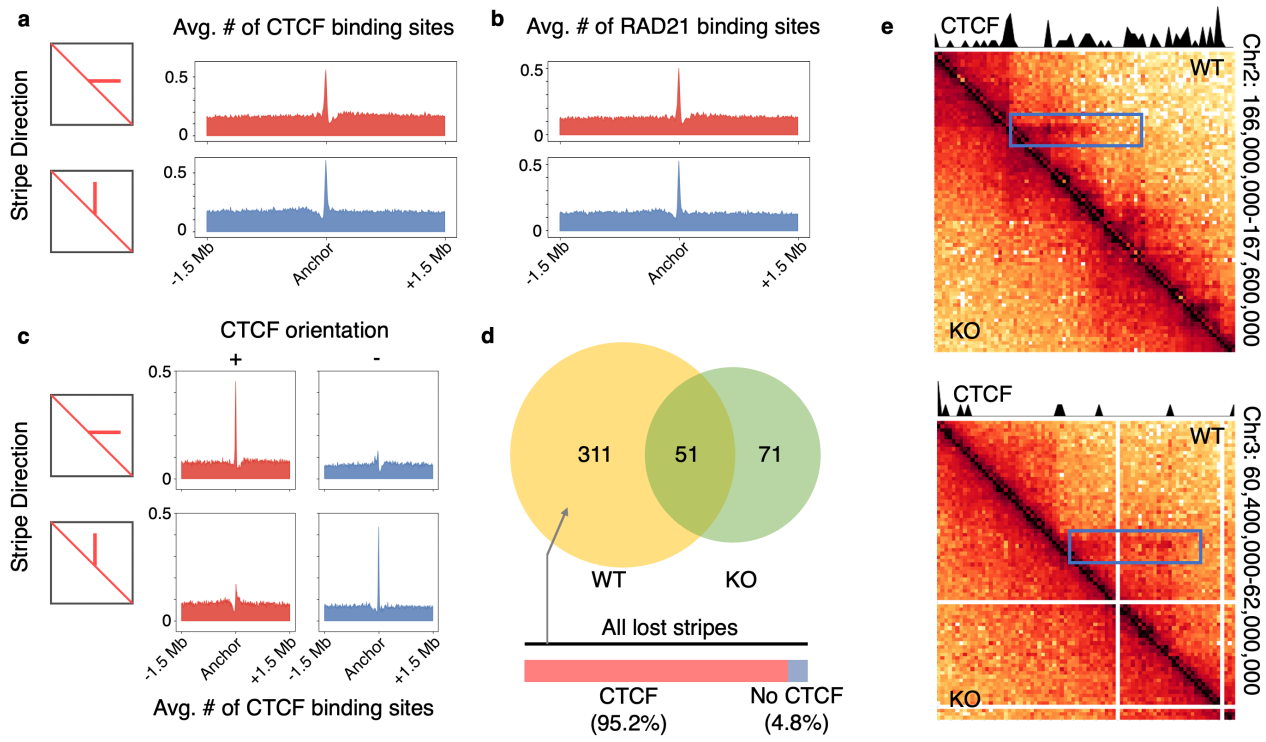

Figure 2.5: **Hi-C stripes called by Quagga are closely related to CTCF/RAD21 extrusion.**
a-b. For both 3' and 5' stripes, CTCF and RAD21 are enriched at GM12878 stripe anchors. **c.** The 3' stripe anchors are more enriched in positive-strand CTCF, and the 5' stripes are more enriched in negative-strand CTCF. **d.** Quagga identifies that the majority of Hi-C stripes will disappear after CTCF knockout, in which 95.2 % of the lost stripes are anchored at CTCF binding sites. **e.** Two example regions illustrate that Quagga identified the stripe loss after CTCF knockout.

2.5.5 Conclusion

Quagga offers a flexible, powerful method that gives researchers control to tighten the tool to search and statistically test for stripes on a variety of chromatin contact data. Quagga will allow researchers to efficiently and extensively extract features from Hi-C/Micro-C contact maps.

CHAPTER 3

Connecting High-resolution 3D Chromatin Organization with Epigenomics

Although chromosome conformation capture (3C) technologies [21] have been evolving for almost two decades, 3D chromatin structures are rarely utilized by computational models for understanding chromatin states [35, 57] due to 1) the lack of high-resolution and high-quality chromatin contact maps and 2) difficulties for jointly analyzing 1D epigenetic features with 3D chromatin structure in one model.

In this chapter, we will introduce how to use deep learning models to impute high-resolution chromatin contacts with 1D epigenomic features, and how the imputed contact maps help reveal regulatory processes in the human genome [37].

3.1 Introduction

Whereas 3D chromatin organization at the large scale of topologically associating domains (TADs) and compartments has been well characterized in many cell and tissue types by Hi-C technology [110], our understanding of fine-scale 3D chromatin organization at the nucleosome resolution has just begun [61, 73, 107]. With the increasing evidence that fine-scale chromatin organization at the nucleosome resolution is closely related to epigenomic state [125, 71], one intriguing question to ask is whether we can accurately extrapolate such high-resolution chromatin contact maps from epigenomic features such as chromatin accessibility, histone modifications, and transcription factor binding profiles. To explore this, we proposed CAESAR (Chromosomal structure And EpigenomicS AnalyzeR), a deep learning approach to predict nucleosome-resolution 3D chromatin contact maps from existing epigenomic features and lower-resolution Hi-C contact maps.

Our model leverages cutting-edge deep learning approaches to identify representations relevant to high-resolution chromatin organization. In particular, 1D convolutional and graph

convolutional layers [70] identify epigenomic patterns over the linear chromatin fiber and over the 3D spatial chromatin organization that is relevant to impute high-resolution chromatin contact maps. With existing high-resolution Micro-C contact maps, Hi-C contact maps, and a number of cell-type matched epigenomic data on human H1-hESC (hESC), mouse ESC (mESC), and human foreskin fibroblasts (HFF), we systematically evaluated the model's performance across different chromosomes, across different cell types, and across different species. In the experiments, the model accurately imputes many fine-scale chromosomal structures that Hi-C sequencing fails to detect, including short-range chromatin loops and stripes. The model is more accurate at imputing evolutionarily conserved regions, active A compartment, and early-replicating regions, which indicates that the fine-scale 3D chromatin organization is strongly influenced by the nature of the epigenomic factors in these regions. The imputed chromatin contacts also recapitulate enhancer activities previously elucidated by CRISPRi experiments [44], and manifest expression quantitative trait loci (eQTLs) previously profiled by GTEx project [85]. CAESAR is also coupled with an attribution method which identifies epigenomic features explanatory to these fine-scale 3D chromatin structures. The explanatory features help to further subtype fine-scale chromatin structures and elucidate the interplay between histone modifications and nucleosome-level chromatin organization.

CAESAR connects 3D genome organization with epigenomics at nucleosome resolution and unprecedented scale. First, compared with previous computational models for imputing Hi-C contact maps, such as HiCPlus [149], HiCGAN [84], and HiC-Reg [146], CAESAR reaches a much higher resolution. Since the majority of epigenomic activities (TF binding and histone modifications) take place at the nucleosome resolution, it is desirable to develop the predictive model that connects epigenomics and chromatin organization at the nucleosome resolution. Second, although previous models EpiTensor [151] and DeepTACT [77] also reconstruct sparse 3D chromatin interactions from epigenomics at an ultra-high resolution, CAESAR learns from real Micro-C contact maps and predicts all chromatin contacts within a distance range, which reveals diverse fine-scale structures such as stripes, TADs, and polycomb interactions between repressive regions. Third, different from Akita [41] and DeepC [123] which predict chromatin contact maps from conserved DNA sequences, CAESAR generates tissue-specific or cell line-specific predictions from epigenomic features. Therefore, it imputes an unprecedented number of high-resolution human chromatin contact maps, including 57 tissue samples, 16 cell lines, 12 primary cells, and 6 *in vitro* differentiated cells. The imputed high-resolution contact maps are shared on a web server (https://nucleome.dcmb.med.umich.edu/), which allows users to easily navigate these fine-scale chromatin structures and the corresponding explanatory epigenomic features. In ad-

dition, CAESAR includes an attribution component, which reveals detailed relationships between 3D chromatin organization and epigenomic features.

3.2 A deep learning model imputing high-resolution chromatin contact maps

We proposed CAESAR, a supervised deep learning model to impute chromatin contact maps at nucleosome resolution. CAESAR's inputs include a lower-resolution Hi-C contact map and a number of histone modification features (e.g., H3K4me1, H3K4me3, H3K27ac, and H3K27me3), chromatin accessibility (e.g., ATAC-seq), and protein binding profiles (e.g., CTCF) (Supplementary Note 2). CAESAR captures the Hi-C contact map as a graph $\mathcal{G}$ with nodes representing genomic regions of 200 bp long, weighted edges representing chromatin contacts between the regions, and N epigenomic features modeled as N-dimensional node attributes. The architecture of CAESAR (Figure 3.1a) includes ordinary 1D convolutional layers which extract local epigenomic patterns along the 1D chromatin fiber, and graph convolutional layers which extract spatial epigenomic patterns over the neighborhood specified by $\mathcal{G}$. The concatenated outputs from the convolutional layers capture all relevant features for one particular 200 bp bin, which are further fed into two parallel output layers — a fully-connected layer predicts the contact profile for each 200 bp bin, and an inner product layer predicts loops between bins. The outputs from the fully-connected layer and the inner product layer are summed up as CAESAR's final output. Using Micro-C contact maps from hESC, mESC, and HFF as the prediction target, the model was trained with backpropagation [137], in which the aforementioned convolutional features were learned adaptively. Other than leveraging a number of epigenomic features, our model architecture differs from HiCPlus [149] and DeepHiC [59] which treats Hi-C contact maps as images and performs grid-convolution to improve the resolution. With the graph convolutional networks and additional epigenomic features, CAESAR not only enhances the resolution of contact maps, but also predicts the structures which are not captured by Hi-C, including polycomb repressive regions, short-range loops and stripes (Figure 3.1b).

3.2.1 Detailed model architecture

The model includes two major parts — one for predicting chromatin loops, and the other for predicting contact profile. Each part includes consecutive input layers, convolutional layers, and output layers. CAESAR captures the interpolated Hi-C contact map as a graph $\mathcal{G}$ with nodes representing genomic regions of 200 bp long, and weighted edges representing

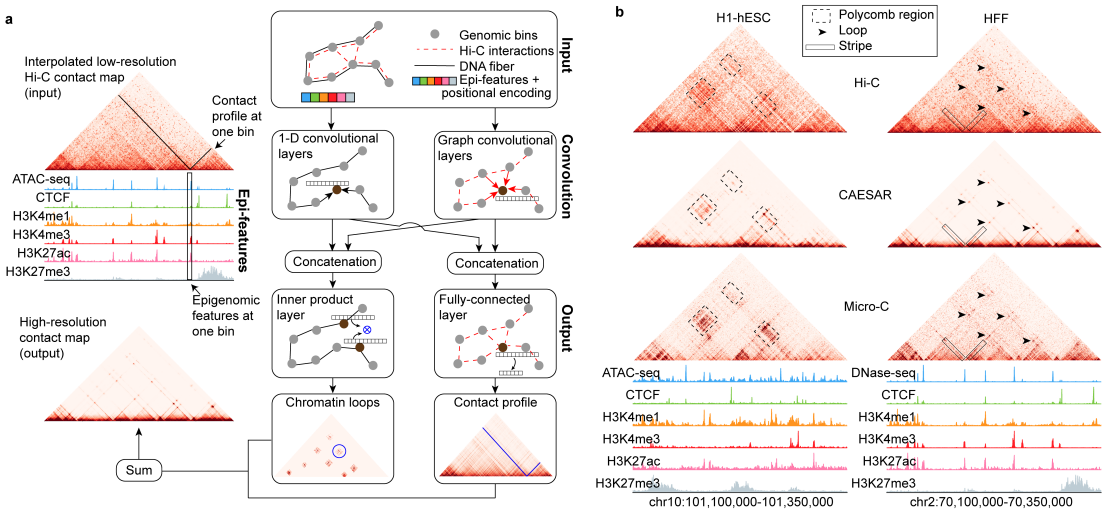

Figure 3.1: **Overview of the model.**
a, Model architecture. The model inputs are a Hi-C contact map and a number of epigenomic features including histone modifications, chromatin accessibility, and protein binding profiles. The lower-resolution Hi-C contact map is first interpolated into a 200 bp-resolution contact map, and then transformed into a graph $\mathcal{G}$ in which the nodes represent 200 bp genomic bins and the edges represent the interpolated contacts between the nodes. Positional encoding is unrelated to Hi-C or epigenomic data and only encodes node order in the genome. The epigenomic features and positional encoding are assigned to the corresponding nodes as node attributes. The inputs are fed into 1D convolutional and graph convolutional layers to generate hidden representations, which extract features from both nearby genomic regions along the 1D DNA sequence and spatially-contacting regions specified by $\mathcal{G}$. The output layers take input the hidden representations and predict the contact profile at each 200 bp bin as well as the chromatin contacts between bins. **b**, In an example region, the polycomb interactions are accurately predicted by CAESAR. In another example region, loops and stripes undetected by Hi-C are accurately predicted by CAESAR.

chromatin contacts. A is the adjacency matrix of $\mathcal{G}$. For both parts, the inputs include the graph adjacency matrix A and the epigenomic features X. As one 250 kb region is fed into the model each time, the dimension of the input adjacency matrix is 1250×1250. In a 6-epigenomic model, the size of the epigenomic feature matrix is 6×1250. In addition, eight positional encoding dimensions are concatenated to the epigenomic features. The positional encoding is calculated with the following method, in which *pos* is from 0 to 1249 and i is from 0 to 7 [133].

$$PE_{(pos,2i)} = \sin\left(pos/10000^{2i/8}\right)$$
$$PE_{(pos,2i+1)} = \cos\left(pos/10000^{2i/8}\right)$$

In deep learning models, convolutional kernels are small filters sliding through the input to extract certain patterns. When the filter is applied to an input element, it calculates the weighted sum of the element with its local neighbors. In a convolutional layer, multiple kernels work in parallel to learn different sets of weights and extract different patterns. There are two types of convolutional layers, 1-D convolutional (Conv1D) and graph convolutional (GC) layers in CAESAR. Conv1D layers operate along the genome fiber, aggregating the epigenomic features from nearby bins. GC layers extract spatial epigenomic patterns over the spatial neighborhood specified by $\mathcal{G}$. Here, we use the GC layer

$$Y = \sigma(\tilde{A}XW)$$

in which X and Y are the input and output, $\tilde{A}$ is the normalized graph adjacency matrix, W is the trainable parameters, and σ is the *relu* activation function [100]. GC layers provide additional structural patterns for imputing high-resolution chromatin architecture. For example, if two distant loci i and j are in the same TAD, then nodes i and j are neighbors on the graph. Therefore, when we predict the contact profile of i, the information flows from j to i in the GC layers, so that the features at j contribute to the prediction of i, and *vice versa*. The window size for each 1-D convolution kernel is 15 in the contact profile predicting part and 5 in the loop predicting part, which captures relevant features from a 3 kb and 1 kb neighborhood, respectively.

For the contact profile predicting part, the output layer is a fully-connected layer. The input of this layer is the concatenation of convolutional layers' outputs and the Hi-C contact profile, and the output is the imputed contact profile of each 200 bp bin. For the loop predicting part, the output layer is an inner product layer. This layer also takes the concatenation of convolutional layers' outputs as input, and calculates the inner product between each bin pairs' representation to predict the chromatin loops. The outputs of the two output layers are summed up to generate the final imputation result. The model includes 2 million

parameters, which is much fewer than the number of elements (∼15 billion) in the contact matrix.

3.3 Evaluation of the deep learning algorithm

3.3.1 Accurately predicting high-resolution chromatin contact maps

With existing Micro-C data on mESC, hESC, and HFF, we evaluated CAESAR in three different sets of experiments, including a cross-chromosome experiment, a cross-cell type experiment, and a cross-species experiment, so as to evaluate the model's generalizability in different scenarios. In the cross-validation experiment on hESC, we divided the human chromosomes into a train set, a test set, and a tune set of similar sizes. CAESAR and two baseline models, including HiCPlus [149] which only used low-resolution chromatin contact maps, and HiC-Reg [146] which only used epigenomic features, were trained with the train set and evaluated with the test set. We used the tune set to tune hyperparameters. For CAESAR and HiC-Reg, 6 epigenomic features were used, including ATAC-seq, CTCF, H3K4me1, H3K4me3, H3K27ac, and H3K27me3. CAESAR outperformed HiCPlus and HiC-Reg in terms of the stratum-adjusted correlation coefficient (SCC) with the observed Micro-C contact map (Figure 3.2a). The results demonstrated that it is necessary to leverage both the contact maps and epigenomic features in the prediction of high-resolution contact maps. In the cross-cell type experiment, we used the same train set of chromosomes to build a model on HFF, and then tested it on hESC with the same test set of chromosomes as in the cross-chromosome experiments. The HFF-trained model imputed almost as well as the hESC-trained model for chromatin contacts within 100 kb and 200 kb range (Figure 3.2b). In the cross-species experiment, we trained the model on mESC and tested the performance on hESC. In order to stay consistent with cross-chromosome and cross-cell-type evaluation, we also divided mouse chromosomes into train, tune, and test sets of similar sizes. We trained the model with mESC's train set and then tested its performance on the same aforementioned test set of hESC. It was observed that the model trained on mESC also moderately generalized to hESC, and the generalization deteriorates as the contact distance increases.

In addition, we tested CAESAR's performance in predicting fine-scale structures including loops and stripes. In the test set of HFF, CAESAR captured 72% of the loops and 63% of the stripes from Micro-C contact maps, whereas only less than 1% were captured from the input Hi-C contact maps (Figures 3.2c and 3.2e). Since loops called from two Hi-C replicates only agree ∼60% [114], we believe that our imputed contact map recovers a good

portion of these fine-scale structures. By piling up all the loop and stripe regions called from the Micro-C contact maps, we observed comparable enrichment from our predicted high-resolution contact maps and the observed Micro-C contact maps, but the pile-up results from the input Hi-C contact maps showed little enrichment (Figures 3.2d and 3.2f). Chromatin contact maps imputed by CAESAR also show comparable cell-type variability as real Micro-C contact maps in terms of SCC and cell type-specific fine-scale structures including chromatin loops and stripes.

3.3.2 Factors influencing CAESAR's performance

In order to optimize CAESAR's efficiency, we next explored the factors influencing its performance. As CAESAR's principle inputs are epigenomic and Hi-C data, we began by evaluating the minimum required number of datasets to achieve good imputed results. Four sets of epigenomic features were chosen based on common availability (Figure 3.3a), and we observed comparable performance among the 13-epi, 7-epi, 6-epi, and 3-epi models (Figure 3.3b). Although the SCC of the 3-epi model (including ATAC-seq, CTCF, and H3K27ac) did not drop significantly, it over-predicted fine-scale structures. Therefore, we recommend using the commonly profiled 6 epigenomic features in CAESAR. We also asked what is the requirement for input Hi-C contact maps. Using Hi-C data from Rao et. al. [110] and Krietenstein et. al. [73], we tested four contact maps, including the original Hi-C contact maps with around 1 billion contacts, two down-sampled Hi-C contact maps with 100 million and 10 million contacts, and a surrogate Hi-C contact map with 1 billion contacts aggregated from four unmatched cell lines. The surrogate contact map acts as a replacement when no chromatin contact map is available for a particular cell type. Although the SCC curve does not drop significantly with the down-sampled contact maps, surrogate Hi-C performs better (Figure 3.3c). The model trained with surrogate Hi-C can still capture 69% of the loops and 61% of the stripes from Micro-C contact maps in the test set. Therefore, if the matched Hi-C contact map is unavailable to complement the epigenomic data in a particular analysis, a surrogate contact map can be used in CAESAR.

We further investigated the relationship between CAESAR's performance, measured with Spearman's correlation between the imputed and the observed Micro-C contact maps, and evolutionary conservation, measured with phastCons scores. It was observed that the model imputed more accurately in the regions with higher evolutionary conservation (Figure 3.3d). In addition, we also discovered that the model imputes more accurately in A compartment than B compartment, and in early-replicating regions than late-replicating regions (Figures 3.3e and 3.3f). The results indicate that fine-scale chromatin organization is more closely

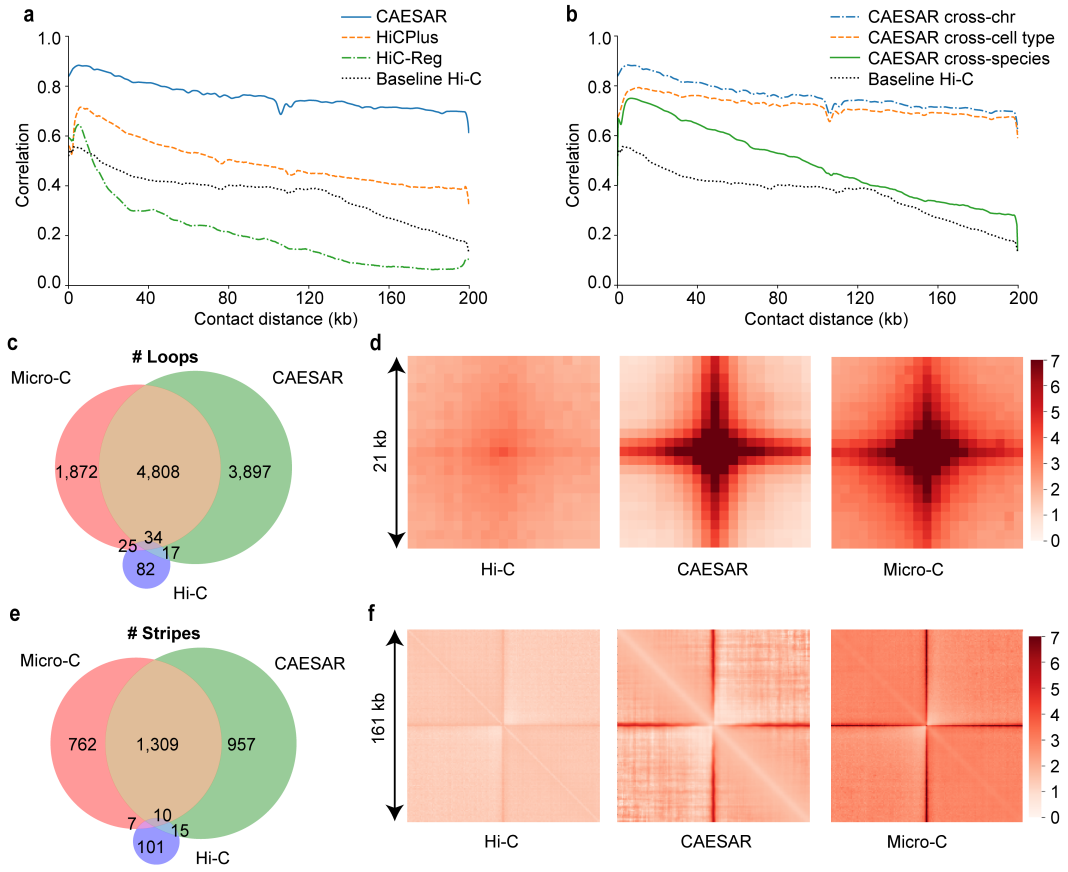

Figure 3.2: **Evaluating CAESAR's performance in multiple tasks.**
a, The distance-stratified Pearson's correlation with the observed Micro-C contact map from CAESAR and two baselines, HiC-Reg and HiCPlus, in a cross-chromosome experiment. The black dotted lines in **a** and **b** are the correlation between the input Hi-C contact map and the observed Micro-C contact map. **b**, The distance-stratified Pearson's correlation with the observed Micro-C contact map from CAESAR in 1) a cross-chromosome experiment (train on hESC train set and test on hESC test set), 2) a cross-cell type experiment (train on HFF train set and test on hESC test set), and 3) a cross-species experiment (train on mESC train set and test on hESC test set). **c**, The Venn diagram of the loops called from 1) the input Hi-C contact map, 2) the CAESAR-imputed contact map, and 3) the observed Micro-C contact map. **d**, The pile-up visualization of the loops called from 1) the input Hi-C contact map, 2) the CAESAR-imputed contact map, and 3) the observed Micro-C contact map. **e**, The Venn diagram of the stripes called from 1) the input Hi-C contact map, 2) the CAESAR-imputed contact map, and 3) the observed Micro-C contact map. **f**, The pile-up visualization of the stripes called from 1) the input Hi-C contact map, 2) the CAESAR-imputed contact map, and 3) the observed Micro-C contact map.

related to the 6 epigenomic factors at evolutionarily conserved regions, A compartment, and early-replicating regions.

3.3.3 Recapitulating CRISPRi-validated enhancer activities

With publicly available epigenomic data, we imputed high-resolution chromatin contact maps for 15 human cancer cell lines. In some cancer cell lines, noncoding regions with their regulating genes have been interrogated by CRISPR interference (CRISPRi) technology [44]. The profiled CRISPRi score indicates genomic loci's capability to regulate an essential gene, and the peaks (both positive and negative) often correspond to enhancers and promoters.

We used the CRISPRi scores profiled near two essential genes - *MYC* and *GATA1*, to validate our imputed contact maps. On the imputed contact maps for the chronic myelogenous leukemia cell K562, *MYC* gene strongly interacts with *PVT1*, which matches with the peaks of CRISPRi scores at *PVT1* locus (Figure 3.4a). The imputed contact map also showed a significant interaction between *GATA1* and *HDAC6*, which matches the CRISPRi score peak at *HDAC6* locus (Figure 3.4b). The matching of chromatin contacts and CRISPRi score peaks demonstrates our model recapitulates gene-enhancer interactions in cancer cell lines.

3.3.4 Recovering eQTL-gene interactions

With the large-scale epigenomic data available from ENCODE and Roadmap Epigenomics Project, we imputed the high-resolution contact maps for 57 human tissue samples and 2 cell lines – IMR-90 and GM12878 (Supplementary Tables 4a and 4b). With eQTLs profiled by GTEx [85], we asked whether our imputed chromatin contacts are enriched between genes and their eQTLs in the corresponding tissue or cell line. Previous works [145] have shown eQTLs are enriched in tissue-specific frequently interacting regions on Hi-C contact maps at 40 kb resolution, but a large portion of eQTLs reside too close to their gene transcriptional start sites (TSS) to be seen on a low-resolution contact map (Figure 3.5a). For example, two eQTLs that are specific in pancreas and lung respectively both locate in chr16:57,950,000-58,050,000. The loop between the pancreas-specific eQTL and its target *USB1* gene can only be called from the CAESAR-imputed contact maps of pancreas. The loop between lung-specific eQTL and its target *TEPP* gene can be called from the CAESAR-imputed contact maps of both lung and pancreas, which demonstrates some tissue-specific eQTLs do not necessarily correspond to exclusive loops in the tissue (Figure 3.5a).

To evaluate the overall contact enrichment between eQTLs-TSS pairs, we piled up the contact regions between tissue-specific eQTLs and their gene TSS. In the pile-up results

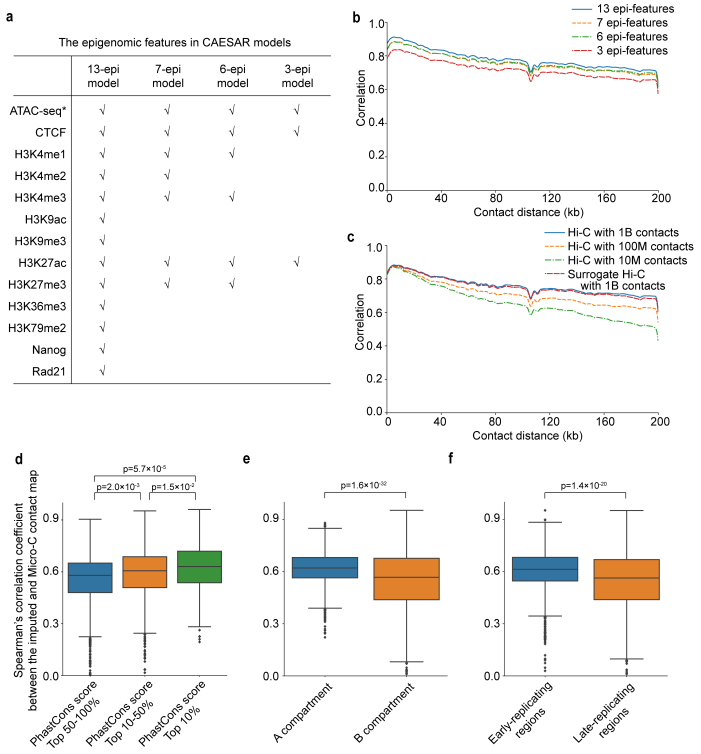

Figure 3.3: **The relationships between CAESAR's performance with Hi-C quality, the number of epigenomic features, evolutionary conservation, A/B compartments, and early/late replication timing.**
a, The epigenomic features in 13-epi, 7-epi, 6-epi, and 3-epi CAESAR models are listed in the table, which are chosen based on common availability. **b**, The distance-stratified Pearson's correlation with the observed Micro-C contact map from CAESAR in a cross-cell type experiment with different numbers of epigenomic features (i.e., 13, 7, 6, and 3). **c**, The distance-stratified Pearson's correlation with the observed Micro-C contact map from CAESAR in a cross-cell type experiment when 1) using the original Hi-C contact map with about 1 billion contacts, 2) randomly down-sampling the Hi-C contact map at different down-sampling rates (resulting in 100 million and 10 million chromatin contacts), and 3) using a surrogate Hi-C contact map with 1 billion contacts aggregated from HFF, GM12878, IMR-90, and K562 with equal proportions. **d**, The model performance in a specific region is quantified by the Spearman's correlation coefficient between the CAESAR-imputed and the Micro-C contact map. In cross-chromosome and cross-cell-type experiments, the model performance (i.e., Spearman's correlation coefficient) is significantly correlated with evolutionary conservation evaluated by sequence alignment scores (n[regions]=1,203, 960, and 240, one-sided t-test). In all the boxplots, the center line indicates median; the box limits are upper and lower quartiles; the whiskers are 1.5×interquartile range; the points are outliers. **e**, In cross-chromosome and cross-cell-type experiments, the correlation coefficient is significantly larger in A compartment than in B compartment (n[regions]=1,018 and 1,388, one-sided t-test). **f**, In cross-chromosome and cross-cell-type experiments, the correlation coefficient is significantly larger in early-replicating regions than in late-replicating regions (n[regions]=1,203, 960, and 240, one-sided t-test).

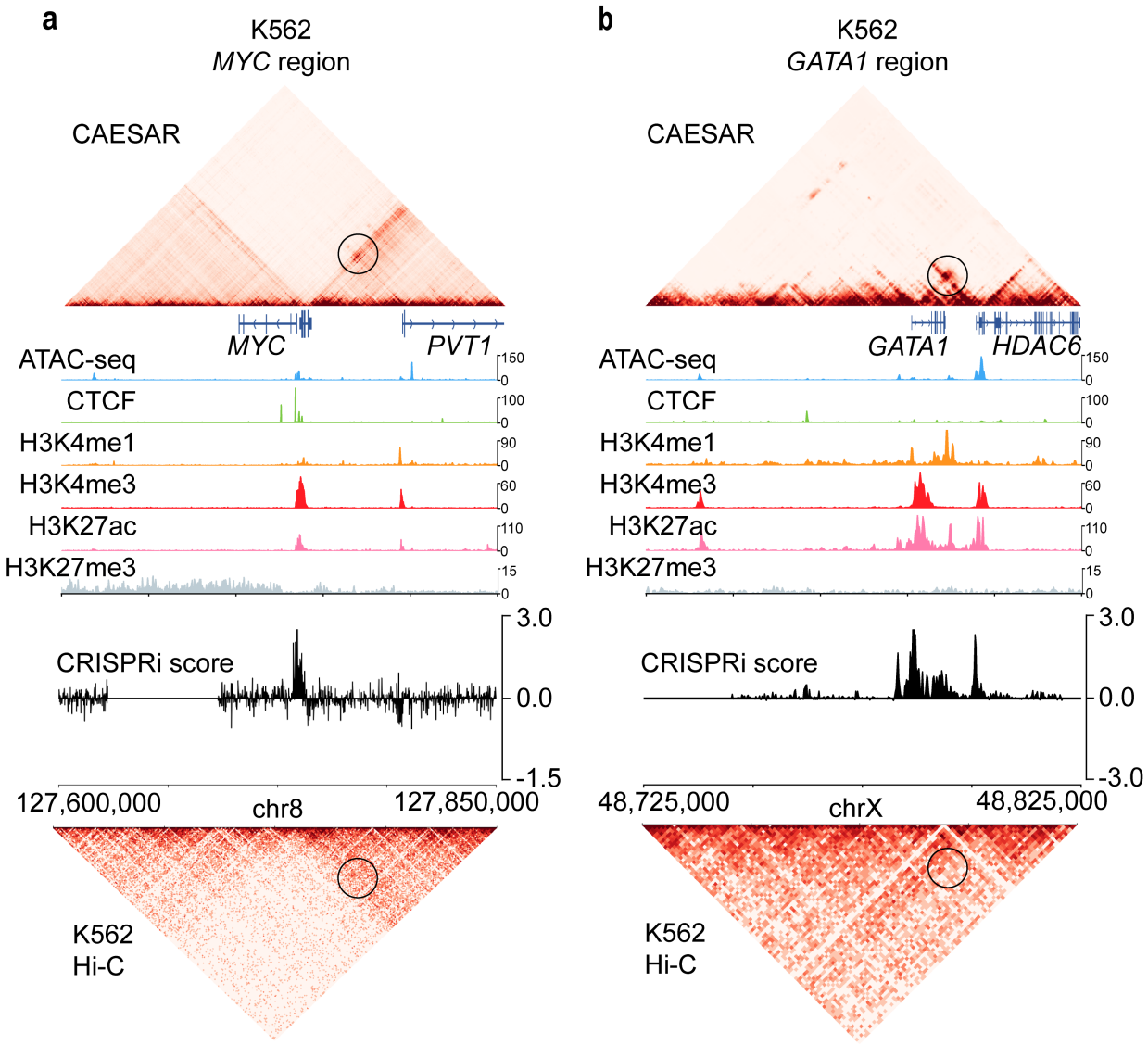

Figure 3.4: **The interactions between genes and their CRISPRi-validated enhancers in CAESAR-imputed contact maps.**
a, The CAESAR-imputed contact map of K562 at *MYC* region (chr8: 127,600,000-127,850,000) demonstrates significant contacts between *MYC* and *PVT1*, which agree with with CRISPRi score peaks, but are not shown on the original input Hi-C contact map. The magnitude of the epigenomic features is the observed value divided by the genome-wide average. **b**, The CAESAR-imputed contact map of K562 at *GATA1* region (chrX: 48,725,000-48,825,000) demonstrates significant contacts between *GATA1* and *HDAC6*, which agree with with CRISPRi score peaks, but are not shown on the original input Hi-C contact map.

of twelve tissue/cell lines, seven CAESAR-imputed contact maps (adrenal gland, heart left ventricle, IMR-90, pancreas, sigmoid colon, spleen, and transverse colon) have the highest contact values for their tissues/cell line-specific eQTL-TSS interactions. Another four CAESAR-imputed contact maps (GM12878, lung, stomach, and tibial nerve) also have close-to-highest contact values for their tissues/cell line-specific eQTL-TSS interactions. These results demonstrate that tissue/cell line-specific enhancer-promoter interactions are recovered by CAESAR. In addition, the moderate enrichment on Micro-C and CAESAR-imputed contact maps from unmatched tissue/cell lines further demonstrates the eQTL-TSS interactions are not necessarily exclusive even if the eQTLs are tissue or cell line-specific (Figure 3.5b). This suggests that some fine structural interactions are conserved across tissues or cell types but the regulatory functions remain specific.

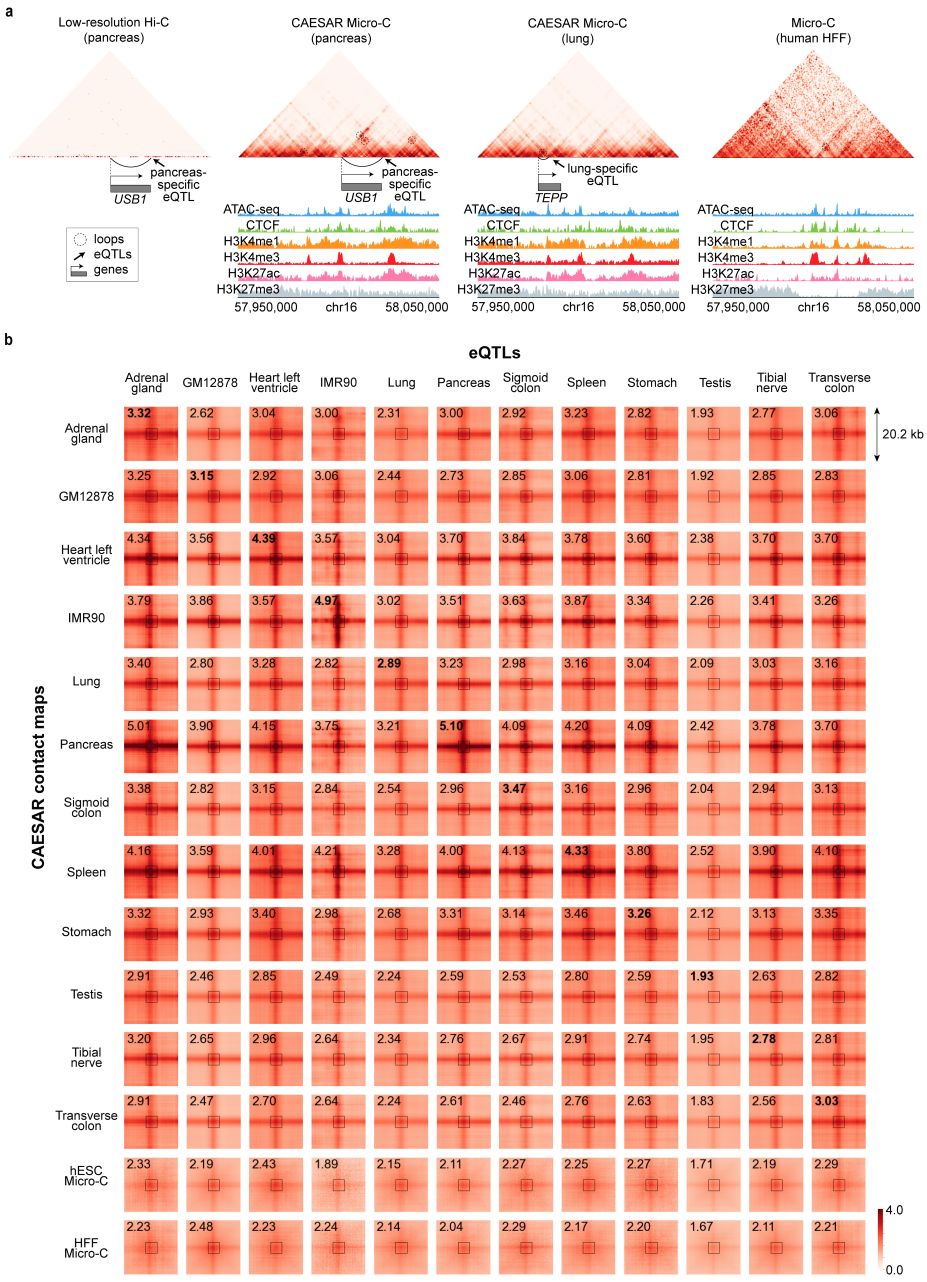

Figure 3.5: **The enrichment of eQTL-gene interactions in CAESAR-imputed contact maps.**
a, The loop between gene *USB1*'s TSS and its pancreas-specific eQTL, which cannot be observed on the original Hi-C contact map, appears on the CAESAR-imputed contact map for pancreas. Although gene *TEPP*'s eQTL is lung-specific, the corresponding loop can be called from the CAESAR-imputed contact maps for both lung and pancreas. **b**, Pile-up analysis of the chromatin contacts between eQTLs and their corresponding gene TSS from twelve different human tissues and cell lines on CAESAR-imputed contact maps and Micro-C contact maps. The average contact values in the central 5×5 squares are marked on the plots, in which the bold fonts indicate that eQTLs and CAESAR-imputed contact maps are from the same tissue/cell line.

3.4 Imputing high-resolution contact maps for more than 90 human samples

As the cross-cell type model is validated, we used the trained model to impute high-resolution chromatin contact maps for other human tissues and cell lines. We collected the epigenomic signals from a total number of 57 tissue samples, 16 cell lines, 12 primary cells, and 6 *in vitro* differentiated cells (Table 3.1). If the ATAC-seq signal was unavailable, DNase-seq was collected as an alternative. The 6-epi CAESAR model trained with both hESC and HFF's train set was used. For IMR-90, GM12878, and K562, we used their deeply sequenced (above 1B contacts) Hi-C contact maps as input. For cell lines or tissues without Hi-C or with only shallowly sequenced Hi-C, we used the surrogate Hi-C as input.

The imputed high-resolution contact maps are shared on a web server (https://nucleome.dcmb.med.umich.edu/), which allows users to easily navigate these fine-scale chromatin structures, and the corresponding explanatory epigenomic features. The back-end of the server uses python *Flask* with *sqlite*. The front-end of the server uses *bootstrap* framework. The web server utilizes multi-threading to allow multiple users to access it at the same time. Our web server processes host data at multiple ports at localhost. We use *Nginx* to perform the reverse proxy that passes internet requests to them. After contact maps are generated, we run *Nucleome Browser* on our web server. Nucleome Browser is an open platform to integratively and interactively browse coordinate-based genome data. Nucleome Browser extends conventional track-based genome browsing to panel-based genome browsing, thus breaks the linear limitation of stacked tracks view mode. Different panel modules host and render different modality data including visualized tracks and reconstructed 3D chromatin structures.

3.5 Identifying epigenomic features relevant to fine-scale 3D chromatin organization

Although deep learning models are often referred to as "black boxes", their outputs can be traced back and interpreted. In our model, we used *integrated gradient* [128] to attribute the predicted chromatin contacts to each genomic locus of each input epigenomic feature. The attribution results illustrate which parts of the epigenomic features are the most determinative for the model's predictions. By attributing the entire contact map to all epigenomic features, we evaluated the overall contribution for each feature, and low attribution is another reason for leaving H3K4me2 out from the 7-epi model besides limited availability.

Table 3.1: CAESAR-imputed tissues and cell lines

Tissue		
Adrenal gland	Ascending aorta	Body of pancreas
Breast epithelium	Esophagus muscularis mucosa	Esophagus squamous epithelium
Gastrocnemius medialis	Gastroesophageal sphincter	Heart left ventricle
Lung	Ovary	Pancreas
Peyer's patch	Prostate gland	Right atrium auricular region
Sigmoid colon	Spleen	Stomach
Suprapubic skin	Testis	Thoracic aorta
Thyroid gland	Tibial artery	Tibial nerve
Transverse colon	Upper lobe of left lung	Uterus
Vagina		
Cell line		
A549	A673	GM12878
GM23338	HCT116	HeLa-S3
HepG2	IMR-90	K562
Karpas-422	MCF-7	MM1S
OCI-LY7	PC-3	PC-9
SK-N-SH		
Primary cell		
B cell	CD14-positive monocyte	Astrocyte
Endothelial cell of umbilical vein	Fibroblast of dermis	Fibroblast of lung
Foreskin fibroblast	Foreskin keratinocyte	Keratinocyte
Mammary epithelial cell	Osteoblast	Skeletal muscle myoblast
In vitro differentiated cell		
Bipolar neuron	Cardiac muscle cell	Hepatocyte
Myotube	Neural progenitor cell	Smooth muscle cell

This method can be applied to arbitrary regions on the contact map, which allows us to connect fine-scale structures with the most explanatory epigenomic features. Surprisingly, many of the peaks in the input epigenomic features do not necessarily help the model to predict fine-scale structures. For example, the H3K27ac peaks showed negative attribution in predicting the stripe in Figure 3.6a and the loop in Figure 3.6b. With attribution calculated by *integrated gradient*, the predicted chromatin structures can be further analyzed and subtypes.

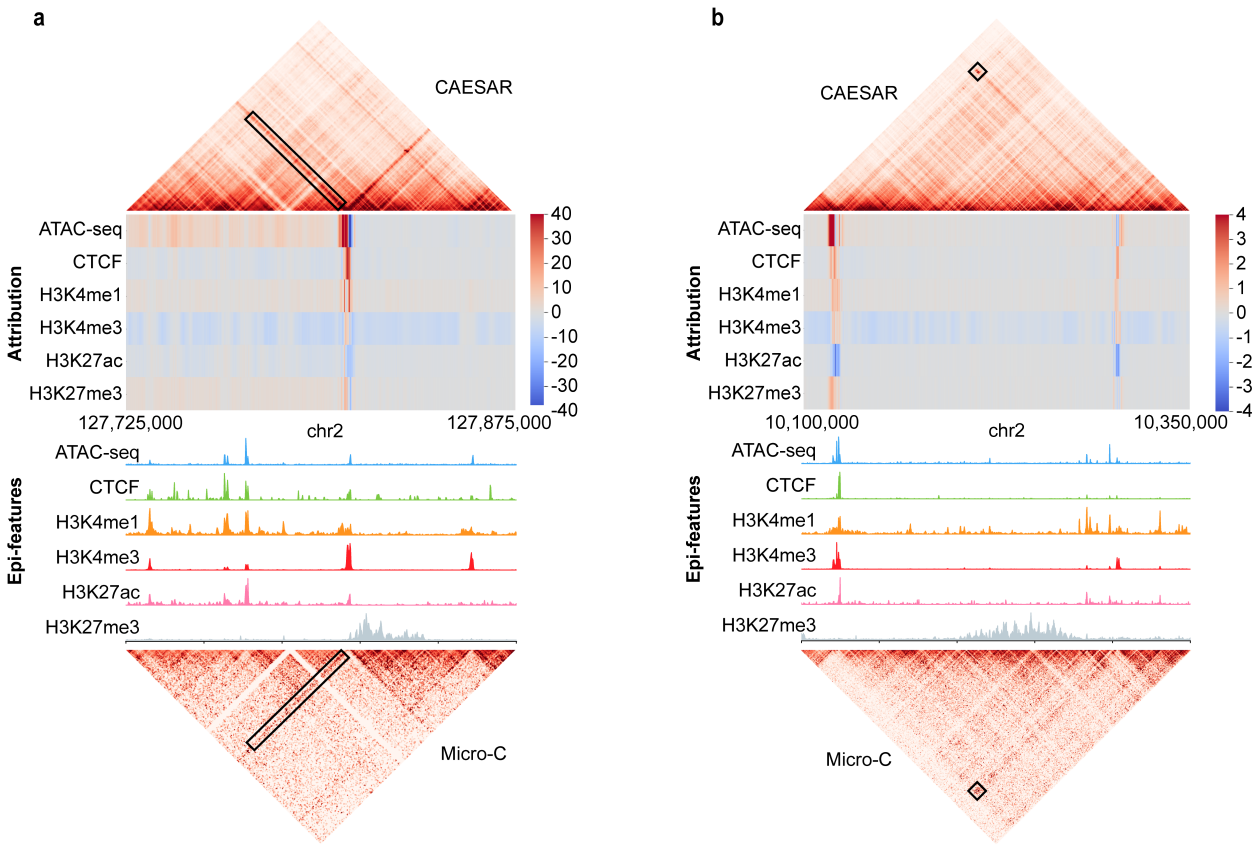

Figure 3.6: Attributing CAESAR outputs to epigenomic features via *integrated gradient*. Larger attribution magnitudes indicate more contribution to the model's prediction.
a, The significant attribution of the particular stripe are from its anchor. Although all 6 epigenomic features have peaks at the anchor locus, the model predicts the stripe mostly from 1) ATAC-seq and CTCF peaks at the anchor, and 2) H3K4me1 modification surrounding the anchor. **b**, The significant attribution of the particular loop are from its two anchors. Although H3K27ac have peaks at the left anchor locus, its contribution is negative towards predicting the loop. The CTCF binding at the anchors and H3K4me1/H3K4me3 modifications next to the anchors have positive attribution in predicting the loop.

3.6 Discussion and future directions

Our study connects nucleosome-resolution chromatin structures with epigenomic features. Leveraging the currently available Micro-C contact maps for hESC, mESC, and HFF from the 4DN consortium and the corresponding epigenomic profiles from ENCODE and Roadmap Epigenomics Project, we systematically mapped 1D epigenomic profiles to fine-scale 3D chromatin structures with CAESAR. The mapping was validated by high SCCs with observed Micro-C contact maps and the accurate capture of fine-scale loops and stripes. CAESAR can be applied to generate high-resolution contact maps for any cell line or tissue as long as their common epigenomic features are profiled. Our model further connects transcriptome with fine-scale structures and epigenomics by identifying the spatial interactions between genes and regulatory elements. Therefore, the imputed high-resolution contact maps will be useful for target finding, hypotheses generating, and other downstream analyses. All imputed human chromatin contact maps across 57 tissues, 16 cell lines, 12 primary cells, and 6 *in vitro* differentiated cells have been made publicly available on our web server (http://nucleome.dcmb.med.umich.edu/) for ease of access by biomedical researchers to perform further analyses.

While CAESAR presents a novel way to investigate fine details of 3D chromatin structure, we note that it is an evolving methodology with certain shortcomings that can be improved. First, since Micro-C data mostly outperforms Hi-C in the detection of short-range interactions, CAESAR also performs best at genomic distances of less than 200 kb. As a result of this, CAESAR-imputed contact maps are not well suited for analyses of large 3D chromatin structures such as compartments. Second, because Micro-C and Hi-C generate short-read sequences, our study is still limited to pairwise chromatin contacts, and therefore higher-order interactions are insufficiently studied. Third, our analyses showed that CAESAR performed well according to multiple evaluation metrics, yet there was a clear bias towards A compartment, evolutionarily conserved regions, and early-replicating regions. This is likely a reflection that the epigenomic features in the study are generally more enriched in these regions. As such, it is possible that including additional epigenomic features may shift this bias effect accordingly. Fourth, though CAESAR demonstrated clear relationships between epigenomic features and 3D fine-scale chromatin organization, we did not observe significant improvement in imputed contact maps with an increasing number of epigenomic datasets. This suggests that epigenomic data may not explain all the features observed in 3D chromatin organization. There may be unexplored layers of genetic and/or epigenetic information that play a role in the organization of chromatin inside the nucleus. So far, CAESAR demonstrated a framework for jointly analyzing 3D chromatin

structures and 1D epigenomic features at a matched resolution, and further integration of 1D DNA sequences is possible. For example, our model can potentially include DNA sequences as features and elucidate 3D QTLs [49] in the context of high-resolution chromatin organization.

CHAPTER 4

GenomicKB: A Knowledge Graph for the Human Genome

This chapter introduces the work of GenomicKB, a knowledge graph for the human genome [36]. Genomic Knowledgebase (GenomicKB) is a relational database for researchers to explore and investigate human genome, epigenome, transcriptome, and 4D nucleome with simple and efficient queries. The database uses a knowledge graph to consolidate genomic datasets and annotations from over 30 consortia and portals, and includes 347 million genomic entities, 1.36 billion relations, and 3.9 billion entity and relation properties. GenomicKB is equipped with a relational query system which allows users to query the knowledge graph with customized graph patterns and specific constraints on entities and relations. Compared with traditional tabular-structured data stored in separate data portals, GenomicKB emphasizes the relations among genomic entities, intuitively connects isolated data matrices, and supports efficient queries for scientific discoveries. GenomicKB transforms complicated analysis among multiple genomic entities and relations into coding-free and interactive queries, and facilitates data-driven genomic discoveries in the future.

4.1 Background: why do we need a knowledge graph?

Since the completion of the Human Genome Project [64], ever-evolving biotechnologies have enabled us to characterize the human genome from different perspectives. Consequently, many landmarking consortia have made tremendous progress towards understanding the functions of human genome in different aspects, such as the Encyclopedia of DNA Elements (ENCODE) [17], Roadmap Epigenomics [13], Genotype-Tissue Expression (GTEx) [51], and 4D Nucleome (4DN) [24], among others. Although these consortia provided different insights at an unprecedented scale and depth, the separately-stored tabular data is inconvenient for genomic research and scientific discoveries. First, merging multi-modal data often requires joining multiple tables, which takes tremendous storage space and efforts. Second, it is

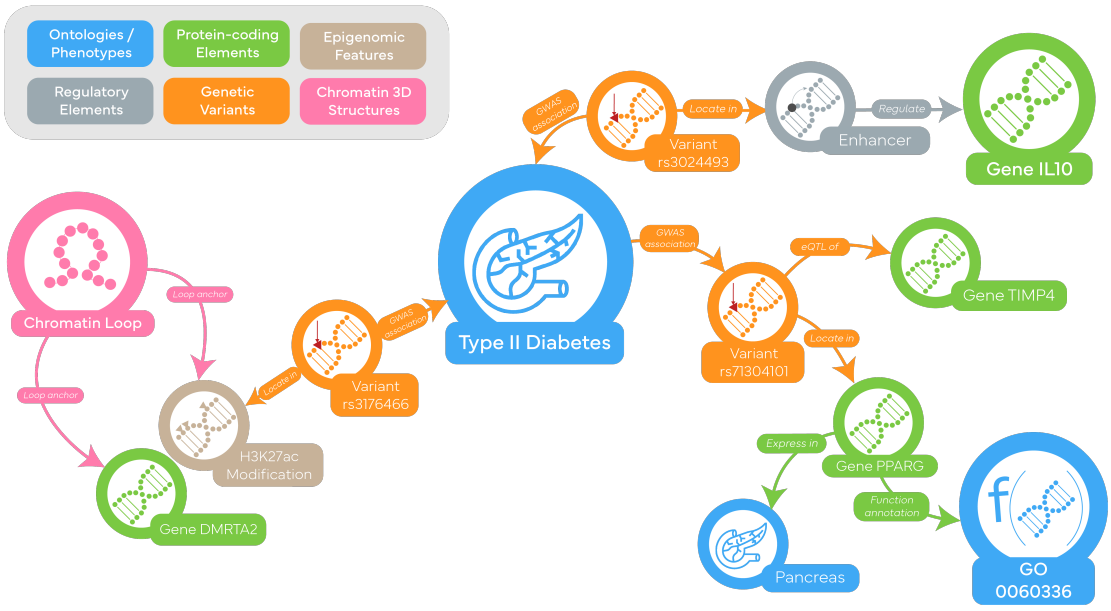

Figure 4.1: **An example subgraph of GenomicKB.**
In this subgraph, three GWAS variants of type II diabetes are connected with entities including genes, tissues, and 3D chromatin structures.

challenging to reconcile multiple data sources for the same topic (e.g., enhancers annotated by ENCODE CCRE [34] and ENdb [5]). In addition, extracting information from these isolated data requires coding skills, making open science and reproducible research difficult.

To solve this problem, we build Genomic Knowledgebase (GenomicKB), which seamlessly integrates datasets and annotations related to the human genome into a knowledge graph. Knowledge graphs intuitively represent connected data entities, and have been applied to biological domains [119, 144, 6, 56, 96, 8]. Compared with traditional tabular-structured data stored at separate portals, GenomicKB emphasizes the relations between genomic entities at multiple resolutions and from multiple tissues and cell types. Entities from each consortium automatically and explicitly cross-link with one another in the knowledge graph without any operations such as table joining and sorting. In addition, our GenomicKB is rigorously built with well-defined schema, identity, and ontology to maintain the data structure, disambiguate genomic concepts, and support future extension. As a result, GenomicKB is not only flexible to adapt updates of nodes, relations, and entire data sources, but can also connect with other knowledge graphs in related biomedical domains.

To support customized user queries, GenomicKB is equiped with a user-friendly web portal (`https://gkb.dcmb.med.umich.edu/`). To the best of our knowledge, this is the first graph pattern query system for the human genome, in which a query does not necessarily

start with a genomic region or a specific genomic entity. Instead, GenomicKB supports customized pattern queries such as "finding two genes which are both related to signal transduction, locate on the same chromosome, and form ligand-receptor pairs". As a result, GenomicKB transforms multi-modal data analysis into intuitive queries, and enables large-scale cross-modality pattern searching and learning in a highly-integrated knowledge graph. With this integrated data source and a robust data-sharing web portal, biomedical scientists can easily query, compare and investigate the high quality, high resolution, and comprehensive knowledge graph regarding chromatin organization, regulatory elements, epigenomic markers, and transcriptional regulation in various human tissues/cell lines at multiple resolutions.

4.2 Building GenomicKB

4.2.1 Collecting data for GenomicKB

Our knowledge graph integrates over 30 well-established data sources, including GENCODE [53], the Eukaryotic Promoter Database (EPD) [31], dbSuper [69], RNAcentral [15], Genotype-Tissue Expression (GTEx) [51], GWAS [89], Database of Genomic Variants (DGV) [90], NCBI dbVar [74], 4D Nucleome (4DN) [24], FIRE studies [120], ENCODE [17], MotifMap [20], NCBO ontologies [92], etc. (Tables 4.1 and 4.2). Each of these consortia incorporates thousands of datasets and provides different insights regarding human genome at an unprecedented scale and depth. To the best of our knowledge, the coverage of GenomicKB exceeds any knowledge graphs in related fields [119, 144, 6, 56, 96, 8]. One vital advantage of our knowledge graph structure is its flexibility which allows easy inclusion of new data in different formats. In addition, the query efficiency only drops insignificantly as we increase data entries.

Table 4.1: Number of entities included in the genomic graph and their data sources

Entity Type	Entity Sub-type	Data source	Number of Entities	
4*Coding elements	Genes	GENCODE	61186	
	Transcripts	GENCODE	236816	
	Exons	GENCODE	643060	
	Proteins	GENCODE	106140	
9*Non-coding elements	4*Enhancers	ENCODE Candidate cis-Regulatory Elements (CCRE)	809429	
		EnhancerAtlas	2895013	
		FANTOM5	32689	
		ENdb	249	
	Insulators	ENCODE Candidate cis-Regulatory Elements (CCRE)	56766	
	2*Promoters	The Eukaryotic Promoter Database (EPD)	21071	
		ENCODE Candidate cis-Regulatory Elements (CCRE)	34803	
	Super-enhancers	dbSuper	38030	
	non_coding_RNA	RNAcentral	474310	
5*Genomic variants	2*SNPs	Genotype-Tissue Expression (GTEx)	4295337	
		GWAS	167191	
	insertion/deletion	Genotype-Tissue Expression (GTEx)	337120	
	2*Structural variants	Database of Genomic Variants (DGV)	808608	
		NCBI dbVar	67718	
4*3D structures	Topological associating domains (TADs)	4D Nucleome (4DN)	44643	
	Chromatin loops	4D Nucleome (4DN)	37892	
	A/B compartments	4D Nucleome (4DN)	7879	
	Frequently interacting regions (FIREs)	FIRE studies	20960	
6*Epigenomic features	ChromHMM states	UCSC genome browser	4143552	
	Replication timing	4D Nucleome (4DN)	354962	
	Transcriptional factor binding profile	ENCODE	219830128	
	Transcriptional factor binding motifs	MotifMap	3996453	
	DNase-hypersensitivity sites	ENCODE	21858996	
	Histone binding profiles	ENCODE	44947427	
6*Ontologies	3*Tissue and cell lines	Cell ontology (CL)	2493*	3*27783*
		Uber-anatomy ontology (UBERON)	15398*	
		BRENDA tissue ontology (BTO)	6520*	
	Experimental factors	Experimental factor ontology (EFO)	11299*	28472*
	Genes	Gene ontology (GO)	50635	
	Transcriptional factors	Human transcriptional factors	2765	

*Ontologies including CL, UBERON, BTO, and EFO have some shared terms. Numbers on the left are the count of terms that starts with the corresponding ontology name (e.g., BTO:0006563), and those on the right indicate terms that are categorized as "tissue and cell lines" or "experimental factors".

4.2.2 Schema, identity, and ontology in GenomicKB

Schema: Schemata prescribe high-level structures and semantics that the knowledge graph follows, which reduces data errors and allows reasoning over the data graph [58]. In GenomicKB, we formally define node schema and edge schema as follows. Nodes are labeled with hierarchical classes. The top level includes six classes, namely chromosome chain, coding element, non-coding element, epigenomic feature, variant, and ontology. Each class also consists of sub-classes (Table 4.1). Edge schema defines the rules of node connections. Edges are categorized into position, regulation, expression, and annotation, and each sub-type has corresponding start and end node types (Table 4.2). For example, an "express in" edge must start from a gene and point to a tissue or cell line, and a "correlate with" edge only corresponds to the correlation between

Table 4.2: Number of relationships included in the genomic graph and their data sources

Relationship type	Relationship subtype		From	To	Data source	Number of relationships
2*Positional	Connect (next_loc/lower_resolution)		Genomic sequence	Genomic sequence	NCBI	19517799
	Locate at (locate_on_chain)		All entities that have a location property	Genomic sequence	All data sources	1057172343
5*Expression	Express into (express_into)		Genes	Transcriptional factors	humanTF	2765
	Express in (express_in)		Genes	Ontologies	Genotype-Tissue Expression (GTEx)	3032424
	Transcribe (transcribe_into)		Genes	Transcripts	Ensembl	236816
	Translate (translate_into)		Transcripts	Proteins	Ensembl	106140
	Include		Transcripts	Exons	Ensembl	1274728
6*Regulatory	2*Regulate		Genes	Genes	RegNetwork	129129
			Enhancers	Genes	EnhancerAtlas/ENdb	9112174
	Expression QTLs (correlate_with)		Variants	Genes/Ontologies	Genotype-Tissue Expression (GTEx), dbVar, GWAS	14058410
	3*SNP and gene	SNP_in_gene	3*Sequence variants	3*Genes	3*GWAS	104178
		SNP_upstream_gene				432326
		SNP_downstream_gene				457807
6*Annotation	2*Belong to		Gene	Ontology	Ensembl	2*343858
			Non coding RNA	Ontology	RNAcentral	
	Gene sub-type		Gene ontology	Gene ontology	Gene ontology (GO)	4*151939
	3*Tissue/cell sub-type		3*Cell/Tissue ontology	3*Cell/Tissue ontology	Cell ontology (CL)	
					Uber-anatomy ontology (UBERON)	
					BRENDA tissue ontology (BTO)	

variants and gene expression or phenotype. Node schema and edge schema are exactly followed during data importing to ensure GenomicKB's structure, semantics, and data types.

Identity: Identity consolidates a set of unique identifiers and disambiguates different genomic identities in the knowledge graph. Since different data sources may follow different conventions to represent the same concept (e.g., ENSG00000223972 and gene *DDX11L1*), or use the same name to describe different concepts (e.g., gene *p53* and protein p53), we use *globally-unique identifiers* and *external identity links* in GenomicKB. For example, for genes, transcripts, and exons, we refer to Ensembl [60] IDs for their external identity links. For epigenomic entities without external identity links such as ChIP-seq peaks, we define their globally-unique identifiers according to their genomic coordinates, cell lines, and histone/TF types.

Ontology: Ontology is a uniform language to describe scientific terms. Concepts such as cell lines and tissues are represented as ontology URLs and IDs instead of common names to ensure disambiguity and future integration with other knowledge graphs. GenomicKB includes well-established ontologies related to genes (GO [18] and HGNC [108]), tissues and cell lines (UBERON [97], BTO [50], CL [26], and EFO [91]). These ontologies serve two roles in GenomicKB. First, some entities directly connect to ontologies and are accessible in queries. For example, users can query all genes linked to the same specified GO term.

Second, scientific terms such as diseases and cell line names are encoded in ontology IDs. Therefore, different conventions of the same concept, such as "IMR-90", "IMR90", and "cells - cultured fibroblasts" are unified in GenomicKB.

4.3 GenomicKB supports graph-based relational queries

We design a web interface (http://gkb.dcmb.med.umich.edu/) that supports customized query of diverse entities, relations and properties. The query system consists of three components - a canvas, an editor panel, and a console. On the canvas, users can draw customized graph patterns by inserting nodes and edges. When adding a node/edge or a node/edge is selected, the corresponding editor panel on the top left activates to enable node/edge configuration, such as edit the type of the node/edge or add property constraints. During the process, the bottom left console shows real-time hints to guide users to create valid queries. After the user specifies the query conditions, the user needs to click the "Submit" button on the bottom to submit the query, which re-directs to a result page.

The result page includes two panels. The left panel displays the result sub-graph with moving and zooming functions. If positional relationships (such as *overlap* and *downstream*) are included in the query, genomic regions that entities locate in are also visualized as connected bins, whereas other entities related to this region are displayed around it. The right panel displays detailed properties when a node is selected. If the retrieved sub-graph is overly large, then only partial results (e.g., five to twenty matched patterns) are visualized, and the complete query result can be downloaded by clicking "export all". The downloaded result is in json or excel format. A video tutorial is also attached on our front page.

4.4 Applications of GenomicKB

4.4.1 GenomicKB simplifies cross-modality analysis as queries over the knowledge graph

GenomicKB integrates complementary data sources into a knowledge graph and simplifies multi-modal analysis as queries over the knowledge graph. For example, to identify genes and genetic variants related to type II diabetes (T2D), traditional approaches require integrating multiple data sources as follows. First, all variants correlated with T2D are retrieved from portals such as GWAS Catalog [89]. Then, variants are linked to genes by identifying intra-

gene variants with gene coordinates from GENCODE [53]. Additional restrictions about the gene may be applied as well, such as the minimum gene expression level in pancreas (from consortia such as ENCODE [17] and GTEx [51]). Lastly, function annotations of the genes are identified from Gene Ontology [18]. With GenomicKB, the aforementioned analysis can be easily completed with a sub-graph query over the knowledge graph (Figure 4.2). All restrictions and sub-graphs can be specified via the user-friendly interface, and we no longer require complex queries in individual data sources or any coding skills. At the backend, the submitted query pattern is automatically translated into a Cypher query [39], and the query results are returned and visualized as graphs (Figure 4.2). With consolidated data and an intuitive query process, GenomicKB makes it easier for researchers to discover new genomic insights.

4.4.2 GenomicKB encodes positional relations among different genomic entities

Most genomic entities locate on specific regions on the chromosome with positional relations between each other. GenomicKB supports queries based on positional relations including *locate_in* (one entity is completely included by another), *overlap* (two entities have a coordinate overlap), *upstream/downstream* (one entity does not overlap and is upstream/downstream of another on the same chromosome), and *same_chr* (two entities are on the same chromosome). For example, to investigate transcription factor (TF) binding at chromatin loop anchors called at 5 Kb resolution, the traditional approach is to call loops from chromatin contact maps available at 4DN data portal and collect TF binding profile from epigenome consortia such as ENCODE and Roadmap Epigenomics, and then identify their overlap with computational tools. In GenomicKB, a query "*TF_binding_site* overlap *loop*" provides the same result (Figure 4.3). When restricting the query to GM12878 cell line and TF name to CTCF, 4,724 distinct loops are returned. As a comparison, 5,758 are returned from the query of all GM12878 loops without specifying the overlap with TF binding sites. Therefore, 82% of loops in GM12878 have at least one anchor bound by CTCF. A similar query of loops overlapping two different CTCF binding sites results in 2,680 returned entries, indicating that 47% of the 5,758 loops are between two CTCF binding sites.

To represent positional relations in GenomicKB, we first split all chromosomes into regions of a particular size (i.e., resolution), represent each region as a node, and connect them with edges. The series of nodes and edges are referred to as "chromosome chains", which are constructed in 200, 1000, 5000, 10000, and 50000 base-pair resolutions. Afterwards, entities that locate on specific regions are connected to the corresponding chromosome chain nodes.

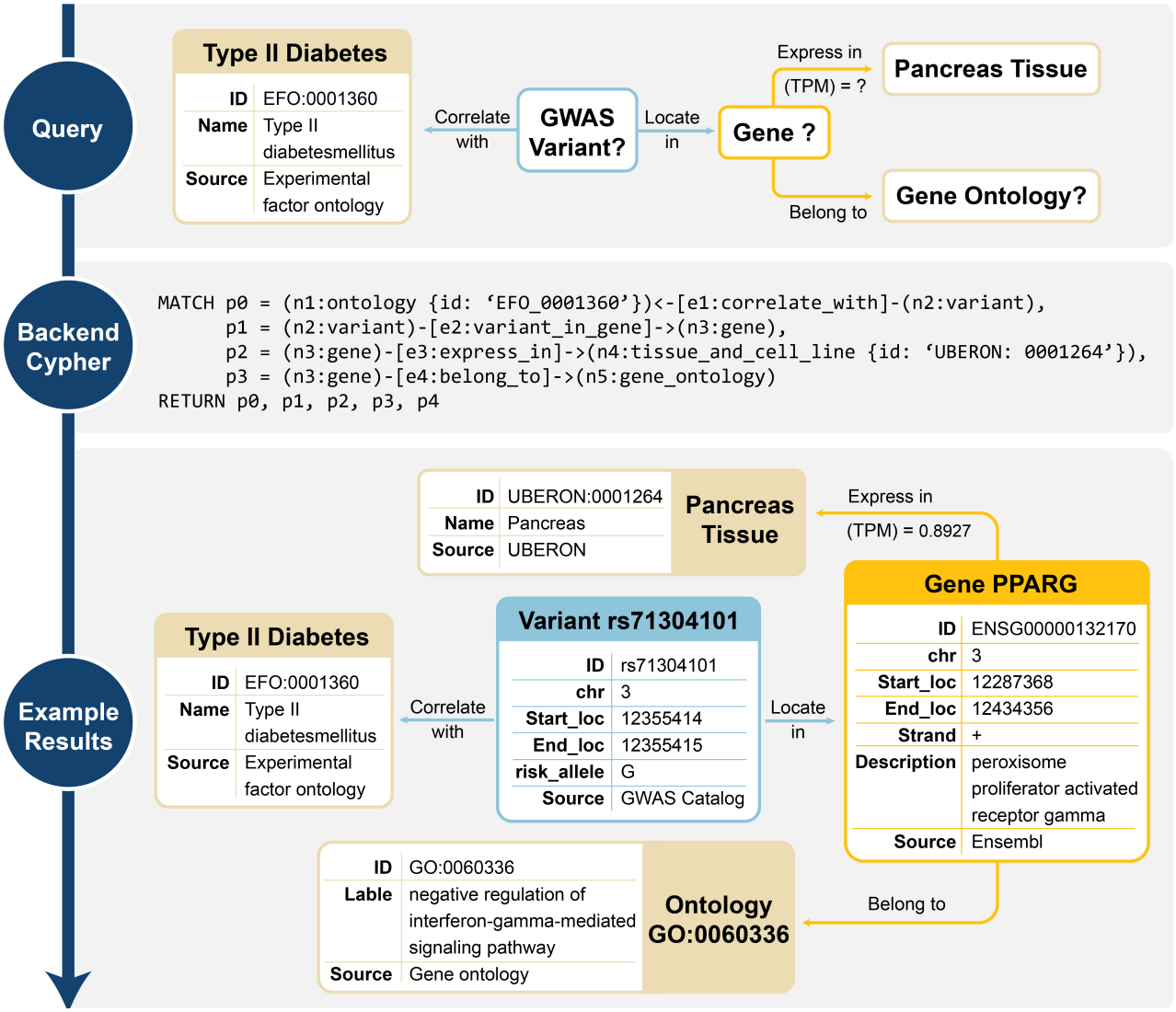

Figure 4.2: **GenomicKB simplifies cross-modality analysis as queries over the knowledge graph.**
If a user is interested in relations between T2D and genes, then instead of searching multiple databases including GWAS, ENCODE, and GO, a sub-graph query over GenomicKB returns all variants, genes, and gene ontologies that satisfy the query criteria.

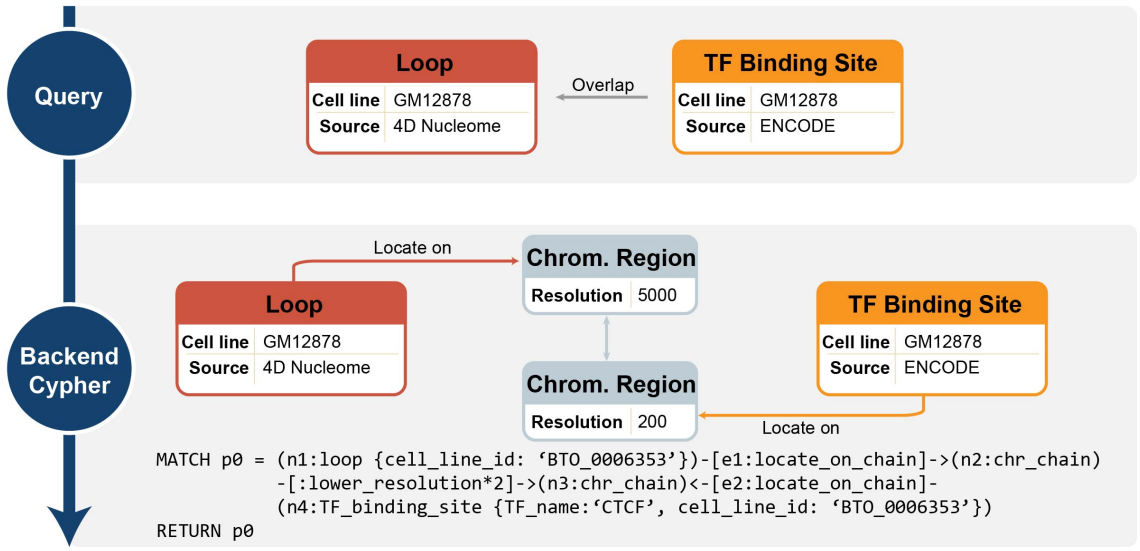

Figure 4.3: **GenomicKB supports queries related to positional relations between genomic entities.**
An example query of CTCF binding to loop anchors is illustrated.

The chromosome chains are intermediate nodes for capturing any positional relations between genomic entities (Figure 4.3).

4.4.3 GenomicKB reconciles consensus or conflicting data sources of the same problem

For some genomic entity, multiple data sources may provide either consensus or conflicting evidence. Knowledge graphs are able to reconcile mof facts in the light of well-defined schema, identities, and ontologies. We use the example of enhancers to show that GenomicKB reconciles multiple data sources for the same problem. As key regulatory elements, enhancers are annotated by several data sources, such as ENdb [5], EnhancerAtlas [46], ENCODE CCRE [17], and FANTOM5 [2]. To identify enhancers from one database in GenomicKB, users can query the node "enhancer" with restrictions such as "data_source = FANTOM5". By defining enhancers from different data sources with coordinate overlaps as consensus ones, one can also query how many enhancers from two sources (e.g., CCRE and EnhancerAtlas) agree with each other (Query 1 in Figure 4.4). In addition, relations from one data source can be cross-validated by other data sources. For example, EnhancerAtlas provides enhancer-gene interactions, which can be validated by other approaches that map enhancers to genes such

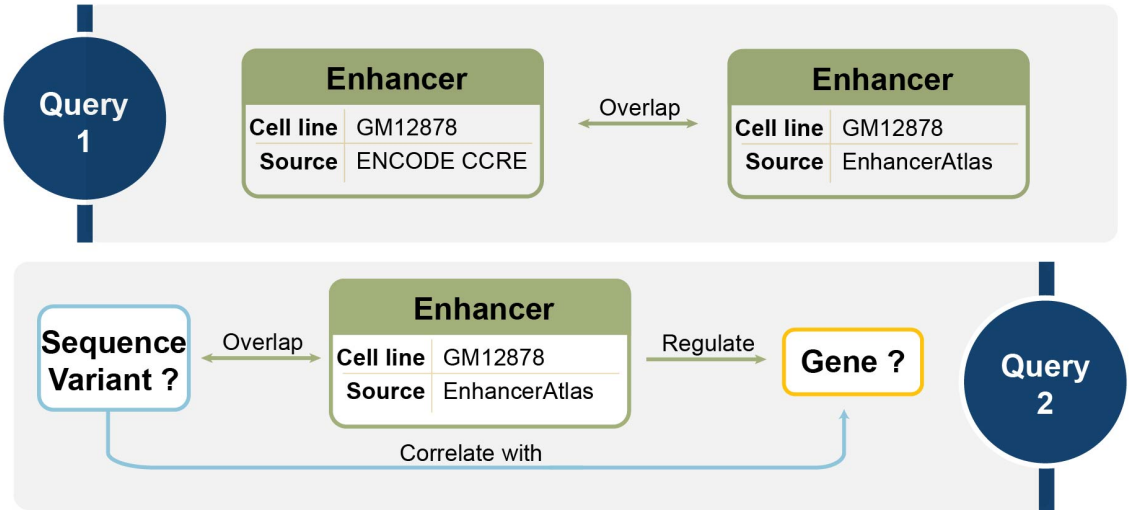

Figure 4.4: **GenomicKB reconciles multiple data sources for the same problem, such as identifying enhancers and mapping enhancers to genes.**
Query 1 demonstrates how GenomicKB evaluates the consensus enhancers between CCRE and EnhancerAtlas. Query 2 illustrates how enhancer-gene mapping from EnhancerAtlas is validated by eQTL-gene pairs in GenomicKB.

as eQTL-gene correlation as follows. First, a query "*enhancer* regulate *gene*" with restriction "cell_line=GM12878" and "data_source=EnhancerAtlas" returns 118,610 enhancer-gene pairs from EnhancerAtlas. Then, we can identify the eQTLs of the gene locating in the enhancer, which can be represented as "*variant* overlap *enhancer*", *enhancer* regulate *gene*", and "*variant* correlate_with *gene*" (Query 2 in Figure 4.4). The number of distinct enhancer-gene pairs decreases to 16,871 in the result, indicating that 16,871 enhancer-gene pairs from EnhancerAtlas can be validated by GTEx eQTLs.

4.5 Discussion and future directions

In conclusion, GenomicKB integrates our existing knowledge regarding human genome, epigenome, transcriptome, and 4D nucleome in a large knowledge graph. Different from traditional tabular-structured data, it emphasizes the relations between different perspectives and provides explicit connections between entities of interest. With the flexibility, well-defined schemata and ontologies used in the knowledge graph, it is quite easy to update the existing entities and relations and incrementally add more entities and relations. Since GenomicKB adapts external unique identifiers for nodes and edges, it is convenient to

connect it with other biomedical knowledge graphs. To increase accessibility, GenomicKB is equipped with a web portal (http://gkb.dcmb.med.umich.edu/) for users to specify and submit intuitive graph-based queries. With this portal, GenomicKB is capable of answering human genomics-related questions and conducting multi-modal analysis with coding-free and interactive queries. Therefore, we expect that GenomicKB can attract researchers with diverse backgrounds and enhance open science in genomic research.

In recent years, artificial intelligence plays increasingly important roles in problems related to transcription regulation [150, 68, 67, 4], chromatin 3D structures [42, 11, 147, 78], and single-cell genomics [40, 87]. Nevertheless, we are still looking for a "universal model" that captures large-scale genomic data from different perspectives and comprehensively decodes the human genome. Similar to the field of natural language processing in which new language models and question-answering systems are based on large knowledge graphs [16, 101] (e.g., the Wiki knowledge graph), we expect that genomic research becomes increasingly data-driven, and GenomicKB provides high-quality and integrated data for large-scale machine learning methods and facilitates scientific discoveries.

CHAPTER 5

Conclusion

5.1 Dissertation summary: graph-based representations of genomic knowledge

This dissertation introduces the intricate landscape of transcriptional regulation in the human genome and proposes that the human genome can be deciphered with graphs.

The introductory chapter elucidates the complexity of transcriptional regulation, emphasizing the collaboration of enhancers, promoters, transcriptional factors, and chromatin structures. It emphasizes the need to move beyond a simplistic 1-D sequence understanding and explores the intricate network of the human genome from diverse data sources.

Chromatin 3D structure is the first graph view for the human genome (Chapters 2 and 3), in which genomic regions are nodes, chromatin contacts between regions are relationships, and genomic/epigenomic features are node properties. Chapter 2 introduces Chromatin Conformation Capture (3C) technology, which uncovers hierarchical structural features through chromatin contact maps. It focuses on the relationships between transcriptional regulation and 3D chromatin organization, showcasing the role of CTCF-mediated chromatin loops in regulating fetal hemoglobin expression. Additionally, we present 3D genome-related computational tools, including scHiCTools for analyzing single-cell Hi-C data and Quagga for identifying stripes from chromatin contact maps. Chapter 3 introduces the work connecting high-resolution 3D chromatin organization with epigenomics using a deep-learning model This model, CAESAR, imputes chromatin contact maps for more than 90 human samples. The chapter evaluates the algorithm's accuracy, factors influencing its performance, and its ability to recapitulate CRISPRi-validated enhancer activities and recover eQTL-gene interactions. It also explores the identification of epigenomic features relevant to fine-scale 3D chromatin organization.

Chapter 4 introduces another graph view of the human genome - knowledge graphs. Published data from landmarking consortia including ENCODE, Roadmap, GTEx, 4DN, and

HubMAP have provided multifaceted insights into the human genome's functions, connections, and relationships. The chapter highlights the importance of considering the human genome as a complex network, proposing a shift from the aforementioned isolated datasets to a graph-based representation - GenomicKB. It details the process of building GenomicKB, including data collection, schema, identity, and ontology. The chapter highlights how GenomicKB supports graph-based relational queries and discusses its applications in simplifying cross-modality analysis, encoding positional relations among genomic entities, and reconciling consensus or conflicting data sources.

The dissertation aims to contribute a comprehensive understanding of transcriptional regulation in the human genome through a graph-based representation. This novel approach has the potential to unlock new insights into the relationships between genomic, epigenomic, and transcriptomic entities, fostering a deeper comprehension of the regulatory mechanisms governing gene expression in human biology.

5.2 Perspective: data mining, data integration, and hypothesis-generating for human genomics

Following this dissertation, we can propose potential next steps in the research community.

5.2.1 Data mining and data integration: accumulating the knowledge of human genome at a larger scale

Knowledge and expertise are limited for a single biologist or research team. However, in the past half a century, millions of research papers, posters, and abstracts have been published in all domains of the human genome. By accumulating knowledge from the entire community, we expect more scientific discoveries can be made.

GenomicKB is an early effort to uniformly integrate data and provide user-friendly access functions. In the future, we plan to extend the work in the following directions.

5.2.1.1 Using artificial intelligence to extract a wider range of knowledge

Although GenomicKB integrates data from more than 30 databases and data consortia, it only covers part of the discoveries from the research community. Another direct way of extracting knowledge is text mining from research papers. It is possible to apply the state-of-the-art named-entity recognition (NER) and relation extraction (RE) algorithms, and leverage descriptions of biological terms and their biologically interactive information to

achieve entity and relation extraction [12, 76]. The extracted knowledge can be consolidated with GenomicKB.

5.2.1.2 Integrate GenomicKB with other knowledge graphs

GenomicKB is the pioneer knowledge graph for the human genome, but not the only knowledge graph in the biomedical domain. Previous knowledge graphs focused more on proteins, chemicals, and drugs [119, 144, 6, 56, 96, 8]. It is possible to integrate them together as long as the same identity and ontology system is applied.

Our ultimate goal is to build a reliable data source for the biomedical domain. This data source not only helps researchers perform cross-modality queries but also provides efficient representations for computers and artificial intelligence models. Similar to the field of natural language processing in which new language models and question-answering systems are based on large knowledge graphs [16, 101] (e.g., the Wiki knowledge graph), we expect that genomic research becomes increasingly data-driven, and GenomicKB provides high-quality and integrated data for large-scale machine learning methods and facilitates scientific discoveries.

5.2.2 Integrated analysis and hypothesis generation for the human genome

Once our computational framework has integrated enough data, we expect to develop computational pipelines to let them work like biologists. Two example applications are introduced in this chapter.

5.2.2.1 Knowledge-driven integrated analysis

Modern artificial intelligence models, such as ChatGPT, have made it possible to generate sentences and figures according to user-specified prompts. This is achieved by both well-designed deep learning architecture and large-scale training corpus. Similarly, with large-scale machine-readable knowledge integrated into our knowledge graphs, we also propose to develop a knowledge-based or prompt-based analysis platform. For example, the user could provide a list of genes that are upregulated in their experiments, and ask "please summarize the common features of this group of genes". We expect that with the integrated knowledge, our model could generate meaningful responses such as "these genes are enriched in chromosome 6", "a large proportion of these genes are related to inflammation response", or even "a previous paper generated a similar gene list as you provided".

5.2.2.2 Hypothesis generation for human genomics

Beyond conventional analysis, our computational framework also has the potential to generate hypotheses in human genomics. Computational models have been published to learn rules from knowledge graphs and generate novel hypotheses of undiscovered relationships [48, 1, 127]. Similarly, our system will leverage its comprehensive understanding of integrated datasets to propose novel hypotheses. For example, if a researcher is interested in a disease, our computational approaches could predict which genes, variants, and biological pathways should be paid attention to when investigating the disease. Since large models accumulate the knowledge and expertise of millions of researchers from the research community, we expect the accuracy of the hypotheses to be significantly higher than the experience-based decision from a single biologist.

BIBLIOGRAPHY

[1] Uchenna Akujuobi, Michael Spranger, Sucheendra K Palaniappan, and Xiangliang Zhang. T-pair: Temporal node-pair embedding for automatic biomedical hypothesis generation. *IEEE Transactions on Knowledge and Data Engineering*, 34(6):2988–3001, 2020.

[2] Robin Andersson, Claudia Gebhard, Irene Miguel-Escalada, Ilka Hoof, Jette Bornholdt, Mette Boyd, Yun Chen, Xiaobei Zhao, Christian Schmidl, Takahiro Suzuki, et al. An atlas of active enhancers across human cell types and tissues. *Nature*, 507(7493):455–461, 2014. [PubMed:24670763] [PubMed Central:PMC5215096] [doi:10.1038/nature12787].

[3] Abbas Roayaei Ardakany, Ferhat Ay, and Stefano Lonardi. Selfish: Discovery of differential chromatin interactions via a self-similarity measure. *bioRxiv*, 2019.

[4] Žiga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R Ledsam, Agnieszka Grabska-Barwinska, Kyle R Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, and David R Kelley. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature methods*, 18(10):1196–1203, 2021. [PubMed:34608324] [PubMed Central:PMC8490152] [doi:10.1038/s41592-021-01252-x].

[5] Xuefeng Bai, Shanshan Shi, Bo Ai, Yong Jiang, Yuejuan Liu, Xiaole Han, Mingcong Xu, Qi Pan, Fan Wang, Qiuyu Wang, et al. Endb: a manually curated database of experimentally supported enhancers for human and mouse. *Nucleic Acids Research*, 48(D1):D51–D57, 2020. [PubMed:31665430] [PubMed Central:PMC7145688] [doi:10.1093/nar/gkz973].

[6] Irina Balaur, Alexander Mazein, Mansoor Saqi, Artem Lysenko, Christopher J Rawlings, and Charles Auffray. Recon2neo4j: applying graph database technologies for managing comprehensive genome-scale networks. *Bioinformatics*, 33(7):1096–1098, 2017. [PubMed:27993779] [PubMed Central:PMC5408918] [doi:10.1093/bioinformatics/btw731].

[7] Arthur Bank. Regulation of human fetal hemoglobin: new players, new complexities. *Blood*, 107(2):435–443, 2006.

[8] Albert-László Barabási, Natali Gulbahce, and Joseph Loscalzo. Network medicine: a network-based approach to human disease. *Nature reviews genetics*, 12(1):56–68, 2011. [PubMed:21164525] [PubMed Central:PMC3140052] [doi:10.1038/nrg2918].

[9] Daniel E Bauer, Sophia C Kamran, Samuel Lessard, Jian Xu, Yuko Fujiwara, Carrie Lin, Zhen Shao, Matthew C Canver, Elenoe C Smith, Luca Pinello, et al. An erythroid enhancer of bcl11a subject to genetic variation determines fetal hemoglobin level. *Science*, 342(6155):253–257, 2013.

[10] Daniel E Bauer, Sophia C Kamran, and Stuart H Orkin. Reawakening fetal hemoglobin: prospects for new therapies for the β-globin disorders. *Blood, The Journal of the American Society of Hematology*, 120(15):2945–2953, 2012.

[11] Polina S Belokopytova, Miroslav A Nuriddinov, Evgeniy A Mozheiko, Daniil Fishman, and Veniamin Fishman. Quantitative prediction of enhancer–promoter interactions. *Genome research*, 30(1):72–84, 2020. [PubMed:31804952] [PubMed Central:PMC6961579] [doi:10.1101/gr.249367.119].

[12] Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*, 2019.

[13] Bradley E Bernstein, John A Stamatoyannopoulos, Joseph F Costello, Bing Ren, Aleksandar Milosavljevic, Alexander Meissner, Manolis Kellis, Marco A Marra, Arthur L Beaudet, Joseph R Ecker, et al. The NIH roadmap epigenomics mapping consortium. *Nature biotechnology*, 28(10):1045, 2010.

[14] Laura Breda, Irene Motta, Silvia Lourenco, Chiara Gemmo, Wulan Deng, Jeremy W Rupon, Osheiza Y Abdulmalik, Deepa Manwani, Gerd A Blobel, and Stefano Rivella. Forced chromatin looping raises fetal hemoglobin in adult sickle cells to higher levels than pharmacologic inducers. *Blood, The Journal of the American Society of Hematology*, 128(8):1139–1143, 2016.

[15] RNA central. Rnacentral 2021: secondary structure integration, improved sequence search and new member databases. *Nucleic acids research*, 49(D1):D212–D220, 2021. [PubMed:33106848] [PubMed Central:PMC7779037] [doi:10.1093/nar/gkaa921].

[16] Xiaojun Chen, Shengbin Jia, and Yang Xiang. A review: Knowledge reasoning over knowledge graph. *Expert Systems with Applications*, 141:112948, 2020. [doi:10.1016/j.eswa.2019.112948].

[17] ENCODE Project Consortium et al. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57, 2012. [PubMed:22955616] [PubMed Central:PMC3439153] [doi:10.1038/nature11247].

[18] Gene Ontology Consortium. The gene ontology (go) database and informatics resource. *Nucleic acids research*, 32(suppl_1):D258–D261, 2004. [PubMed:33137190] [PubMed Central:PMC7778975] [doi:10.1093/nar/gkh036].

[19] HuBMAP Consortium et al. The human body at cellular resolution: the nih human biomolecular atlas program. *Nature*, 574(7777):187, 2019.

[20] Kenneth Daily, Vishal R Patel, Paul Rigor, Xiaohui Xie, and Pierre Baldi. Motifmap: integrative genome-wide maps of regulatory motif sites for model species. *BMC bioinformatics*, 12(1):1–13, 2011.

[21] Elzo de Wit and Wouter De Laat. A decade of 3C technologies: insights into nuclear organization. *Genes & development*, 26(1):11–24, 2012.

[22] J. Dekker, A. S. Belmont, M. Guttman, V. O. Leshyk, J. T. Lis, S. Lomvardas, L. A. Mirny, C. C. O'Shea, P. J. Park, B. Ren, J. C. Ritland Politz, J. Shendure, S. Zhong, and the 4D Nucleome Network. The 4D nucleome project. *Nature*, 549:219–226, 2017.

[23] Job Dekker, Frank Alber, Sarah Aufmkolk, Brian J Beliveau, Benoit G Bruneau, Andrew S Belmont, Lacramioara Bintu, Alistair Boettiger, Riccardo Calandrelli, Christine M Disteche, et al. Spatial and temporal organization of the genome: Current state and future aims of the 4d nucleome project. *Molecular Cell*, 2023.

[24] Job Dekker, Andrew S Belmont, Mitchell Guttman, Victor O Leshyk, John T Lis, Stavros Lomvardas, Leonid A Mirny, Clodagh C O'shea, Peter J Park, Bing Ren, et al. The 4D nucleome project. *Nature*, 549(7671):219, 2017. [PubMed:28905911] [PubMed Central:PMC5617335] [doi:10.1038/nature23884].

[25] Wulan Deng, Jongjoo Lee, Hongxin Wang, Jeff Miller, Andreas Reik, Philip D Gregory, Ann Dean, and Gerd A Blobel. Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. *Cell*, 149(6):1233–1244, 2012.

[26] Alexander D Diehl, Terrence F Meehan, Yvonne M Bradford, Matthew H Brush, Wasila M Dahdul, David S Dougall, Yongqun He, David Osumi-Sutherland, Alan Ruttenberg, Sirarat Sarntivijai, et al. The cell ontology 2016: enhanced content, modularization, and ontology interoperability. *Journal of biomedical semantics*, 7(1):1–10, 2016. [PubMed:27377652] [PubMed Central:PMC4932724] [doi:10.1186/s13326-016-0088-7].

[27] J. R. Dixon, S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. S. Liu, and B. Ren. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376–380, 2012.

[28] Jesse R Dixon, Inkyung Jung, Siddarth Selvaraj, Yin Shen, Jessica E Antosiewicz-Bourget, Ah Young Lee, Zhen Ye, Audrey Kim, Nisha Rajagopal, Wei Xie, et al. Chromatin architecture reorganization during stem cell differentiation. *Nature*, 518(7539):331, 2015.

[29] Jesse R Dixon, Siddarth Selvaraj, Feng Yue, Audrey Kim, Yan Li, Yin Shen, Ming Hu, Jun S Liu, and Bing Ren. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376, 2012.

[30] Jill M Dowen, Zi Peng Fan, Denes Hnisz, Gang Ren, Brian J Abraham, Lyndon N Zhang, Abraham S Weintraub, Jurian Schuijers, Tong Ihn Lee, Keji Zhao, et al. Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes. *Cell*, 159(2):374–387, 2014.

[31] René Dreos, Giovanna Ambrosini, Rouayda Cavin Périer, and Philipp Bucher. Epd and epdnew, high-quality promoter resources in the next-generation sequencing era. *Nucleic acids research*, 41(D1):D157–D164, 2013. [PubMed:23193273] [PubMed Central:PMC3531148] [doi:10.1093/nar/gks1233].

[32] Ian Dunham, Anshul Kundaje, Shelley F. Aldred, Patrick J. Collins, Carrie A. Davis, Francis Doyle, Charles B. Epstein, Seth Frietze, Jennifer Harrow, Rajinder Kaul, Jainab Khatun, Bryan R. Lajoie, Stephen G. Landt, Bum-Kyu Lee, Florencia Pauli, Kate R. Rosenbloom, Peter Sabo, Alexias Safi, Amartya Sanyal, Noam Shoresh, Jeremy M. Simon, Lingyun Song, Nathan D. Trinklein, Robert C. Altshuler, Ewan Birney, James B. Brown, Chao Cheng, Sarah Djebali, Xianjun Dong, Ian Dunham, Jason Ernst, Terrence S. Furey, Mark Gerstein, Belinda Giardine, Melissa Greven, Ross C. Hardison, Robert S. Harris, Javier Herrero, Michael M. Hoffman, Sowmya Iyer, Manolis Kellis, Jainab Khatun, Pouya Kheradpour, Anshul Kundaje, Timo Lassmann, Qunhua Li, Xinying Lin, Georgi K. Marinov, Angelika Merkel, Ali Mortazavi, Stephen C. J. Parker, Timothy E. Reddy, Joel Rozowsky, Felix Schlesinger, Robert E. Thurman, Jie Wang, Lucas D. Ward, Troy W. Whitfield, Steven P. Wilder, Weisheng Wu, Hualin S. Xi, Kevin Y. Yip, Jiali Zhuang, Bradley E. Bernstein, Ewan Birney, Ian Dunham, Eric D. Green, Chris Gunter, Michael Snyder, Michael J. Pazin, Rebecca F. Lowdon, Laura A. L. Dillon, Leslie B. Adams, Caroline J. Kelly, Julia Zhang, Judith R. Wexler, Eric D. Green, Peter J. Good, Elise A. Feingold, Bradley E. Bernstein, Ewan Birney, Gregory E. Crawford, Job Dekker, Laura Elnitski, Peggy J. Farnham, Mark Gerstein, Morgan C. Giddings, Thomas R. Gingeras, Eric D. Green, Roderic Guigó, Ross C. Hardison, Timothy J. Hubbard, Manolis Kellis, W. James Kent, Jason D. Lieb, Elliott H. Margulies, Richard M. Myers, Michael Snyder, John A. Stamatoyannopoulos, Scott A. Tenenbaum, Zhiping Weng, Kevin P. White, Barbara Wold, Jainab Khatun, Yanbao Yu, John Wrobel, Brian A. Risk, Harsha P. Gunawardena, Heather C. Kuiper, Christopher W. Maier, Ling Xie, Xian Chen, Morgan C. Giddings, Bradley E. Bernstein, Charles B. Epstein, Noam Shoresh, Jason Ernst, Pouya Kheradpour, Tarjei S. Mikkelsen, Shawn Gillespie, Alon Goren, Oren Ram, Xiaolan Zhang, Li Wang, Robbyn Issner, Michael J. Coyne, Timothy Durham, Manching Ku, Thanh Truong, Lucas D. Ward, Robert C. Altshuler, Matthew L. Eaton, Manolis Kellis, Sarah Djebali, Carrie A. Davis, Angelika Merkel, Alex Dobin, Timo Lassmann, Ali Mortazavi, Andrea Tanzer, Julien Lagarde, Wei Lin, Felix Schlesinger, Chenghai Xue, Georgi K. Marinov, Jainab Khatun, Brian A. Williams, Chris Zaleski, Joel Rozowsky, Maik Röder, Felix Kokocinski, Rehab F. Abdelhamid, Tyler Alioto, Igor Antoshechkin, Michael T. Baer, Philippe Batut, Ian Bell, Kimberly Bell, Sudipto Chakrabortty, Xian Chen, Jacqueline Chrast, Joao Curado, Thomas Derrien, Jorg Drenkow, Erica Dumais, Jackie Dumais, Radha Duttagupta, Megan Fastuca, Kata Fejes-Toth, Pedro Ferreira, Sylvain Foissac, Melissa J. Fullwood, Hui Gao, David Gonzalez, Assaf Gordon, Harsha P. Gunawardena, Cédric Howald, Sonali Jha, Rory Johnson, Philipp Kapranov, Brandon King, Colin Kingswood, Guoliang Li, Oscar J. Luo, Eddie Park, Jonathan B. Preall, Kimberly Presaud, Paolo Ribeca, Brian A. Risk, Daniel Robyr, Xiaoan Ruan, Michael Sammeth, Kuljeet Singh Sandhu, Lorain Schaeffer, Lei-Hoon See, Atif Shahab, Jorgen Skancke, Ana Maria Suzuki, Hazuki Takahashi, Hagen Tilgner, Diane Trout, Nathalie Walters,

Huaien Wang, John Wrobel, Yanbao Yu, Yoshihide Hayashizaki, Jennifer Harrow, Mark Gerstein, Timothy J. Hubbard, Alexandre Reymond, Stylianos E. Antonarakis, Gregory J. Hannon, Morgan C. Giddings, Yijun Ruan, Barbara Wold, Piero Carninci, Roderic Guigó, Thomas R. Gingeras, Kate R. Rosenbloom, Cricket A. Sloan, Katrina Learned, Venkat S. Malladi, Matthew C. Wong, Galt P. Barber, Melissa S. Cline, Timothy R. Dreszer, Steven G. Heitner, Donna Karolchik, W. James Kent, Vanessa M. Kirkup, Laurence R. Meyer, Jeffrey C. Long, Morgan Maddren, Brian J. Raney, Terrence S. Furey, Lingyun Song, Linda L. Grasfeder, Paul G. Giresi, Bum-Kyu Lee, Anna Battenhouse, Nathan C. Sheffield, Jeremy M. Simon, Kimberly A. Showers, Alexias Safi, Darin London, Akshay A. Bhinge, Christopher Shestak, Matthew R. Schaner, Seul Ki Kim, Zhuzhu Z. Zhang, Piotr A. Mieczkowski, Joanna O. Mieczkowska, Zheng Liu, Ryan M. McDaniell, Yunyun Ni, Naim U. Rashid, Min Jae Kim, Sheera Adar, Zhancheng Zhang, Tianyuan Wang, Deborah Winter, Damian Keefe, Ewan Birney, Vishwanath R. Iyer, Jason D. Lieb, Gregory E. Crawford, Guoliang Li, Kuljeet Singh Sandhu, Meizhen Zheng, Ping Wang, Oscar J. Luo, Atif Shahab, Melissa J. Fullwood, Xiaoan Ruan, Yijun Ruan, Richard M. Myers, Florencia Pauli, Brian A. Williams, Jason Gertz, Georgi K. Marinov, Timothy E. Reddy, Jost Vielmetter, E. Partridge, Diane Trout, Katherine E. Varley, Clarke Gasper, and The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, 2012.

[33] Kyle P Eagen, Erez Lieberman Aiden, and Roger D Kornberg. Polycomb-mediated chromatin loops revealed by a subkilobase-resolution chromatin interaction map. *Proceedings of the National Academy of Sciences*, 114(33):8764–8769, 2017.

[34] ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) project. *Science*, 306(5696):636–640, 2004.

[35] Jason Ernst and Manolis Kellis. ChromHMM: automating chromatin-state discovery and characterization. *Nature methods*, 9(3):215, 2012.

[36] Fan Feng, Feitong Tang, Yijia Gao, Dongyu Zhu, Tianjun Li, Shuyuan Yang, Yuan Yao, Yuanhao Huang, and Jie Liu. Genomickb: a knowledge graph for the human genome. *Nucleic Acids Research*, 51(D1):D950–D956, 2023.

[37] Fan Feng, Yuan Yao, Xue Qing David Wang, Xiaotian Zhang, and Jie Liu. Connecting high-resolution 3D chromatin organization with epigenomics. *Nature Communications*, 13(1):1–10, 2022.

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

[39] Nadime Francis, Alastair Green, Paolo Guagliardo, Leonid Libkin, Tobias Lindaaker, Victor Marsault, Stefan Plantikow, Mats Rydberg, Petra Selmer, and Andrés Taylor. Cypher: An evolving query language for property graphs. In *Proceedings of the 2018 international conference on management of data*, pages 1433–1445, 2018.

[40] Laiyi Fu, Lihua Zhang, Emmanuel Dollinger, Qinke Peng, Qing Nie, and Xiaohui Xie. Predicting transcription factor binding in single cells through deep learning. *Science Advances*, 6(51):eaba9031, 2020. [PubMed:33355120] [doi:10.1126/sciadv.aba9031].

[41] Geoff Fudenberg, David R Kelley, and Katherine S Pollard. Predicting 3D genome folding from DNA sequence with akita. *Nature Methods*, 17(11):1111–1117, 2020.

[42] Geoff Fudenberg, David R Kelley, and Katherine S Pollard. Predicting 3d genome folding from dna sequence with akita. *Nature methods*, 17(11):1111–1117, 2020. [PubMed:33046897] [PubMed Central:PMC8211359] [doi:10.1038/s41592-020-0958-x].

[43] Geoffrey Fudenberg, Maxim Imakaev, Carolyn Lu, Anton Goloborodko, Nezar Abdennur, and Leonid A Mirny. Formation of chromosomal domains by loop extrusion. *Cell reports*, 15(9):2038–2049, 2016.

[44] Charles P Fulco, Mathias Munschauer, Rockwell Anyoha, Glen Munson, Sharon R Grossman, Elizabeth M Perez, Michael Kane, Brian Cleary, Eric S Lander, and Jesse M Engreitz. Systematic mapping of functional enhancer–promoter connections with CRISPR interference. *Science*, 354(6313):769–773, 2016.

[45] Charles P. Fulco, Joseph Nasser, Thouis R. Jones, Glen Munson, Drew T. Bergman, Vidya Subramanian, Sharon R. Grossman, Rockwell Anyoha, Benjamin R. Doughty, Tejal A. Patwardhan, Tung H. Nguyen, Michael Kane, Elizabeth M. Perez, Neva C. Durand, Caleb A. Lareau, Elena K. Stamenova, Erez Lieberman Aiden, Eric S. Lander, and Jesse M. Engreitz. Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. *Nature Genetics*, 51(12):1664–1669, 2019.

[46] Tianshun Gao and Jiang Qian. Enhanceratlas 2.0: an updated resource with enhancer annotation in 586 tissue/cell types across nine species. *Nucleic acids research*, 48(D1):D58–D64, 2020. [PubMed:31740966] [PubMed Central:PMC7145677] [doi:10.1093/nar/gkz980].

[47] Yury Goltsev, Nikolay Samusik, Julia Kennedy-Darling, Salil Bhate, Matthew Hale, Gustavo Vazquez, Sarah Black, and Garry P Nolan. Deep profiling of mouse splenic architecture with CODEX multiplexed imaging. *Cell*, 174(4):968–981, 2018.

[48] Vishrawas Gopalakrishnan, Kishlay Jha, Aidong Zhang, and Wei Jin. Generating hypothesis: Using global and local features in graph to discover new knowledge from medical literature. In *Proceedings of the 8th International Conference on Bioinformatics and Computational Biology, BICOB*, volume 2016, pages 23–30, 2016.

[49] David U Gorkin, Yunjiang Qiu, Ming Hu, Kipper Fletez-Brant, Tristin Liu, Anthony D Schmitt, Amina Noor, Joshua Chiou, Kyle J Gaulton, Jonathan Sebat, et al. Common DNA sequence variation influences 3-dimensional conformation of the human genome. *Genome Biology*, 20(1):1–25, 2019.

[50] Marion Gremse, Antje Chang, Ida Schomburg, Andreas Grote, Maurice Scheer, Christian Ebeling, and Dietmar Schomburg. The brenda tissue ontology (bto): the first

all-integrating ontology of all organisms for enzyme sources. *Nucleic acids research*, 39(suppl_1):D507–D513, 2010. [PubMed:21030441] [PubMed Central:PMC3013802] [doi:10.1093/nar/gkq968].

[51] GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature*, 550:204–213, 2017. [PubMed:29022597] [PubMed Central:PMC5776756] [doi:10.1038/nature24277].

[52] Lars LP Hanssen, Mira T Kassouf, A Marieke Oudelaar, Daniel Biggs, Chris Preece, Damien J Downes, Matthew Gosden, Jacqueline A Sharpe, Jacqueline A Sloane-Stanley, Jim R Hughes, et al. Tissue-specific ctcf–cohesin-mediated chromatin architecture delimits enhancer interactions and function in vivo. *Nature cell biology*, 19(8):952, 2017.

[53] Jennifer Harrow, Adam Frankish, Jose M Gonzalez, Electra Tapanari, Mark Diekhans, Felix Kokocinski, Bronwen L Aken, Daniel Barrell, Amonida Zadissa, Stephen Searle, et al. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Research*, 22(9):1760–1774, 2012.

[54] Kathryn L Hassell. Population estimates of sickle cell disease in the us. *American journal of preventive medicine*, 38(4):S512–S521, 2010.

[55] Pamela Himadewi, Xue Qing David Wang, Fan Feng, Haley Gore, Yushuai Liu, Lei Yu, Ryo Kurita, Yukio Nakamura, Gerd P Pfeifer, Jie Liu, et al. 3 hs1 ctcf binding site in human β-globin locus regulates fetal hemoglobin expression. *Elife*, 10:e70557, 2021.

[56] Daniel Scott Himmelstein, Antoine Lizee, Christine Hessler, Leo Brueggeman, Sabrina L Chen, Dexter Hadley, Ari Green, Pouya Khankhanian, and Sergio E Baranzini. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *Elife*, 6:e26726, 2017. [PubMed:28936969] [PubMed Central:PMC5640425] [doi:10.7554/eLife.26726].

[57] Michael M Hoffman, Orion J Buske, Jie Wang, Zhiping Weng, Jeff A Bilmes, and William Stafford Noble. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nature methods*, 9(5):473, 2012.

[58] Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard de Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, et al. Knowledge graphs. *Synthesis Lectures on Data, Semantics, and Knowledge*, 12(2):1–257, 2021. [doi:10.2200/S01125ED1V01Y202109DSK022].

[59] Hao Hong, Shuai Jiang, Hao Li, Guifang Du, Yu Sun, Huan Tao, Cheng Quan, Chenghui Zhao, Ruijiang Li, Wanying Li, et al. DeepHiC: A generative adversarial network for enhancing Hi-C data resolution. *PLoS Computational Biology*, 16(2):e1007287, 2020.

[60] Kevin L Howe, Premanand Achuthan, James Allen, Jamie Allen, Jorge Alvarez-Jarreta, M Ridwan Amode, Irina M Armean, Andrey G Azov, Ruth Bennett, Jyothish Bhai, et al. Ensembl 2021. *Nucleic acids research*, 49(D1):D884–D891, 2021. [PubMed:33137190] [PubMed Central:PMC7778975] [doi:10.1093/nar/gkaa942].

[61] Tsung-Han S Hsieh, Claudia Cattoglio, Elena Slobodyanyuk, Anders S Hansen, Oliver J Rando, Robert Tjian, and Xavier Darzacq. Resolving the 3D landscape of transcription-linked mammalian chromatin folding. *Molecular Cell*, 2020.

[62] Tsung-Han S Hsieh, Elena Slobodyanyuk, Anders S Hansen, Claudia Cattoglio, Oliver J Rando, Robert Tjian, and Xavier Darzacq. Resolving the 3d landscape of transcription-linked mammalian chromatin folding. *bioRxiv*, page 638775, 2019.

[63] Elizabeth Ing-Simmons, Vlad C Seitan, Andre J Faure, Paul Flicek, Thomas Carroll, Job Dekker, Amanda G Fisher, Boris Lenhard, and Matthias Merkenschlager. Spatial enhancer clustering and regulation of enhancer-proximal genes by cohesin. *Genome research*, 25(4):504–513, 2015.

[64] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001. [PubMed:11237011] [doi:10.1038/35057062].

[65] Biola M Javierre, Oliver S Burren, Steven P Wilder, Roman Kreuzhuber, Steven M Hill, Sven Sewitz, Jonathan Cairns, Steven W Wingett, Csilla Várnai, Michiel J Thiecke, et al. Lineage-specific genome architecture links enhancers and non-coding disease variants to target gene promoters. *Cell*, 167(5):1369–1384, 2016.

[66] Xiong Ji, Daniel B Dadon, Benjamin E Powell, Zi Peng Fan, Diego Borges-Rivera, Sigal Shachar, Abraham S Weintraub, Denes Hnisz, Gianluca Pegoraro, Tong Ihn Lee, et al. 3D chromosome regulatory landscape of human pluripotent cells. *Cell stem cell*, 18(2):262–275, 2016.

[67] David R Kelley. Cross-species regulatory sequence activity prediction. *PLoS computational biology*, 16(7):e1008050, 2020. [PubMed:32687525] [PubMed Central:PMC7392335] [doi:10.1371/journal.pcbi.1008050].

[68] David R Kelley, Yakir A Reshef, Maxwell Bileschi, David Belanger, Cory Y McLean, and Jasper Snoek. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome research*, 28(5):739–750, 2018. [PubMed:29588361] [PubMed Central:PMC5932613] [doi:10.1101/gr.227819.117].

[69] Aziz Khan and Xuegong Zhang. dbSUPER: a database of super-enhancers in mouse and human genome. *Nucleic acids research*, 44(D1):D164–D171, 2015. [PubMed:26438538] [PubMed Central:PMC4702767] [doi:10.1093/nar/gkv1002].

[70] T.N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. arXiv:1609.02907, 2016.

[71] Sandy L Klemm, Zohar Shipony, and William J Greenleaf. Chromatin accessibility and the regulatory epigenome. *Nature Reviews Genetics*, 20(4):207–220, 2019.

[72] Nils Krietenstein, Sameer Abraham, Sergey Venev, Nezar Abdennur, Johan Gibcus, Tsung-Han Hsieh, Krishna Mohan Parsi, Liyan Yang, Rene Maehr, Leonid Mirny, et al. Ultrastructural details of mammalian chromosome architecture. *bioRxiv*, page 639922, 2019.

[73] Nils Krietenstein, Sameer Abraham, Sergey V Venev, Nezar Abdennur, Johan Gibcus, Tsung-Han S Hsieh, Krishna Mohan Parsi, Liyan Yang, René Maehr, Leonid A Mirny, et al. Ultrastructural details of mammalian chromosome architecture. *Molecular Cell*, 2020.

[74] Ilkka Lappalainen, John Lopez, Lisa Skipper, Timothy Hefferon, J Dylan Spalding, John Garner, Chao Chen, Michael Maguire, Matt Corbett, George Zhou, et al. Dbvar and dgva: public archives for genomic structural variation. *Nucleic acids research*, 41(D1):D936–D941, 2012. [PubMed:23193291] [PubMed Central:PMC3531204] [doi:10.1093/nar/gks1213].

[75] Matthew H. Larson, Luke A. Gilbert, Xiaowo Wang, Wendell A. Lim, Jonathan S. Weissman, and Lei S. Qi. CRISPR interference (CRISPRi) for sequence-specific control of gene expression. *Nature Protocols*, 8(11):2180–2196, 2023.

[76] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

[77] Wenran Li, Wing Hung Wong, and Rui Jiang. DeepTACT: predicting 3D chromatin contacts via bootstrapping deep learning. *Nucleic acids research*, 47(10):e60–e60, 2019.

[78] Wenran Li, Wing Hung Wong, and Rui Jiang. Deeptact: predicting 3d chromatin contacts via bootstrapping deep learning. *Nucleic acids research*, 47(10):e60–e60, 2019. [PubMed:30869141] [PubMed Central:PMC6547469] [doi:10.1093/nar/gkz167].

[79] Xinjun Li, Fan Feng, Hongxi Pu, Wai Yan Leung, and Jie Liu. schictools: A computational toolbox for analyzing single-cell hi-c data. *PLoS computational biology*, 17(5):e1008978, 2021.

[80] E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander, and J. Dekker. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–293, 2009.

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

[82] J. Liu, D. Lin, G. Yardimci, and W. S. Noble. Unsupervised embedding of single-cell Hi-C data. *Bioinformatics (Proceedings of the 27th Conference on Intelligent Systems for Molecular Biology)*, 34:96–104, 2018.

[83] Nan Liu, Victoria V Hargreaves, Qian Zhu, Jesse V Kurland, Jiyoung Hong, Woojin Kim, Falak Sher, Claudio Macias-Trevino, Julia M Rogers, Ryo Kurita, et al. Direct promoter repression by bcl11a controls the fetal to adult hemoglobin switch. *Cell*, 173(2):430–442, 2018.

[84] Qiao Liu, Hairong Lv, and Rui Jiang. hicGAN infers super resolution Hi-C data with generative adversarial networks. *Bioinformatics*, 35(14):i99–i107, 2019.

[85] John Lonsdale, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, Gary Walters, Fernando Garcia, Nancy Young, et al. The genotype-tissue expression (GTEx) project. *Nature Genetics*, 45(6):580–585, 2013.

[86] Renhe Luo, Jielin Yan, Jin Woo Oh, Wang Xi, Dustin Shigaki, Wilfred Wong, Hyein S Cho, Dylan Murphy, Ronald Cutler, Bess P Rosen, et al. Dynamic network-guided crispri screen identifies ctcf-loop-constrained nonlinear enhancer gene regulatory activity during cell state transitions. *Nature genetics*, pages 1–11, 2023.

[87] Qin Ma and Dong Xu. Deep learning shapes single-cell data analysis. *Nature Reviews Molecular Cell Biology*, pages 1–2, 2022. [PubMed:35197610] [PubMed Central:PMC8864973] [doi:10.1038/s41580-022-00466-x].

[88] Wenxiu Ma, Ferhat Ay, Choli Lee, Gunhan Gulsoy, Xinxian Deng, Savannah Cook, Jennifer Hesson, Christopher Cavanaugh, Carol B Ware, Anton Krumm, et al. Using DNase Hi-C techniques to map global and local three-dimensional genome architecture at high resolution. *Methods*, 142:59–73, 2018.

[89] Jacqueline MacArthur, Emily Bowler, Maria Cerezo, Laurent Gil, Peggy Hall, Emma Hastings, Heather Junkins, Aoife McMahon, Annalisa Milano, Joannella Morales, et al. The new nhgri-ebi catalog of published genome-wide association studies (gwas catalog). *Nucleic acids research*, 45(D1):D896–D901, 2017. [PubMed:27899670] [PubMed Central:PMC5210590] [doi:10.1093/nar/gkw1133].

[90] Jeffrey R MacDonald, Robert Ziman, Ryan KC Yuen, Lars Feuk, and Stephen W Scherer. The database of genomic variants: a curated collection of structural variation in the human genome. *Nucleic acids research*, 42(D1):D986–D992, 2014. [PubMed:24174537] [PubMed Central:PMC3965079] [doi:10.1093/nar/gkt958].

[91] James Malone, Ele Holloway, Tomasz Adamusiak, Misha Kapushesky, Jie Zheng, Nikolay Kolesnikov, Anna Zhukova, Alvis Brazma, and Helen Parkinson. Modeling sample variables with an experimental factor ontology. *Bioinformatics*, 26(8):1112–1118, 2010. [PubMed:20200009] [PubMed Central:PMC2853691] [doi:10.1093/bioinformatics/btq099].

[92] Marcos Martínez-Romero, Clement Jonquet, Martin J O'connor, John Graybeal, Alejandro Pazos, and Mark A Musen. Ncbo ontology recommender 2.0: an enhanced approach for biomedical ontology recommendation. *Journal of biomedical semantics*, 8(1):1–22, 2017.

[93] Gabriella E Martyn, Beeke Wienert, Lu Yang, Manan Shah, Laura J Norton, Jon Burdach, Ryo Kurita, Yukio Nakamura, Richard CM Pearson, Alister PW Funnell, et al. Natural regulatory mutations elevate the fetal globin gene via disruption of bcl11a or zbtb7a binding. *Nature genetics*, 50(4):498–503, 2018.

[94] Rachel Patton McCord, Noam Kaplan, and Luca Giorgetti. Chromosome conformation capture and beyond: toward an integrative view of chromosome structure and function. *Molecular cell*, 77(4):688–708, 2020.

[95] Jill E. Moore, Michael J. Purcaro, Henry E. Pratt, Charles B. Epstein, Noam Shoresh, Jessika Adrian, Trupti Kawli, Carrie A. Davis, Alexander Dobin, Rajinder Kaul, Jessica Halow, Eric L. Van Nostrand, Peter Freese, David U. Gorkin, Yin Shen, Yupeng He, Mark Mackiewicz, Florencia Pauli-Behn, Brian A. Williams, Ali Mortazavi, Cheryl A. Keller, Xiao-Ou Zhang, Shaimae I. Elhajjajy, Jack Huey, Diane E. Dickel, Valentina Snetkova, Xintao Wei, Xiaofeng Wang, Juan Carlos Rivera-Mulia, Joel Rozowsky, Jing Zhang, Surya B. Chhetri, Jialing Zhang, Alec Victorsen, Kevin P. White, Axel Visel, Gene W. Yeo, Christopher B. Burge, Eric Lécuyer, David M. Gilbert, Job Dekker, John Rinn, Eric M. Mendenhall, Joseph R. Ecker, Manolis Kellis, Robert J. Klein, William S. Noble, Anshul Kundaje, Roderic Guigó, Peggy J. Farnham, J. Michael Cherry, Richard M. Myers, Bing Ren, Brenton R. Graveley, Mark B. Gerstein, Len A. Pennacchio, Michael P. Snyder, Bradley E. Bernstein, Barbara Wold, Ross C. Hardison, Thomas R. Gingeras, John A. Stamatoyannopoulos, and Zhiping Weng. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature*, 583(7818):699–710, 2020.

[96] Sajid Mughal, Ismail Moghul, Jing Yu, Tristan Clark, David S Gregory, and Nikolas Pontikos. Pheno4j: a gene to phenotype graph database. *Bioinformatics*, 33(20):3317–3319, 2017. [PubMed:28633344] [doi:10.1093/bioinformatics/btx397].

[97] Christopher J Mungall, Carlo Torniai, Georgios V Gkoutos, Suzanna E Lewis, and Melissa A Haendel. Uberon, an integrative multi-species anatomy ontology. *Genome biology*, 13(1):1–20, 2012. [PubMed:22293552] [PubMed Central:PMC3334586] [doi:10.1186/gb-2012-13-1-r5].

[98] T. Nagano, Y. Lubling, C. Várnai, C. Dudley, W. Leung, Y. Baran, N. M. Cohen, S. Wingett, P. Fraser, and A. Tanay. Cell-cycle dynamics of chromosomal organization at single-cell resolution. *Nature*, 547:61–67, 2017.

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

[100] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 807–814, 2010.

[101] David N Nicholson and Casey S Greene. Constructing knowledge graphs and their biomedical applications. *Computational and structural biotechnology journal*, 18:1414–1428, 2020. [PubMed:32637040] [PubMed Central:PMC7327409] [doi:10.1016/j.csbj.2020.05.017].

[102] E. P. Nora, B. R. Lajoie, E. G. Schulz, L. Giorgetti, I. Okamoto, N. Servant, T. Piolot, N. L. van Berkum, J. Meisig, J. Sedat, J. Gribnau, E. Barillot, N. Bluthgen, J. Dekker, and E. Heard. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*, 485(7398):381–385, 2012.

[103] Elphège P Nora, Anton Goloborodko, Anne-Laure Valton, Johan H Gibcus, Alec Uebersohn, Nezar Abdennur, Job Dekker, Leonid A Mirny, and Benoit G Bruneau. Targeted degradation of ctcf decouples local insulation of chromosome domains from genomic compartmentalization. *Cell*, 169(5):930–944, 2017.

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

[105] James K. Nuñez, Jin Chen, Greg C. Pommier, J. Zachery Cogan, Joseph M. Replogle, Carmen Adriaens, Gokul N. Ramadoss, Quanming Shi, King L. Hung, Avi J. Samelson, Angela N. Pogson, James Y. S. Kim, Amanda Chung, Manuel D. Leonetti, Howard Y. Chang, Martin Kampmann, Bradley E. Bernstein, Volker Hovestadt, Luke A. Gilbert, and Jonathan S. Weissman. Genome-wide programmable transcriptional memory by CRISPR-based epigenome editing. *Cell*, 184(9):2503–2519.e17, 2021.

[106] Soohwan Oh, Jiaofang Shao, Joydeep Mitra, Feng Xiong, Matteo D'Antonio, Ruoyu Wang, Ivan Garcia-Bassets, Qi Ma, Xiaoyu Zhu, Joo-Hyung Lee, et al. Enhancer release and retargeting activates disease-susceptibility genes. *Nature*, 595(7869):735–740, 2021.

[107] Masae Ohno, Tadashi Ando, David G Priest, Vipin Kumar, Yamato Yoshida, and Yuichi Taniguchi. Sub-nucleosomal genome structure reveals distinct nucleosome folding motifs. *Cell*, 176(3):520–534, 2019.

[108] Sue Povey, Ruth Lovering, Elspeth Bruford, Mathew Wright, Michael Lush, and Hester Wain. The hugo gene nomenclature committee (hgnc). *Human genetics*, 109(6):678–680, 2001. [PubMed:11810281] [doi:10.1007/s00439-001-0615-0].

[109] V. Ramani, X. Deng, R. Qiu, K. L. Gunderson, F. J. Steemers, C. M. Disteche, W. S. Noble, Z. Duan, and J. Shendure. Massively multiplex single-cell Hi-C. *Nature Methods*, 14(3):263–266, 2017.

[110] S. S. P. Rao, M. H. Huntley, N. Durand, C. Neva, E. K. Stamenova, I. D. Bochkov, J. T. Robinson, A. L. Sanborn, I. Machol, A. D. Omer, E. S. Lander, and E. L. Aiden. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 59(7):1665–1680, 2014.

[111] Suhas SP Rao, Su-Chen Huang, Brian Glenn St Hilaire, Jesse M Engreitz, Elizabeth M Perez, Kyong-Rim Kieffer-Kwon, Adrian L Sanborn, Sarah E Johnstone, Gavin D Bascom, Ivan D Bochkov, et al. Cohesin loss eliminates all loop domains. *Cell*, 171(2):305–320, 2017.

[112] Suhas SP Rao, Miriam H Huntley, Neva C Durand, Elena K Stamenova, Ivan D Bochkov, James T Robinson, Adrian L Sanborn, Ido Machol, Arina D Omer, Eric S Lander, et al. A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159(7):1665–1680, 2014.

[113] Niels J. Rinzema, Konstantinos Sofiadis, Sjoerd J. D. Tjalsma, Marjon J. A. M. Verstegen, Yuva Oz, Christian Valdes-Quezada, Anna-Karina Felder, Teodora Filipovska, Stefan van der Elst, Zaria de Andrade dos Ramos, Ruiqi Han, Peter H. L. Krijger, and Wouter de Laat. Building regulatory landscapes reveals that an enhancer can recruit cohesin to create contact domains, engage CTCF sites and activate distant genes. *Nature Structural & Molecular Biology*, 29(6):563–574, 2022.

[114] Abbas Roayaei Ardakany, Halil Tuvan Gezer, Stefano Lonardi, and Ferhat Ay. Mustache: multi-scale detection of chromatin loops from Hi-C and Micro-C maps using scale-space representation. *Genome Biology*, 21, 2020.

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

[116] Vijay G Sankaran, Tobias F Menne, Jian Xu, Thomas E Akie, Guillaume Lettre, Ben Van Handel, Hanna KA Mikkola, Joel N Hirschhorn, Alan B Cantor, and Stuart H Orkin. Human fetal hemoglobin expression is regulated by the developmental stage-specific repressor bcl11a. *Science*, 322(5909):1839–1842, 2008.

[117] Vijay G Sankaran, Jian Xu, Rachel Byron, Harvey A Greisman, Chris Fisher, David J Weatherall, Daniel E Sabath, Mark Groudine, Stuart H Orkin, Anuja Premawardhena, et al. A functional element necessary for fetal hemoglobin silencing. *New England Journal of Medicine*, 365(9):807–814, 2011.

[118] Vijay G Sankaran, Jian Xu, Tobias Ragoczy, Gregory C Ippolito, Carl R Walkley, Shanna D Maika, Yuko Fujiwara, Masafumi Ito, Mark Groudine, MA Bender, et al. Developmental and species-divergent globin switching are driven by bcl11a. *Nature*, 460(7259):1093–1097, 2009.

[119] Alberto Santos, Ana R Colaço, Annelaura B Nielsen, Lili Niu, Maximilian Strauss, Philipp E Geyer, Fabian Coscia, Nicolai J Wewer Albrechtsen, Filip Mundt, Lars Juhl

Jensen, et al. A knowledge graph to interpret clinical proteomics data. *Nature Biotechnology*, NONE:1–11, 2022. [PubMed:35102292] [doi:10.1038/s41587-021-01145-6].

[120] A. D. Schmitt, M. Hu, I. Jung, Z. Xu, Y. Qiu, C. L. Tan, Y. Li, S. Lin, Y. Lin, C. L. Barr, and B. Ren. A compendium of chromatin contact maps reveals spatially active regions in the human genome. *Cell Reports*, 17:2042–2059, 2016. [PubMed:27851967] [PubMed Central:PMC5478386] [doi:10.1016/j.celrep.2016.10.061].

[121] Stefan Schoenfelder, Robert Sugar, Andrew Dimond, Biola-Maria Javierre, Harry Armstrong, Borbala Mifsud, Emilia Dimitrova, Louise Matheson, Filipe Tavares-Cadete, Mayra Furlan-Magaril, et al. Polycomb repressive complex PRC1 spatially constrains the mouse embryonic stem cell genome. *Nature genetics*, 47(10):1179, 2015.

[122] Wibke Schwarzer, Nezar Abdennur, Anton Goloborodko, Aleksandra Pekowska, Geoffrey Fudenberg, Yann Loe-Mie, Nuno A Fonseca, Wolfgang Huber, Christian H Haering, Leonid Mirny, et al. Two independent modes of chromatin organization revealed by cohesin removal. *Nature*, 551(7678):51, 2017.

[123] Ron Schwessinger, Matthew Gosden, Damien Downes, Richard C Brown, A Marieke Oudelaar, Jelena Telenius, Yee Whye Teh, Gerton Lunter, and Jim R Hughes. DeepC: predicting 3D genome folding using megabase-scale transfer learning. *Nature Methods*, 17(11):1118–1124, 2020.

[124] T. Sexton, E. Yaffe, E. Kenigsberg, F. Bantignies, B. Leblanc, M. Hoichman, H. Parrinello, A. Tanay, and G. Cavalli. Three-dimensional folding and functional organization principles of the *Drosophila* genome. *Cell*, 148(3):458–472, 2012.

[125] Andrew B Stergachis, Brian M Debo, Eric Haugen, L Stirling Churchman, and John A Stamatoyannopoulos. Single-molecule regulatory architectures captured by chromatin fiber sequencing. *Science*, 368(6498):1449–1454, 2020.

[126] Fei Sun, Jianhong Ou, Adam R Shoffner, Yu Luan, Hongbo Yang, Lingyun Song, Alexias Safi, Jingli Cao, Feng Yue, Gregory E Crawford, et al. Enhancer selection dictates gene expression responses in remote organs during tissue regeneration. *Nature cell biology*, 24(5):685–696, 2022.

[127] Haohai Sun, Jialun Zhong, Yunpu Ma, Zhen Han, and Kun He. Timetraveler: Reinforcement learning for temporal knowledge graph forecasting. *arXiv preprint arXiv:2109.04101*, 2021.

[128] M. Sundararajan, A. Taly, and Q. Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, 2017.

[129] Orsolya Symmons, Leslie Pan, Silvia Remeseiro, Tugce Aktas, Felix Klein, Wolfgang Huber, and François Spitz. The shh topological domain facilitates the action of remote enhancers by reducing the effects of genomic distances. *Developmental cell*, 39(5):529–543, 2016.

[130] Orsolya Symmons, Veli Vural Uslu, Taro Tsujimura, Sandra Ruf, Sonya Nassari, Wibke Schwarzer, Laurence Ettwiller, and François Spitz. Functional and topological characteristics of mammalian regulatory domains. *Genome research*, 24(3):390–400, 2014.

[131] Elizabeth A Traxler, Yu Yao, Yong-Dong Wang, Kaitly J Woodard, Ryo Kurita, Yukio Nakamura, Jim R Hughes, Ross C Hardison, Gerd A Blobel, Chunliang Li, et al. A genome-editing strategy to treat β-hemoglobinopathies that recapitulates a mutation associated with a benign genetic condition. *Nature medicine*, 22(9):987–990, 2016.

[132] Taro Tsujimura, Osamu Takase, Masahiro Yoshikawa, Etsuko Sano, Matsuhiko Hayashi, Kazuto Hoshi, Tsuyoshi Takato, Atsushi Toyoda, Hideyuki Okano, and Keiichi Hishikawa. Controlling gene activation by enhancers through a drug-inducible topological insulator. *eLife*, 9:e47980, may 2020.

[133] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

[134] Laura Vian, Aleksandra Pekowska, Suhas SP Rao, Kyong-Rim Kieffer-Kwon, Seolkyoung Jung, Laura Baranello, Su-Chen Huang, Laila El Khattabi, Marei Dose, Nathanael Pruett, et al. The energetics and physiological impact of cohesin extrusion. *Cell*, 173(5):1165–1178, 2018.

[135] Bo Wang, Armin Pourshafeie, Marinka Zitnik, Junjie Zhu, Carlos D Bustamante, Serafim Batzoglou, and Jure Leskovec. Network enhancement as a general method to denoise weighted biological networks. *Nature Communications*, 9(1):3108, 2018.

[136] Xiaotao Wang, Jie Xu, Baozhen Zhang, Ye Hou, Fan Song, Huijue Lyu, and Feng Yue. Genome-wide detection of enhancer-hijacking events from chromatin interaction data in rearranged genomes. *Nature methods*, 18(6):661–668, 2021.

[137] Paul J Werbos. Backpropagation through time: what it does and how to do it. *Proceedings of the IEEE*, 78(10):1550–1560, 1990.

[138] Iain Williamson, Laura A Lettice, Robert E Hill, and Wendy A Bickmore. Shh and ZRS enhancer colocalisation is specific to the zone of polarising activity. *Development*, 143(16):2994–3001, 2016.

[139] Gordana Wutz, Csilla Várnai, Kota Nagasaka, David A Cisneros, Roman R Stocsits, Wen Tang, Stefan Schoenfelder, Gregor Jessberger, Matthias Muhar, M Julius Hossain, et al. Topologically associating domains and chromatin loops depend on cohesin and are regulated by ctcf, wapl, and pds5 proteins. *The EMBO journal*, 36(24):3573–3599, 2017.

[140] Jordan Yupeng Xiao, Antonina Hafner, and Alistair N Boettiger. How subtle changes in 3d structure can create large changes in transcription. *eLife*, 10:e64320, jul 2021.

[141] Dapeng Yang, Hyunwoo Cho, Zakieh Tayyebi, Abhijit Shukla, Renhe Luo, Gary Dixon, Valeria Ursu, Stephanie Stransky, Daniel M Tremmel, Sara D Sackett, et al. Crispr screening uncovers a central requirement for hhex in pancreatic lineage commitment and plasticity restriction. *Nature Cell Biology*, 24(7):1064–1076, 2022.

[142] T. Yang, F. Zhang, G. G. Yardımcı, F. Song, R. C. Hardison, W. S. Noble, F. Yue, and Q. Li. HiCRep: assessing the reproducibility of Hi-C data using a stratum-adjusted correlation coefficient. *Genome Research*, 27(11):1939–1949, 2017.

[143] Lin Ye, Jiaming Wang, Yuting Tan, Ashley I Beyer, Fei Xie, Marcus O Muench, and Yuet Wai Kan. Genome editing using crispr-cas9 to create the hpfh genotype in hspcs: An approach for treating sickle cell disease and β-thalassemia. *Proceedings of the National Academy of Sciences*, 113(38):10661–10665, 2016.

[144] Byoung-Ha Yoon, Seon-Kyu Kim, and Seon-Young Kim. Use of graph database for the integration of heterogeneous biological data. *Genomics & informatics*, 15(1):19, 2017. [PubMed:28416946] [PubMed Central:PMC5389944] [doi:10.5808/GI.2017.15.1.19].

[145] Jingting Yu, Ming Hu, and Chun Li. Joint analyses of multi-tissue Hi-C and eQTL data demonstrate close spatial proximity between eQTLs and their target genes. *BMC Genetics*, 20(1):43, 2019.

[146] Shilu Zhang, Deborah Chasman, Sara Knaack, and Sushmita Roy. In silico prediction of high-resolution Hi-C interaction matrices. *Nature Communications*, 10(1):1–18, 2019.

[147] Shilu Zhang, Deborah Chasman, Sara Knaack, and Sushmita Roy. In silico prediction of high-resolution hi-c interaction matrices. *Nature communications*, 10(1):1–18, 2019. [PubMed:31811132] [PubMed Central:PMC6898380] [doi:10.1038/s41467-019-13423-8].

[148] Y. Zhang, C. H. Wong, R. Y. Birnbaum, G. Li, R. Favaro, C. Y. Ngan, J. Lim, E. Tai, H. M. Poh, E. Wong, F. H. Mulawadi, W. K. Sung, S. Nicolis, N. Ahituv, Y. Ruan, and C. L. Wei. Chromatin connectivity maps reveal dynamic promoter-enhancer long-range associations. *Nature*, 504(7479):306–10, 2013.

[149] Yan Zhang, Lin An, Jie Xu, Bo Zhang, W Jim Zheng, Ming Hu, Jijun Tang, and Feng Yue. Enhancing Hi-C data resolution with deep convolutional neural network HiCPlus. *Nature Communications*, 9(1):750, 2018.

[150] Jian Zhou, Chandra L Theesfeld, Kevin Yao, Kathleen M Chen, Aaron K Wong, and Olga G Troyanskaya. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nature genetics*, 50(8):1171–1179, 2018. [PubMed:30013180] [PubMed Central:PMC6094955] [doi:10.1038/s41588-018-0160-6].

[151] Yun Zhu, Zhao Chen, Kai Zhang, Mengchi Wang, David Medovoy, John W Whitaker, Bo Ding, Nan Li, Lina Zheng, and Wei Wang. Constructing 3D interaction maps from 1D epigenomes. *Nature communications*, 7(1):1–11, 2016.

[152] Jessica Zuin, Gregory Roth, Yinxiu Zhan, Julie Cramard, Josef Redolfi, Ewa Piskadlo, Pia Mach, Mariya Kryzhanovska, Gergely Tihanyi, Hubertus Kohler, Mathias Eder, Christ Leemans, Bas van Steensel, Peter Meister, Sebastien Smallwood, and Luca Giorgetti. Nonlinear control of transcription through enhancer–promoter interactions. *Nature*, 604(7906):571–577, 2022.