McGill University

Doctoral Thesis

Reference-panel based 3D genomics analysis

Author:

Yanlin Zhang

Supervisor:

Dr. Mathieu Blanchette

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Doctor of Philosophy

School of Computer Science

McGill University, Montreal

December 8, 2023

To my wife Weiwei Liu and my son Aiden Zhang for their love, accompaniment, and support.

Abstract

The complex and dynamic three-dimensional genome structure is crucial for regulating cellular activity. Recent advancements in chromosome conformation capture (3C) methods and high-resolution imaging techniques allow us to study the 3D genome at unprecedented scales. Among these techniques, Hi-C has emerged as a prominent tool over the past decade. The widespread usage of Hi-C has revealed the hierarchical structures of the genome, thereby deepening our understanding of the organization and function of 3D genomes. However, analyzing Hi-C data remains a challenging task, mainly due to the sequencing coverage of data produced in most Hi-C experiments is insufficient.

In this thesis, we proposed a reference panel enabled framework to tackle the data insufficiency issue in Hi-C data analysis. This pioneering approach represents the first instance of harnessing the vast amount of existing Hi-C datasets while analyzing a given study Hi-C dataset. Within this framework, we developed three applications to enhance a Hi-C contact map, annotate chromatin loops, and identify nested topologically associating domains (TADs) from insufficiently sequenced Hi-C data. Algorithms developed in this thesis leverage ideas from attention mechanisms, representation learning, dynamic programming, and non-parametric statistics. The introduction of a panel of reference Hi-C samples significantly improved prediction accuracy across three diverse Hi-C data analysis tasks under a wide spectrum of benchmarking scenarios. Applying our tools to Hi-C data from various cells deepened our understanding of the formation of TADs and chromatin loops, unraveling key insights into these essential genomic features. Taken together, this thesis provides a new paradigm to perform 3D genomics study at high resolution.

Abrégé

La structure tridimensionnelle complexe et dynamique du génome est cruciale pour la régulation de l'activité cellulaire. Les progrès récents dans les méthodes de capture de conformation chromosomique (3C) et les techniques d'imagerie à haute résolution nous permettent d'étudier le génome 3D à des échelles sans précédent. Parmi ces techniques, Hi-C est devenue un outil important au cours de la dernière décennie. L'application généralisée du Hi-C a révélé les structures hiérarchiques du génome, approfondissant ainsi notre compréhension de l'organisation et de la fonction des génomes 3D. Cependant, l'analyse des données Hi-C reste une tâche difficile, en grande partie parce que les données produites dans la plupart des expériences Hi-C sont insuffisantes.

Dans cette thèse, nous avons proposé un cadre activé par un panel de référence pour résoudre le problème de l'insuffisance des données dans l'analyse des données Hi-C. Cette approche pionnière représente le premier exemple d'exploitation de la grande quantité d'ensembles de données Hi-C existants tout en analysant un ensemble de données Hi-C d'étude donné. Dans ce cadre, nous avons développé trois applications pour améliorer une carte de contact Hi-C, annoter les boucles de chromatine et identifier les *topologically associating domains* (TAD, domaines d'association topologique) imbriqués à partir de données Hi-C insuffisamment séquencées. Les algorithmes développés dans cette thèse exploitent des idées issues des mécanismes d'attention, de l'apprentissage des représentations, de la programmation dynamique et des statistiques non paramétriques. L'introduction d'un panel d'échantillons de référence Hi-C a considérablement amélioré la précision des prévisions dans trois tâches d'analyse de données Hi-C diverses dans un large éventail de scénarios d'analyse comparative. L'application de nos outils aux données Hi-C de diverses cellules a approfondi notre compréhension de la formation des TAD et des boucles

de chromatine, révélant ainsi des informations clés sur ces caractéristiques génomiques essentielles. Dans l'ensemble, cette thèse fournit un nouveau paradigme pour réaliser des études génomiques 3D à haute résolution.

Acknowledgements

I would not have been able to complete this thesis without the support of many individuals. I would like to thank the following individuals for their support over the years:

Christopher JF Cameron, Zichao Yan, Audrey Baguette, Dongjoon Lim, Chris Drogaris, Elliot Layne, Jeanne Tous, Bowei Xiao, Ming Yang Zhou, Faizy Ahsan, Samy Coulombe and many more forgotten but appreciated nonetheless.

examiners of my comprehensive and proposal exams:

Jérôme Waldispühl, Xue Liu, David Rolnick, and Jörg Kienzle.

members of my Ph.D. Supervisory Committee:

Jacek Majewski, and Yue Li

internal and external examiners of my Ph.D. thesis:

Jian Ma, and Yue Li

members of my Ph.D. Defence Committee:

Luc Devroye, Mathieu Blanchette, Yue Li, Jérôme Waldispuhl, and Robert Sladek

and, of course, my Ph.D. advisor Mathieu Blanchette for his unwavering support and patience in addressing my countless questions, even on the minutest scientific details. His mentorship not only introduced me to the captivating field of computational biology, but also left an indelible mark on all facets of my life.

Contents

List of Figures

List of Tables

List of Abbreviations

ADN	**A**cide **D**ésoxyribo**N**ucléique
bp	**b**ase **p**air
ChIA-PET	**Ch**romatin **I**nteraction **A**nalysis by **P**aired-**E**nd **T**ag Sequencing
ChIP	**Ch**romatin **I**mmuno**P**recipitation
ChIP-seq	**ChIP-seq**uencing
CTCF	**CCCTC**-binding **F**actor
DNA	**D**eoxyribo**n**ucleic **A**cid
E-P	**E**nhancer-**P**romoter
FDR	**F**alse **d**iscovery **r**ate
FISH	**F**luorescence **I**n **S**itu **H**ybridization
GAM	**G**enome **A**rchiture **M**apping
GPU	**G**raphics **P**rocessing **U**nit
ICE	**I**terative **C**orrection and **E**igenvalue decomposition
kb	**k**ilo**b**ase
Mb	**M**ega**b**ase
MAE	**M**ean **a**bsolute **e**rror
MSE	**M**ean **s**quared **e**rror
PolII	RNA **Pol**ymerase **II**
PCC	**P**earson **c**orrelation **c**oefficient
PSNR	**P**eak **s**ignal to **n**oise **r**atio
SRCC	**S**pearman **r**ank **c**orrelation **c**oefficient

SSIM **S**tructural **s**imilarity **i**ndex **m**easurement

TAD **T**opologically **A**ssociating **D**omain

Chapter 1

Introduction

1.1 Overview

Over the past decades, scientists have increasingly realized the importance of the three-dimensional (3D) genome architecture in cellular activity [1]. Recent advances in imaging techniques [2] and in chromosome conformation capture methods [3, 4, 5] have revealed the complex and dynamic 3D configuration of the genome during cell differentiation and development, which significantly deepen our understanding of 3D genomics [1, 6]. Although these new techniques are widely used in 3D genomics studies, analyzing these experimental outputs is still in its infancy.

The human genome is very tiny but contains billions of base pairs. We are unable to use a microscope to visualize the whole genome at high resolutions [2]. The most widely used technology in 3D genome study is Hi-C [3], which is a sequencing-based technique. In a Hi-C experiment, DNA fragments in close proximity are ligated and identified through massively parallel sequencing, and the number of ligated fragments spanning two genomic regions reflecting loci proximity are stored in a matrix known as a contact map. Due to the unavoidable high sequencing cost in carrying out high-coverage Hi-C experiment, most published Hi-C experiment only captures hundreds of millions of contact pairs, which is small in comparison to the number of contact pairs that need to be

used to accurately estimate the whole genome level contact frequency in a Hi-C contact map. A typical Hi-C experiment only produces a very sparse Hi-C contact map, which represents a significant barrier in the downstream analyses.

Since the introduction of the original Hi-C protocol (i.e., dilution Hi-C) [3], researchers have developed improved protocols of the chromosomal conformation capture experiment, such as in-situ Hi-C [4] and micro-C [5]. These improved protocols are still inefficient in producing a high-coverage Hi-C contact map. We anticipate that this challenge will remain at least until the sequencing cost drops significantly. Hence, the processing and analysis of Hi-C raw data, characterized by its sparse and large volume, will continue to require sophisticated algorithms and models.

Current tools in computational genomics focus on the analysis of individual Hi-C data sets of interest, without taking advantage of the fact that several hundred Hi-C contact maps are publicly available. Their performance is limited by the sequencing coverage of the Hi-C data in the analysis. Although several tools [7, 8, 9] have been proposed to handle very low coverage Hi-C contact maps, there is still room for improvement. In this thesis, we introduce a reference-panel enabled data analysis framework into computational 3D genomics and develop three applications for different tasks using this framework. Each application outperformed alternative tools at the time of publication.

This chapter begins with an overview of the biological background on the 3D genome and wet lab experiments involved in capturing 3D genome conformation. We then provide an extensive review of existing tools in computational 3D genomics. Last, we recap bioinformatics applications in other domains that motivate our research and essential machine learning background for this thesis.

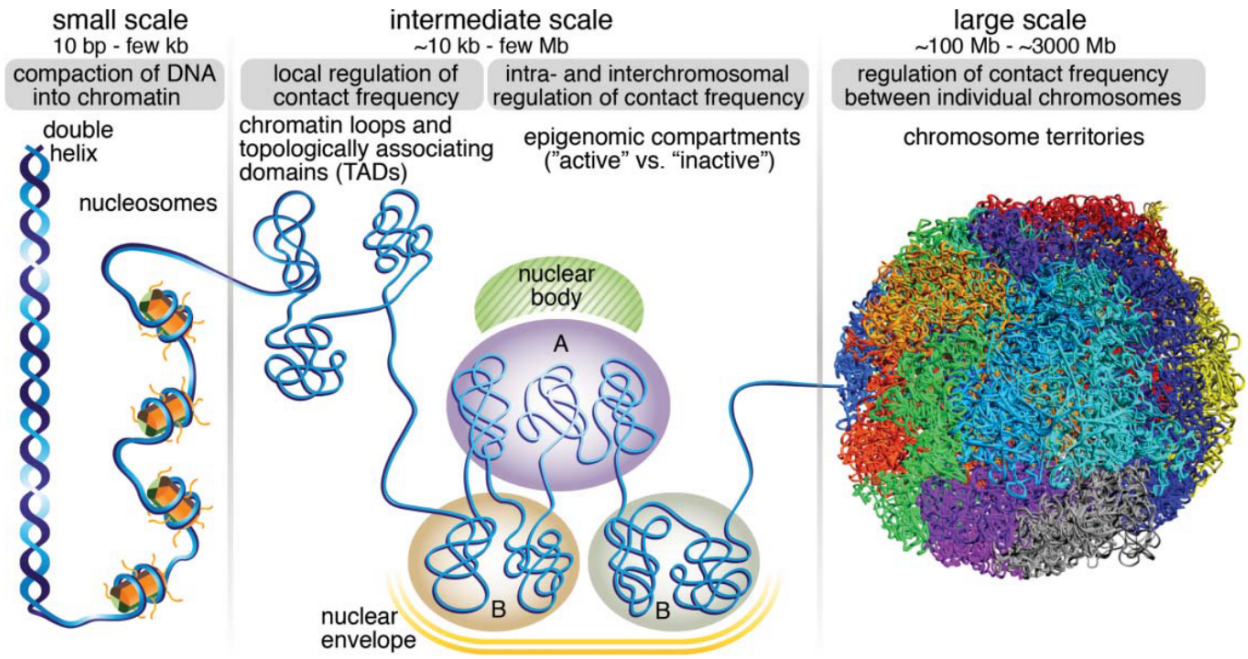


Figure 1.1: The hierarchical organization of the 3D genome. This figure is reproduced from Hansen *et al.* [10].

1.2 Multiscale 3D genome organization

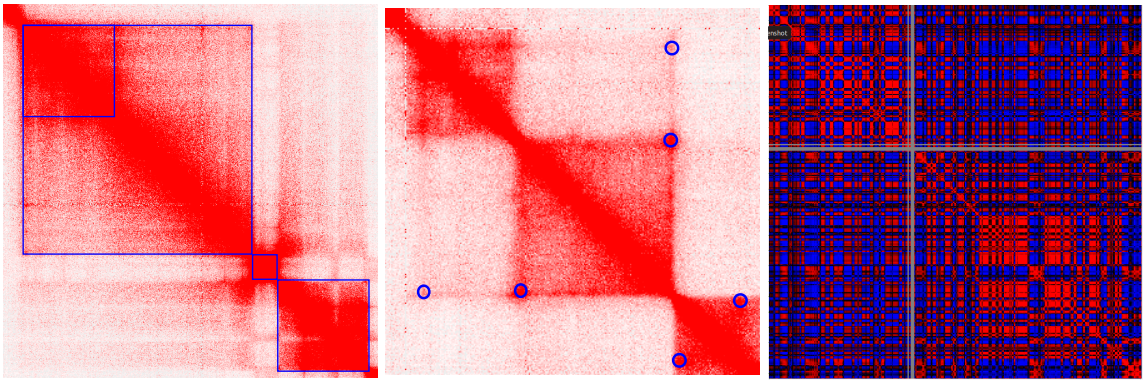
The DNA sequences (i.e., chromosomes) of the human genome would span approximately 2 meters if stretched end-to-end. As illustrated in Fig. 1.1, these linear pieces need to undergo extensive packaging to fit within the micron-scale nucleus and are folded at multi-scales. At the small scale, nucleotides are folded into a double helix. The first condensation level happens when 147 base pairs (bps) of nucleotides wrap around a histone octamer to form a 11 nm nucleosome. At the intermediate scale of 10 kb to several Mb, chromosomes can fold into topologically associating domains (TADs), chromatin loops, and A/B compartments [1].

TADs are kilo- to mega-scale genomic regions with strong interactions among DNA fragments within the same domain, accompanied by weaker interactions across adjacent domains [11, 4]. These domains manifest as square-shaped regions with enriched interaction frequencies along the main diagonal of Hi-C contact maps generated through Hi-C

experiments (Section 1.4, Fig. 1.2a). TADs are first identified as mega-scale regions from low resolution Hi-C contact maps [11, 4]. As high-resolution Hi-C contact maps and improved computational tools become available, it is now clear that smaller subTADs, spanning sub-megabase scales and characterized by higher interaction frequencies, are nested within the TADs in mammalian genomes [6]. This hierarchical organization highlights the intricate and multi-scale nature of chromosomal architecture [1]. Although mechanisms underpinning TAD formation and their functional roles remain poorly understood, a prevailing hypothesis suggests that some TADs are created by the loop extrusion mechanism (which will be introduced in a later paragraph) [12]. Other mechanisms such as compartmentalization can also contribute to TAD formation [6].

Chromatin loops are defined as pairs of genomic loci that are sequentially distant but come into spatial proximity through a mechanism hypothesized to be loop extrusion [1]. These loops manifest as blob-shaped patterns associated with increased interaction frequencies in the off-diagonal region of Hi-C contact maps (Fig. 1.2b). This long-range interaction can be mediated by various proteins, including CTCF and cohesin, which play essential roles in loop formation. The occurrence of convergent CTCF binding motifs leads to the termination of the loop extrusion process. Thus, pairs of loop anchors often colocalize with convergent CTCF binding sites [1]. Additionally, CTCF motif pairs in other directionality are also observed at loop anchors [4]. Some loops that occur at TAD corners contribute to the formation of TADs, which are referred to as "loop TADs" [6].

Two types of compartment, denoted as A and B, were initially identified as mega-scale structures based on both spatial proximity and epigenomic features of chromatin at 1 Mb resolution [3]. Within these compartments, genomic fragments in one type of compartment tend to interact more often with other fragments belonging to the same type of compartment. Thus, a plaid pattern emerges if we transform a Hi-C contact map into a correlation matrix by defining entry (i, j) as the Pearson correlation between column i



(a) (b) (c)

Figure 1.2: Examples of Hi-C contact map annotated with (a) TADs, and (b) chromatin loops. (c) a correlation matrix transformed from a Hi-C contact map.

and *j* of the Hi-C contact map (Fig. 1.2c). A compartments ("active" compartments) often associate with open chromatins and are gene-rich and transcriptionally active regions. In contrast, B compartments ("inactive" compartments) are gene-poor, transcriptionally inactive regions characterized by more condensed chromatin [3]. More recently, higher-resolution Hi-C data suggested that A/B compartments can be further subdivided into a minimum of six subcompartments [4, 13].

At the largest scale, each chromosome tends to occupy a particular region within the nucleus to form chromosome territories [3, 10]. These chromosome territories enable genomic regions of one chromosome to interact more frequently with regions from the same chromosome than from other chromosomes.

Despite the hypothesis of how DNA polymer organized in 3D space was described at early as in 1882 [15], and many computational [16, 17, 18, 19, 20] and wet-lab experiments [3, 4, 5, 21] have been developed to study this question, the way chromatin folds within the nuclei remains poorly understood. Several hypotheses have been proposed to model 3D genome structures, among which, the loop extrusion model (Fig. 1.3) [12, 22, 19] is the most widely accepted hypothesis. Though the loop extrusion model is still unproven, this model succeeds in explaining the formation of TADs and loops identified from Hi-C

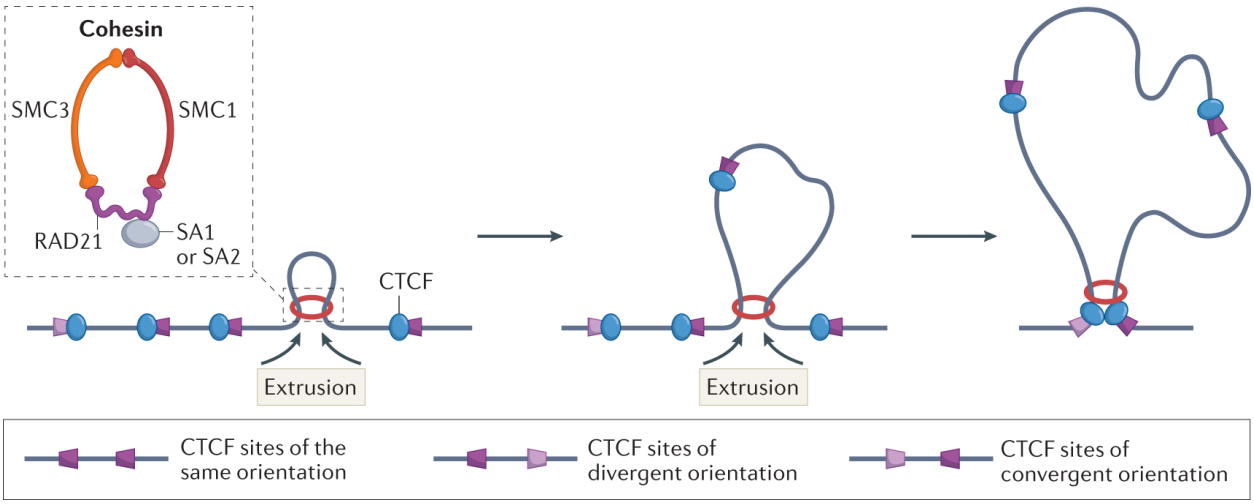


Figure 1.3: An overview of the cohesin-mediated loop extrusion model. This figure is reproduced from Zhang *et al.* [14].

experiments [22]. In the loop extrusion model, a cohesin complex (composed of SMC1, SMC3 and RAD21) binds to a DNA molecule and start to reel flanking regions of the binding site into a loop. Conseqently, the formation of loops is a dynamic process and the size of a chromatin loop would gradually increase as the cohesin complex progressively extrude more nucleotides into the loop. The loop extrusion process terminates when the cohesin complex reaches convergently oriented CTCF [14].

1.3 Biological Importance of 3D genome organization

These multi-scale structures play a vital role in gene-gene interaction and gene regulation, both within cells and throughout cellular differentiation [1]. The primary functions of TADs are self-interaction and insulation. They restrict chromatin interactions, such as enhancer-promoter interactions, within their respective domains (i.e., self-interaction) [23]. their boundaries are barriers that inhibit enhancers within one TAD from establishing interactions with promoters situated outside of their designated domain (i.e., insulation). In addition, TAD boundaries correlate with many linear genomic features. For example, they are enriched for CTCF (CCCTC-binding factor), transcription factors, and

histone marks such as H3K4me3 and H3K36me3 [1]. Chromatin loops can also facilitate the regulation of gene expression. They enable enhancer and promoter interactions by bringing distant genomic elements into close contact.

Apart from controlling transcription, the spatial arrangement of the genome is vital in cell development and disease, such as DNA replication, cell division, [24] and cancers [25]. Aberrations in chromatin folding, modifications in interactions among regulatory elements can result in the misregulation of crucial genes linked to cell growth, differentiation, and apoptosis and can cause the development of diverse cancers [26, 25, 27]. For instance, specific TAD fusion or the inversion of a piece of DNA fragment around specific TAD boundaries can cause interactions between enhancers and non-target promoters and consequent impact gene expression [28].

1.4 Capturing Chromosome Conformation

Most early studies of genome organization rely on the combination of fluorescent in-situ hybridization (FISH) and imaging experiments to assess the spatial position of several genomic loci inside the nuclei [30]. While this remains an excellent tool for the direct measurement of spatial locations of genomic fragments, and throughput and resolution has been improved in recent years [2, 31], most studies in the past decade are conducted with chromosome conformation capture (3C) [32] and a series of subsequent approaches including 4C [33], 5C [34], Hi-C [3], micro-C [35], ChIA-PET [36], and HiCHiP [37] (Fig. 1.4). The development of these capture techniques is a major breakthrough in 3D genome study and their applications have deepen our understanding of 3D genome organization. These approaches allow us to study a 3D genome at different scale (i.e., chromatin segments, whole genome, etc.) under various resolutions (from several Kb to several Mb). This section provides an overview of these capture-based approaches.

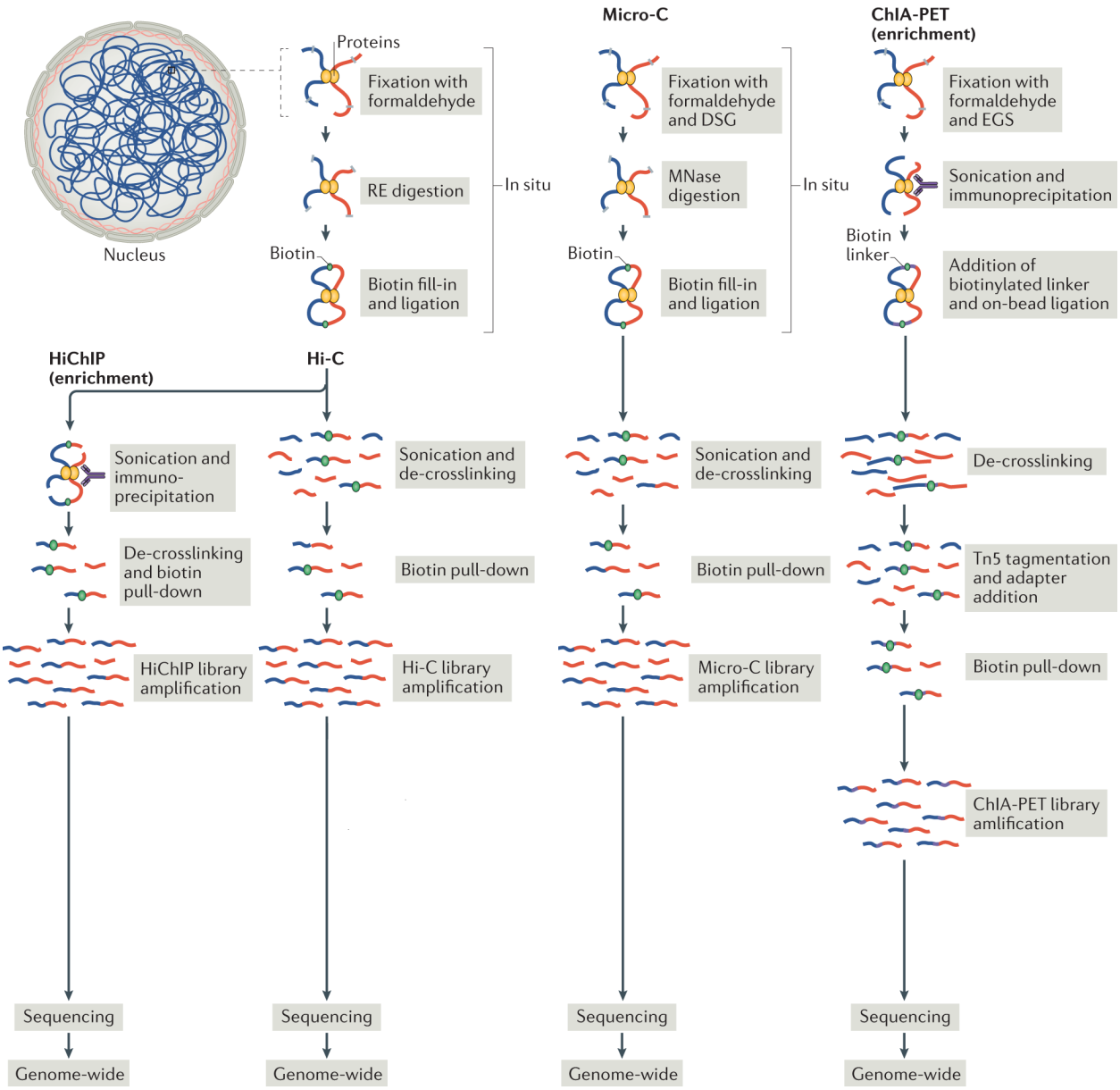


Figure 1.4: The hierarchical organization of the 3D genome. This figure is reproduced from Jerkovic & Cavalli [29].

Hi-C [3] is a tool to capture spatial interactions of DNA fragments at the scale of the entire genome (i.e., all versus all). It and its derivatives have become the most influential techniques in 3D genome organization research in the past decade. Hi-C was initially introduced by Lieberman-Aiden *et. al.* in 2009 [3] and has been improved several times [4, 38]. It is widely used in identifying topologically genome structures including compartments, topologically associating domains, and chromatin loops. Generally, a Hi-C library preparation involves the following six steps: (1) Crosslinking spatially close chromatin segments with formaldehyde; (2) Digesting chromatin with a restriction enzyme. This step is usually performed with a 4-cutter enzyme to achieve a high resolution; (3) Filling the 5'-overhangs of digested fragments with nucleotides including one nucleotide marked with biotin (i.e., biotinylated residue), and ligating the blunt-end fragments; (4) fragmenting all samples by sonication and reversing the crosslinking; (5) Isolating and selecting biotinylated ligation junctions with streptavidin beads; (6) Amplifying selected samples to generate a Hi-C library. The amplified Hi-C library is finally subjected to sequencing. The first three steps are performed in solution in the original Hi-C protocol (i.e., dilution Hi-C) [3], but are performed in intact nuclei (i.e., in situ Hi-C)[4] now.

The in situ protocol enhances the efficiency of capturing true contacts; however, the resolution is constrained by the choice of restriction enzyme in use. The average length of fragments produced by a 4-cutter enzyme is 256 bp, so no matter how we increase the sequencing coverage of a Hi-C experiment, the finer-scale structures that we can study are restricted by the distribution of the enzyme-cutting sites. To overcome the limitation introduced by restriction enzyme digestion, we can perform a micro-C experiment [35]. As illustrated in Fig. 1.4, micro-C differs from Hi-C in the first two steps. It performs double cross-linking and uses micrococcal nuclease (MNase) to digest chromatins. This modification allows micro-C to capture more *cis*-interactions and produce more uniformly distributed fragments with a length as short as $\sim$147 bp.

To study genome-wide chromatin interactions mediated by a protein of interest, we can perform enrichment-based experiments such as HiCHIP (Hi-C Chromatin immunoprecipitation) [37] and ChIA-PET (Chromatin Interaction Analysis with Paired-End Tag sequencing) [36]. HiCHIP differs from the Hi-C protocol by introducing immunoprecipitation after fragmentation. ChIA-PET combines ChIP-based enrichment and chromosome conformation capture (3C) techniques to produce results that are similar to a HiCHIP experiment. To study the fine-scale genome organization at specific regions (i.e., promoter, etc.), techniques such as Capture-C [39] have been developed to capture contacts in specific genomic regions. These approaches can produce very high coverage Hi-C contact map at genomic regions of interest. The above-described techniques are now widely used to study 3D genome conformation at the whole genome level, interactions mediated by a protein of interest, or predefined genomic regions [29].

Recently, several single-cell Hi-C (scHi-C) technologies have emerged for the examination of 3D genome features [40, 41, 42, 43]. These methodologies enable researchers to investigate cell-to-cell variability and the dynamics of chromatin conformation at the single-cell level. Nevertheless, it is noteworthy that contact maps generated through these scHi-C approaches exhibit a high degree of sparsity with a substantial fraction of missing interactions [44].

Meanwhile, other approaches such as GAM [45], SPRITE [46] and GPSeq [47] have also been developed in recent years. Hi-C fails at detecting high-order interactions as well as long distance interactions. In contrast, these non-capturing based approaches do not rely on proximity ligation and overcome these limitations as observed in a Hi-C experiment.

1.5 Processing, storage and visualization of chromatin contact maps

These sequencing-enabled capture approaches generate a tremendous number of reads that require additional steps to create a Hi-C contact map. Although this procedure is sophisticated and computationally intensive. Mature pipelines [48, 49, 50] exist for creating contact maps. Most of these pipelines are built for processing Hi-C data but can be used to analyze other capture-based data with minimal adaption.

The raw output of a Hi-C experiment is a set of read pairs. A typical workflow in Hi-C data analysis involves quality control, aligning reads to a reference genome with short read alignment algorithms [51, 52], contact (i.e., read pair) filtering, splitting the whole genome into fixed-size bins, and counting the number of read pairs spanning two given bins. These read counts indicate pairwise proximity among chromatin fragments. We represent them as a matrix known as a contact map. Several formats exist to store a contact map [49, 48]. Among these formats, *.cool* [49] is most widely used in recent years.

Due to the existence of various biases in a Hi-C experiment, the contact map is only a poor estimation of chromatin interactions. An essential step is to normalize the Hi-C contact map to infer an unbiased estimation of the interaction frequencies from the read count matrix. There are three widely used approaches: Removing bias in the contact map via loci coverage normalization [4]; Modelling bias in a generalized linear model and finding the bias by solving a Poisson or negative binomial regression [53]; Removing bias implicitly via matrix balancing such that sums of each row and column are equal [4]. At least one of these data normalization procedures is included in most Hi-C data analysis pipelines.

In practice, Hi-C contact maps are visualized as heatmaps. For example, Fig. 1.5 shows several Hi-C contact maps with and without normalization at different resolutions (i.e.,

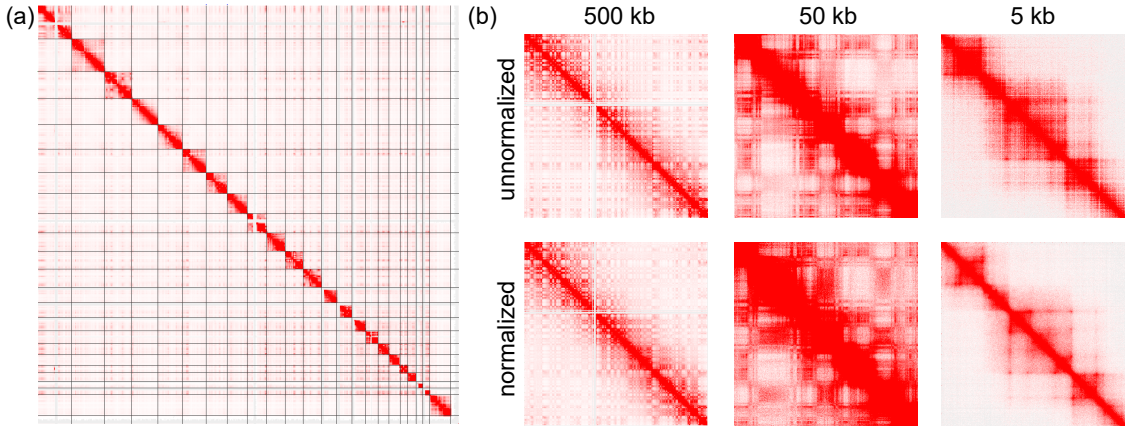


Figure 1.5: **Hi-C contact map examples. a,** a whole genome Hi-C contact map. **b,** Hi-C contact maps for small regions at different resolutions with and without data normalization.

different bin sizes). Hi-C contact map visualization tools such as HiGlass [54], Juicebox [55], and HiCExplorer [56] allow user to visualize Hi-C contact maps (with and without normalization) with annotations (i.e., chromatin loops, TADs, etc.). In addition, tools such as WashU Epigenome Browser [57] and Nucleome Browser [58] allow users to visualize Hi-C contact maps, imaging data, and 3D genome structures simultaneously.

1.6 Computational Tools and Challenges in 3D genomics data analysis

We can classify existing bioinformatics tools for Hi-C (or Hi-C like) data analysis into two categories, basic tools for Hi-C data processing and analysis tools for Hi-C contact maps. This thesis focuses on the second category. A typical coverage (i.e., containing 200M-300M valid read pairs) Hi-C contact map is large and sparse. For example, a genome-wide Hi-C contact map contains more than $600,000$ rows and columns with most of its entries being zero at 5kb resolution. Thus, the analysis of Hi-C contact maps is challenging. There are several tools that exist to analyze Hi-C contact maps. However, in the context of high resolution Hi-C data analysis, most of them only perform well in analyzing

high-coverage (i.e., containing billions of valid read pairs) Hi-C contact maps. Although several approaches have been introduced to address the issue of insufficient sequencing depth [7, 59, 60], there is still room for improvement. In the following paragraphs of this section, we provide an overview of existing tools for three different Hi-C data analysis tasks.

1.6.1 Chromatin loop annotation

Chromatin loops that bring distant loci into close contact play essential roles in biological processes, for example, they permit the interaction between enhancers and promoters. In 3D genome studies, the most widely used approach to study chromatin loops is using a Hi-C experiment. The sequential distance between two loci involved in a chromatin loop ranges from tens to thousands of kilobases. In a Hi-C contact map, a chromatin loop appears as a blob-shaped pattern with more interactions than its surrounding area in off-diagonal regions of the matrix (Fig. 1.2a). There are global enrichment approaches, local enrichment approaches, and non-enrichment approaches that exist for loop detection from Hi-C contact maps. (i) In a global enrichment approach, we define loops as statistical significant chromatin contacts. Existing tools, such as Fit-Hi-C [61] and HiC-DC [62], usually fit a global model to estimate the background distribution of the interaction frequency and identify statistically significant contact pairs by comparing observed values to expected values from the fitted model. The global models are usually defined as generalized linear models with variables including genomic distance between contact pairs, GC content, and mappability of contact pairs, etc. Global enrichment methods assume contact pairs are independent and identically distributed. They do not take surrounding patterns into consideration, thus they identify loop clusters instead of discrete loops. (ii) In contrast, a local enrichment approach takes local patterns into consideration. Most

of the recently developed loop annotation tools are local enrichment approaches. Pioneer tools such as HiCCUPS [4] compare each contact pair to surrounding regions and identify a contact pair with a significantly larger value than its surrounding regions as a loop. These approaches usually require users to set several data coverage sensitive parameters and can only detect loops that satisfy a set of user-defined filtering criteria. (iii) Recently, some researchers started treating loop annotation as a special type of pattern detection problem. They use computer vision or supervised machine learning techniques to detect loops. For example, Mustache [18] treats loop recognition as a blob-shaped object detection problem. Chromosight [16] defines a generic kernel to represent loops and uses this kernel to scan for loops from a Hi-C contact map. These generic pattern-based approaches work well on data with sufficient contact pairs but underperform at low sequencing depths [17]. More recently, several data-driven approaches have been developed to detect loops. For instance, Peakachu [17] is a supervised learning approach trained to recognize loops using data from orthogonal experiments as target values. These data-driven approaches learn loop patterns from data. Although several tools exist for loop annotation, their performances are limited by the sequencing depth of the sample in the study.

1.6.2 Topologically associating domain annotation

TAD annotation has attracted the most attention among all tasks in computational 3D genomics in recent years. TADs are regions with an increasing level of interactions among loci within the same region. The most widely used approach to detect TADs is to use a Hi-C experiment [63]. We can annotate TADs from a Hi-C contact map by detecting squares along the diagonal that associate with more contact pairs than neighboring regions (Fig. 1.2b) [6]. Initially, researchers annotated TAD from Hi-C datasets at low resolution and identified TADs as megabase-scale structural elements [11]. As more and more

high-coverage Hi-C contact maps became available, it became clear that TADs are hierarchically organized and can actually range from tens to hundreds kilobases in mammalian chromosomes [64].

Existing tools [65, 66, 67, 4, 68, 7, 11, 69, 70] for TAD annotation can be classified as either one-dimensional (1D) score-based and or matrix-based approaches. The former, such as TopDom [65], Insulation Score (IS) [69], OnTAD [71] assign each locus a score representing the strength of a potential TAD boundary. Subsequently, these tools detect TAD boundaries by identifying local peaks among these scores. Matrix-based methods directly utilize Hi-C data in two-dimension. For instance, Arrowhead [4] transforms the Hi-C contact map into an arrowhead-shaped feature map and detects TADs by searching for corners in the transformed map. Despite the numerous TAD annotation tools available, the detection of TAD hierarchy and the precise location of TAD boundaries at high resolution remains challenging. Most TAD callers are designed for dense contact maps. As reviewed in previous studies [63, 72, 73, 74], many TAD callers are sensitive to variations in resolution and sequencing coverage. Furthermore, TAD predictions often demonstrate limited consensus among different tools.

1.6.3 Contact map enhancement

Due to sequencing costs, we usually perform Hi-C experiments at the coverage of containing 200M-300M valid read pairs. However, Hi-C data analysis usually requires Hi-C contact maps to contain 500M or even billions of read pairs. To fill the gap, we can produce an *in silico* enhanced contact map using a low-coverage Hi-C contact map as input. This procedure is known as contact map enhancement or super-resolution inference. Several tools exist for contact map enhancement. Contact map enhancement is primarily driven by deep learning algorithms. Most of the existing contact map enhancement

tools are inspired by research in image super-resolution and adopt models originally proposed for image analysis. However, there are significant differences between contact map enhancement and image super-resolution. First, we define resolution as Pixels Per Inch (PPI) in image analysis. The size of the input (i.e., low-resolution image) is smaller than the size of the output (i.e., high-resolution image) in the image super-resolution analysis. In contrast, we define resolution as the number of nucleotides in each fixed bin in the Hi-C contact map. The size of the input and the output in contact map enhancement analysis are equivalent. Both the input and the output are at the same high resolution. Moreover, image super-resolution aims at producing a realistic image that is similar to the low-resolution input; contact map enhancement aims at enhancing signals (i.e., TAD, loop, compartment, etc.) in a Hi-C contact map to facilitate downstream analysis.

Most of existing deep learning tools for contact map enhancement such as HiCPlus [75], HiCNN [76], HiCGAN [9], DeepHiC [8], and DeepLoop [59] are convolutional neural networks. During contact map enhancement, these tools split contact maps into non-overlapping blocks, enhance each block separately, and assemble the enhanced blocks into whole genome predictions. The first tool, HiCPlus [75], only contains one hidden layer and is trained from low-coverage and high-coverage contact map pairs by minimizing the mean square error (MSE) loss. Later, several improved models with more layers and residual connections are proposed [76, 59, 77, 9, 8]. As reviewed in Liu *et al.* [9], HiCPlus's training procedure leads to trained models that overly smooth contact maps. Some of the recent approaches introduced the generative adversarial training loss [9, 8] to alleviate the over-smoothness issue. However, these generative adversarial networks (GAN) tend to introduce artifacts in predictions.

In addition to deep learning tools, several conventional tools [78, 79] also exist. They treat contact map enhancement as imputation and enhance Hi-C signals by fitting a Markov Random Field or performing a random walk on a graph induced from a Hi-C contact

map.

1.7 Reference panel enabled analysis in biology

A common strategy in analyzing biological data, especially for data of insufficient coverage, is to complement the input with external data. This strategy is known as reference panel enabled analysis and plays an important role in analyzing different types of data in bioinformatics [80, 81, 82]. Even though hundreds of Hi-C contact maps have been produced, this strategy was absent in computational 3D genomics prior to the work presented in this thesis. This section presents several motivating examples of reference panel enabled analysis in other biological domains.

1.7.1 Genotype imputation

Genotype imputation [83, 80] is the most well-known reference panel enabled application in bioinformatics. It infers unobserved genotypes for a set of samples. It is an important step in genome-wide association study (GWAS) and population genetics. We achieve imputation by using existing haplotypes as a reference panel. Among a group of (unrelated) individuals, the haplotypes of one individual over a DNA segment are related to other haplotypes by being identical by descent (IBD). We can model IBD with a genealogical tree. This tree is segment specific due to recombination. Although different imputation methods vary in detail [83, 80], the general idea is to identify or utilize the underlying IBD segments between study samples and reference haplotypes and use IBD segments to impute missing alleles in study samples.

1.7.2 Genome resequencing

Resequencing is a special case of the reference panel enabled analysis as its reference panel only contains a single reference genome. Genome sequencing plays an essential role in biology research in the 21^{st} century. To know the genome sequence of a given sample, we can perform *de novo* sequencing, or resequencing [84]. The *de novo* sequencing is computational and memory intensive and expensive. When reference genomes for the species of interest exist, we prefer to perform genome resequencing. Instead of assembling the complete genome from short reads, we can align short reads to a reference genome in resequencing and readout the genome sequence for the sample of interest from the alignment profile [51]. It is sufficient to include a single sample in this reference panel as genome sequences are highly conserved within species. However, using multiple reference genomes (i.e., pangenome) can improve variant detection [85].

1.7.3 Homology modeling

Before deep learning achieved great success in protein structure modeling [86, 87], homology modeling was the most accurate approach for inferring protein structures. Protein structures are more conserved than sequences, thus, proteins with similar sequences often have similar structures [88, 89]. Homology modeling relies on this observation and predicts structures with the use of known homologous structures [88]. It begins with comparing the target sequence to a database of proteins with known structures (i.e., reference panel) to search for all proteins related to the target sequence and selecting protein template(s) from the searching result. One widely used approach to achieve this is to perform pairwise sequence-sequence comparison [90]. Both a single optimal or multiple templates can be used to perform homology modeling with different applications [88, 86]. Then, the target-template alignment is performed to detect regions well-aligned with the template [91]. With the target-template alignment, the 3D modeling step interactively builds the 3D

structures for well-aligned regions with the template's structural information. Unaligned regions are usually built by other techniques such as loop modeling and side-chain optimization. Next, model optimization might be performed to refine the initial structure. The above steps may be repeated several times before reporting and evaluating the final structure [88]. Homology modeling is less popular in recent years, but the idea of using a reference panel of homologous data to improve protein modeling is widely used in many deep learning applications [92, 93, 87].

1.7.4 Reference panel enabled framework in computational 3D genomics

The successful application of each reference panel-enabled approach in these motivating examples is due to the conservation of sequences or structures at different levels. The increase in the size of the reference panel generally leads to improve performance in all examples. 3D genome conformations are cell-type specific, but similar structures may be observed in different cells at specific loci. In addition, previous studies have demonstrated that some of the topological structures are conserved across cells [1, 94, 11]. In this thesis, we proposed a reference panel enabled framework for computational 3D genomics and developed three different applications to solve different tasks in the 3D genome study. This framework takes advantage of the facts that (i) several hundred human Hi-C contact maps are publicly available, and (ii) the vast majority of topological structures are conserved across many cell types. In this framework, we define a collection of existing Hi-C contact maps as a reference panel. In a given region, we combine the study sample and reference samples that display similar structures as the study sample to make predictions.

1.8 Attention mechanism in deep learning

This section provides a brief introduction to the attention mechanism in deep learning neural networks. Two of our reference panel enabled applications in computational 3D genomics rely in part on this mechanism to select samples that are similar to the study sample from the reference panel. We assume that the reader is already familiar with introductory notions in deep learning, which are covered in textbooks such as *Deep Learning* [95]. The attention mechanism in deep learning was initially introduced to improve encoder-decoder RNNs (recurrent neural networks) for sequence-to-sequence tasks [96]. This mechanism allows the decoder to focus on different parts of the input sequence at different decoding steps. This mechanism was soon introduced to other domains [97, 98] and became the key component of transformers [99]. Although we express the attention mechanism as a weighted sum of a set of vectors, different formulas for the attention mechanism exist in the literature. This thesis follows Vaswani *et al.* [99] to define the attention mechanism in terms of key (K), value (V), and query (Q) as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \tag{1.1}$$

This is known as the Scaled Dot-Product Attention. d_k is the dimension of query and key vectors. $\frac{1}{\sqrt{d_k}}$ is the scaling factor. For large d_k, the dot product QK^T could be very large; this scaling factor is used to stabilize the gradient computation.

Given a set of input embeddings X (i.e., embeddings for all words in a sequence-to-sequence model, etc.), and an embedding Z (i.e., the output of the previous decoding step in a sequence-to-sequence model, etc.), the Scaled Dot-Product Attention seaks to compute a transformation of X such that different Z leads to different transformations. To achieve this, we first convert X to (K,V) pair, and Z to Q. This step can be defined as network layers of any type. Then, we use the Scaled Dot-Product Attention to compute

such transformation. The Scaled Dot-Product Attention assigns more weights to values that associate with keys more similar to the query. There is no restriction on how to specify K, V, and Q. In self-attention [99], we derive K, V, and Q from the same set of embeddings. In multimodal learning [100], we can derive K, and V from one type of data (e.g., image), and Q from other modalities of data (e.g., text). We can also compute K and Q in the same way, which allows us to select instances that are similar to the query instance from the key-value pairs in a soft manner.

The attention mechanism selectively focuses on more important and relevant features from many features. Thus, it enables a network layer to have a very large receptive field to capture long range interactions [101].

In biological data analysis, many applications have been developed using the attention mechanism [102]. For example, in protein structure prediction, deep learning tools such as Alphafold [92] and models inspired by Alphafold [103, 104, 105] are transformers with the attention mechanism. In gene expression prediction, tools such as Enformer [106] are also attention models. In single cell data analysis, attention mechanism is also frequently used to perform different tasks [107, 108, 109]. In computational 3D genomics, this mechanism has already been used in applications such as Higashi [60], C.Origami [110], and GrapHiC [111].

1.9 Thesis roadmap

The remaining chapters of this thesis are organized as follows.

In Chapter 2, we propose a novel deep learning approach, RefHiC, that uses a reference panel of Hi-C samples to improve topological structure (TADs, and loops) annotation from a study sample. It is the first application of our reference panel enabled framework in computational 3D genomics. It is also the first tool that can leverage information from additional Hi-C contact maps to facilitate the analysis of a given sample

of interest. We demonstrated that the introduction of a reference panel of Hi-C samples significantly improves TAD and chromatin loop annotations from both low- and high-coverage Hi-C contact maps, with the most striking improvement observed in handling very low-coverage Hi-C data.

In Chapter 3, we explore how to apply our reference panel enabled framework to address other challenges in Hi-C data analysis. Here we focused on the contact map enhancement task. We want to balance local and global information when predicting per-pixel value of the Hi-C contact map. Our approach, called RefHiC-SR, adopts the U-Net [112] as the backbone architecture. We extensively modified the original model to incorporate a reference panel into the computational graph. We demonstrate that the introduction of a reference panel of Hi-C samples can also improve contact map enhancement significantly.

In Chapter 4, we propose RobusTAD, a non-parametric approach to detect nested TAD from a study sample while leveraging information from a reference panel of Hi-C data. We demonstrate that the introduction of a reference panel of Hi-C samples allows RobusTAD to better annotate TADs from a low-coverage Hi-C contact map.

Last, in Chapter 5, we summarize our findings and discussed future directions in computational 3D genomics.

1.10 Author contributions

This thesis includes the full text and figures of three scientific articles, each of which has been published, or is currently under review, in a peer-reviewed journal or conference proceeding. These articles are listed below in the order they appear in this thesis. I am the first author of each one.

Chapter 2

Zhang, Y., & Blanchette, M. (2022). Reference panel guided topological structure annotation of Hi-C data. Nature Communications, 13(1), 7426.

Y.Z. and M.B. conceived the study and designed models. Y.Z. implemented models, performed data analysis, and wrote the manuscript. M.B. supervised the project and edited the manuscript. All authors read and approved the final article.

Chapter 3

Zhang, Y., & Blanchette, M. (2023). Reference panel-guided super-resolution inference of Hi-C data. Bioinformatics, 39(Supplement_1), i386-i393.

Y.Z. and M.B. conceived the study. Y.Z. performed analysis. M.B. supervised the project. Y.Z. wrote the article and M.B. edited it. All authors read and approved the final article.

Chapter 4

Zhang, Y., Dali, R & Blanchette, M. (2023). RobusTAD: Reference panel based annotation of nested topologically associating domains. Submitted to Genome Biology

Y.Z. and M.B. conceived the study. R.D. conceived and implemented the initial version of the domain boundary annotation. Y.Z. implemented RobusTAD, performed data analysis, and wrote the manuscript. M.B. supervised the project and edited the manuscript. All authors read and approved the final article.

Chapter 2

Reference panel guided topological structure annotation of Hi-C data

Yanlin Zhang, and Mathieu Blanchette *

School of Computer Science, McGill University, Montréal, Québec, H3A 0E9, Canada

Preface

The Hi-C approach to mapping chromosome conformation within the nucleus of cells and identifying topological structures (i.e. loops and TADs) is rapidly gaining popularity. It has a significant impact in the fields of 3D genomics, gene regulation, and cancer. Indeed, several high-quality bioinformatics tools have been published recently, aiming to annotate chromatin loops and/or TADs from a given HiC data set. Yet, the precise and reliable identification of these spatial patterns at high resolution remains inadequately resolved. Existing tools only take the sample of interest as input, so their performance is limited by the sequencing coverage of the study sample.

In this chapter, we introduce RefHiC, an attention-based deep learning framework that leverages a reference panel of Hi-C datasets to annotate topological structure from a given study sample. To our knowledge, RefHiC is the first tool to take advantage of

the paradigm of reference-panel aided analysis, which has proved very powerful in a variety of other types of genomics data analyses. Compared to existing approaches, we show RefHiC provides greatly improved accuracy for loop and TAD annotation across different cell types and sequencing depths.

The rest of this chapter is the entire text from the following article:
Zhang, Y., & Blanchette, M. (2022). Reference panel guided topological structure annotation of Hi-C data. Nature Communications, 13(1), 7426.

2.1 Abstract

Accurately annotating topological structures (e.g. loops and topologically associating domains) from Hi-C data is critical for understanding the role of 3D genome organization in gene regulation. This is a challenging task, especially at high resolution, in part due to the limited sequencing coverage of Hi-C data. Current approaches focus on the analysis of individual Hi-C data sets of interest, without taking advantage of the facts that (i) several hundred Hi-C contact maps are publicly available, and (ii) the vast majority of topological structures are conserved across multiple cell types. Here, we present RefHiC, an attention-based deep learning framework that uses a reference panel of Hi-C datasets to facilitate topological structure annotation from a given study sample. We compare RefHiC against tools that do not use reference samples and find that RefHiC outperforms other programs at both topological associating domain and loop annotation across different cell types, species, and sequencing depths.

2.2 Introduction

Chromosome conformation capture assays such as Hi-C [3], and micro-C [5] have been developed to measure the spatial proximity between DNA fragments in genomes as average pairwise contact frequency in cell populations. These approaches have revealed a hierarchical spatial organization of topological structures of the genome inside nuclei. Among them, topological associating domains (TADs) are kilo- to mega-scale regions with strong interactions between DNA fragments within the same domain and weaker interactions across domains [6]. Loops bring into contact distant loci such as promoters and enhancers [4]. These topological structures are dynamic both within cells [113] and during cellular differentiation [1]. They are essential components of gene regulation.

While Hi-C and its variants remain the most popular approaches to map chromatin contacts on a genome-wide scale, the analysis of the data they produce is challenging, in large part due to the moderate sequencing depth (typically 200-500 Million valid read pairs) compared to the size of the contact frequency matrices that need to be estimated. Numerous TAD annotation tools exist that rely on various statistical significance tests [63, 72, 114]. This includes the popular Insulation score (IS) [69], a widely used approach for TAD boundary detection, and more robust variants such as RobusTAD [115]. Still, the performance of all of these approaches is relatively poor, especially at low coverage, due to stochastic noise and biases [63]. Loop detection is even more challenging [4, 17] due to their small size in contact maps. Fit-Hi-C [61] and HiC-DC [62] fit a global model to estimate the background distribution of the contact frequency and identify statistically significant contact pairs by comparing observed values to expected values from the fitted model. These global enrichment approaches evaluate each contact pair independently without modelling neighboring patterns and identify loop clusters instead of discrete loops. In contrast, HiCCUPS [4] compares each contact pair to surrounding regions and identifies locally enriched contact pairs as loops. It requires users to set several sequencing depth sensitive parameters and can only detect loops that satisfy the user defined filtering criteria. Both loop and TAD predictions have been shown to benefit from prior smoothing of Hi-C matrices, e.g. using HIFI [78].

Recent approaches tackle topological structure annotations using computer vision and machine learning techniques. For instance, Mustache [18] treats chromatin loop recognition as a blob-shaped object detection problem. Chromosight [16] employs expert-designed templates to represent each type of topological structures. These generic pattern-based approaches work well on data with sufficient contact pairs but underperform at low sequencing depth. In contrast, Peakachu [17] is a supervised learning approach trained to recognize loops using data from orthogonal experiments as target values.

Many approaches have been introduced to address the issue of insufficient sequencing depth. Grinch [7] proposed a graph-regularized non-negative matrix factorization algorithm to smooth sparse Hi-C contact map while detecting TADs. DeepLoop [59] identifies significant interactions from sparse Hi-C contact maps by denoising and enhancing loop signals with a neural network. Higashi [60], a single cell Hi-C data analysis tool, represents a cohort of scHi-C data as a hypergraph, learns to predict missing hyperedge to impute missing interaction, and then performs structural annotation on imputation.

A common strategy in analyzing biological data is to complement data about the sample of interest with data of the same type obtained previously for other samples. This strategy has proven effective for genotype imputation [80] and phasing [81], as well as protein structure prediction [116], among others. Even though hundreds of Hi-C experiments have been conducted, they have never been analyzed jointly for topological structure annotation. Here we introduce RefHiC, a reference panel informed deep learning approach for topological structure (loop and TAD) annotation from Hi-C data. RefHiC uses a reference panel that contains high-quality Hi-C data of different cell types. For each potential contact in the study sample, it uses an attention mechanism [117] that determines which of the reference samples are most relevant, and then makes a prediction based on the combined study sample and attention-weighted reference samples. We demonstrate that RefHiC enables significant accuracy and robustness gains, across cell types, species, and coverage levels.

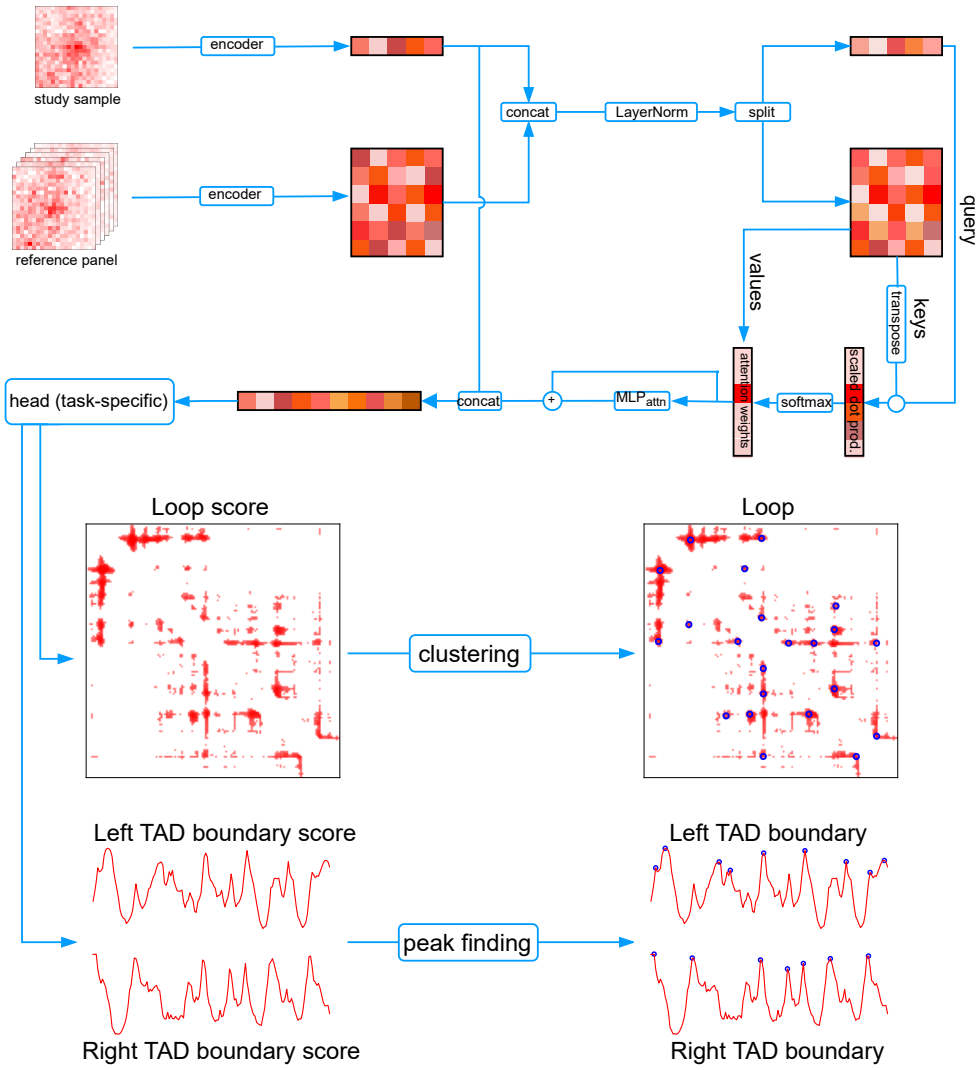


Figure 2.1: **RefHiC architecture.** Overview of the RefHiC neural network for loop and TAD boundary scoring, followed by clustering or peak finding algorithm for discrete loop and TAD predictions (shown as blue circles).

2.3 Results

2.3.1 Overview of RefHiC

RefHiC takes as input a Hi-C contact map for a study sample and a reference panel of Hi-C contact maps (provided with the tool). It produces highly reliable loop or TAD boundary annotations for the study sample. RefHiC is based on two components (Fig. 2.1 and

Methods): (i) a neural network predicts loop (resp. TAD boundary) scores for every candidate pair (resp. locus) based on the local contact sub-matrix, combining information from the study sample and the reference panel; (ii) a task-specific component selects one representative loop/TAD boundary from each high-scoring cluster. For human, the reference panel contains 30 uniformly processed Hi-C contact maps, each with at least 350 million contact pairs (Table 2.1). For mouse, it consists of 20 such maps (Table 2.2). Normalization of reference Hi-C samples is unnecessary as the network automatically learns to handle batch effect and coverage differences from the training data.

To obtain a loop or TAD boundary score for bin pair (i, j) (with $i \neq j$ for loops and $i = j$ for TAD boundaries), an encoder projects the sub-matrices centered at (i, j) in both the study sample and reference panel to low-dimensional embeddings. An attention module [117] computes a combined representation of all reference samples as a weighted sum of their embeddings, with weights based on their local similarity to the study sample's embedding. Finally, a multi-layer perceptron predictor computes loop or TAD boundary score from the concatenation of the study sample's embedding and the attention output. The process is repeated for every pair (i, j) to obtain a scoring matrix (resp. vector), from which discrete loop (resp. TAD boundary) predictions are extracted.

Training RefHiC (i.e. choosing the weights of the encoder, fully connected layers in attention module, and head) is achieved using a variety of downsampled versions of a high-coverage Hi-C data set for GM12878 [4]. Following Salameh et al. [17], we used as prediction targets a set of long-range loops identified at 5 kb resolution by either ChIA-PET on CTCF [118] or RAD21 [119], as well as by HiCHIP on SMC1 [37] or H3K27ac [120]. Using multiple experimental data sets ensures a broad coverage of various types of loops.

Importantly, although RefHiC is trained on GM12878 data, the model learned is not cell-type specific, and we will demonstrate in later sections that the same model can be used to annotate structures in many other cell types without retraining and with similar

accuracy. The same trained model can also be used to make predictions on mouse Hi-C data, based on our reference panel for that species.

In our experiments, we used human chromosomes 11 and 12 for validation, chromosomes 15-17 for testing, and the rest of the autosomes for training. To prevent potential data leakage, all results reported here pertain only to the three test chromosomes.

2.3.2 RefHiC accurately detects chromatin loops from Hi-C contact maps

We first assessed the loop prediction accuracy of RefHiC on a downsampled Hi-C data set (500M valid read pairs) for human GM12878 cells [4]. We then applied Chromosight [16], Peakachu [17], Mustache [18], and HiCCUPS [4] to annotate loops from the same data with default parameters. For all tools, we set the same 5% FDR cutoff whenever possible.

The sets of predicted loops are quite different among tools, with RefHiC making the largest number of unique predictions (Fig. 2.2a and Fig. 2.7). Aggregate peak analysis (Fig. 2.2b) shows that loops detected by Chromosight, RefHiC, and Mustache had a more diffuse loop center compared to those identified by Peakachu and HiCCUPS. Finally, the distribution of distances between loop anchors predicted by RefHiC and Mustache most closely resembled that of ChIA-PET/HiCHIP-supported loops (Fig. 2.2c), whereas Peakachu, HiCCUPS, and Chromosight predicted more short-range interactions.

We then evaluated predicted loops by comparing them to loops revealed by loop-targeting experimental data, allowing up to 5 kb shift. To facilitate interpretation, we considered the top 1700 predictions from each tool by adjusting the FDR or loop score cutoff. Fig. 2.2d-f, and Fig. 2.8a show that RefHiC produced 1250 CTCF-supported loops, 784 RAD21-supported loops, 588 SMC1-supported loops, and 213 H3K27ac-supported loops. In contrast, other tools yielded 20-52% fewer validated loops. Comparison against DeepLoop [59] reveal similar numbers (Section 2.11.1 and Fig. 2.9). Finally, to delineate the impact of using a reference panel, we evaluated a version of RefHiC that operates

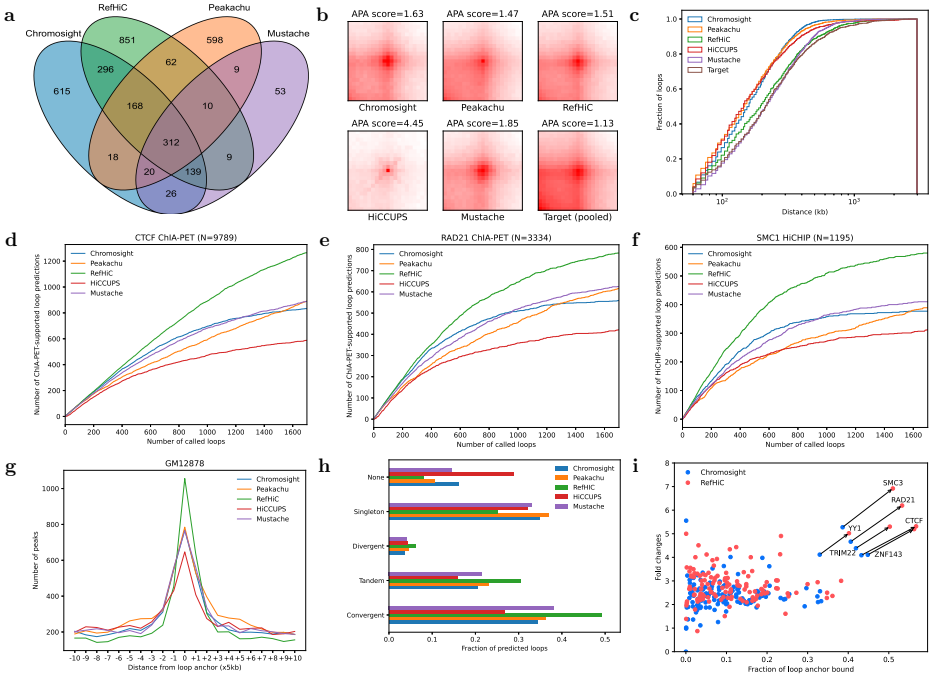


Figure 2.2: **Comparison of RefHiC, Chromosight, Peakachu, HiCCUPS, and Mustache on GM12878 Hi-C data (500M valid read pairs). a,** Venn diagram of loops predicted by different tools. **b,** Aggregate peak analysis profiles for target (ChIA-PET and HiCHIP identified) and annotated loops. **c,** Cumulative distance distributions of predicted loops. RefHiC's predicted loop distance distribution closely resembles that of ChIA-PET/HiCHIP-supported loops (target). **d-f,** Number of ChIA-PET/HiCHIP-supported loop predictions, among the top 1700 predictions made by RefHiC and other tools, for test chromosomes chr15, chr16, and chr17, compared against CTCF ChIA-PET (d), RAD21 ChIA-PET (e), and SMC1 HiCHIP (f). RefHiC's loop predictions matches those experimental data better than predictions made by other tools on test chromosomes. **g,** Occupancy of ChIP-seq identified CTCF binding site as a function of distance to predicted loop anchors. **h,** Orientation of CTCF motifs at predicted loops. **i,** Transcription factor (TF) occupancy at predicted loops (RefHiC and Chromosight only). Each dot is a TF or histone modification (based on 133 ENCODE ChIP-seq data sets for GM12878), whose x-coordinate is the fraction of loop anchors containing a binding/modification site and the y-axis is the fold enrichment against genome-wide frequency. Most TFs are more strongly enriched at RefHiC loop predictions than at Chromosight loop predictions.

exclusively based on the study sample (Section 2.11.2); while this reference-free predictor obtains state-of-the-art performance (or better), it is far from the reference panel based RefHiC (Fig. 2.10). As shown in Fig. 2.2g, predicted loop anchors detected by RefHiC were strongly enriched with the CTCF binding motifs. TAD-forming loops have been previously shown to be associated with the presence of convergent CTCF binding sites

at loop anchors [113]. Indeed, 50% of RefHiC's loop predictions are associated with such pairs of sites; significantly more than for other tools (Fig. 2.2h).

Among loops detected by each tool, 46% of RefHiC, 39% of Chromosight, 50% of Peakachu, and 9% of Mustache were not detected by other tools (Fig. 2.2a). Fig. 2.11 shows that RefHiC-specific predictions are not only more numerous but also more accurate when evaluated against CTCF/RAD21 ChIA-PET, and SMC1 HiCHIP data, though slightly less accurate than Chromosight and Peakachu on H3K27ac HiCHiP data. Chromosight and Peakachu were slightly better than RefHiC when being evaluated against H3K27ac HiCHiP data. A deeper analysis of the properties of loops predicted by each tool is presented in Section 2.11.3 and Figs. 2.12-2.14.

To further study the properties of loops predicted by each tool, we performed transcription factor (TF) and histone modification enrichment analysis around loop anchors. Fig. 2.2i and Fig. 2.8b,c show enrichment for known loop-mediating proteins (SMC3, RAD21, YY1, TRIM22, CTCF, and ZNF143) was strongest for RefHiC compared to Chromosight and Peakachu, and comparable to Mustache.

Combined, these results demonstrate the overall superior prediction accuracy of RefHiC on GM12878 data (500M read pairs) compared to other approaches.

2.3.3 RefHiC performs well across cell types and species

Although RefHiC is trained on human GM12878 data, we demonstrate here that the same trained model performs well across other human and mouse cell types. We applied RefHiC and other tools (5% FDR) to Hi-C data from human K562, IMR90 [4], and cohesin-depleted HCT-116 [121] cell lines (test chromosomes 15-17 only), as well as mouse embryonic stem cells (mESC) [122] (all chromosomes). Since the IMR90 data set has twice the sequencing coverage of the K562 data set, all tools identified more loops in the former, with Chromosight and RefHiC making the largest number of predictions (Fig. 2.3a).

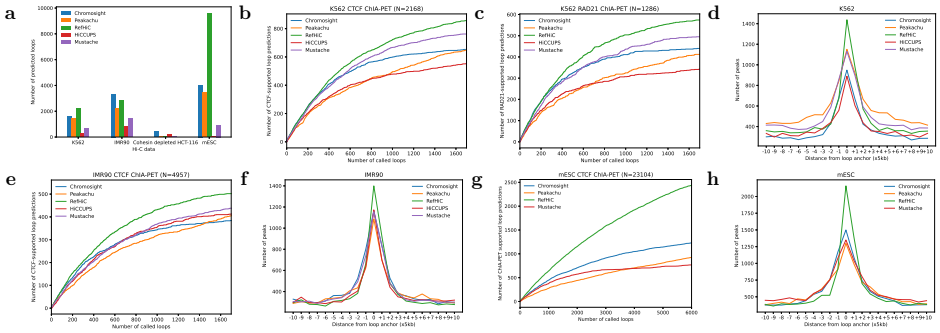


Figure 2.3: **Loop detection in Hi-C data from human K562, IMR90, and cohesin-depleted HCT-116 cells, as well as mouse ESC. a,** Number of loops identified in Hi-C datasets obtained in each data set (human data: test chromosomes 15-17 only; mouse data: all autosomes). Note that in HCT-116, one would not expect any cohesin-mediated loops. **b,c,** Number of ChIA-PET/HiCHIP-supported loop predictions, among the top predictions made by RefHiC and other tools on K562 Hi-C data, for test chromosomes chr15-17, compared against CTCF (b) and RAD21 ChIA-PET (c) data. **d,** Occupancy of ChIP-seq identified CTCF binding sites in K562 cells as a function of distance to predicted loop anchors. **e,f** Same as (b,d), but for data obtained in IMR90 cells. **g,h,** Same as (b,d), but for data obtained in mESC (all autosomes).

However, RefHiC is notably more robust to sequencing depth, with a decrease of only 22% from IMR90 to K562, compared to 34-66% for other tools.

Cohesin-depleted HCT-116 cells are expected not to contain any loop. Indeed, RefHiC, Peakachu, and Mustache made fewer than 24 loop predictions on this data, whereas Chromosight and HiCCUPS had many more likely false positives.

For the mESC data, which contains only 124M valid read pairs, we used the same RefHiC model trained from human GM12878, but with a mouse reference panel made of 20 mouse Hi-C data sets (Table 2.2). Applied to the complete set of autosomes, RefHiC identified more than twice as many loops as any other tool, indicating that it is much more sensitive than other tools on low-coverage data.

We then assessed these tools' accuracy using loops revealed by orthogonal experiments. As before, we included the top 1700 predictions on test chromosomes from each tool by adjusting FDR or loop score cutoff. For K562 data, as shown in Fig. 2.3b,c, RefHiC outperformed other tools as it identified more CTCF- and RAD21-supported loops. The

pileup analysis of CTCF binding sites around predicted loop anchors (Fig. 2.3d) shows occupancy 25% higher for RefHiC than for Peakachu and Mustache, and 51% higher than for Chromosight and HiCCUPS. Similar results are obtained on IMR90 data (Fig. 2.3e,f), although its very high coverage enables competing approaches to get somewhat closer to RefHiC's performance. For the mESC data, Fig. 2.3g,h indicates that RefHiC outperformed alternative tools significantly as it detected as twice as many CTCF-supported loops as alternative tools and its loop anchor predictions are more strongly enriched for CTCF binding sites than other tools. To further study the ability of RefHiC to identify loops in samples that are very different from those present in its reference panel, we generated reference panels excluding samples that are closely related to study sample GM12878, or even entirely unrelated (e.g. from the incorrect chromosome). Section 2.11.4 and Fig. 2.20 show that RefHiC performs comparably or better than other tools even under this less favourable scenario. Together, the results show that RefHiC achieves superior performance across both human and mouse cell types.

2.3.4 RefHiC is robust to sequencing depths

To benchmark RefHiC's ability to detect loops from Hi-C data at different sequencing depths, we produced downsampled versions a high-coverage GM12878 Hi-C combined contact map [4] and applied different loop prediction tools (default parameters; FDR cutoff 0.05 when possible). Although lower sequencing depths led to fewer loop predictions for all tools (Fig. 2.4a), RefHiC was most robust to sequencing depths. For example, RefHiC identified 731 loops from low coverage Hi-C data (62.5M contact pairs) – 32% of the results obtained from 2,000M contact pairs. In contrast, other tools are largely unable to make sensitive loop predictions at this low sequence depth. Fig. 2.4b shows that RefHiC detected highly concordant sets of loops across sequencing depths: ∼85% of

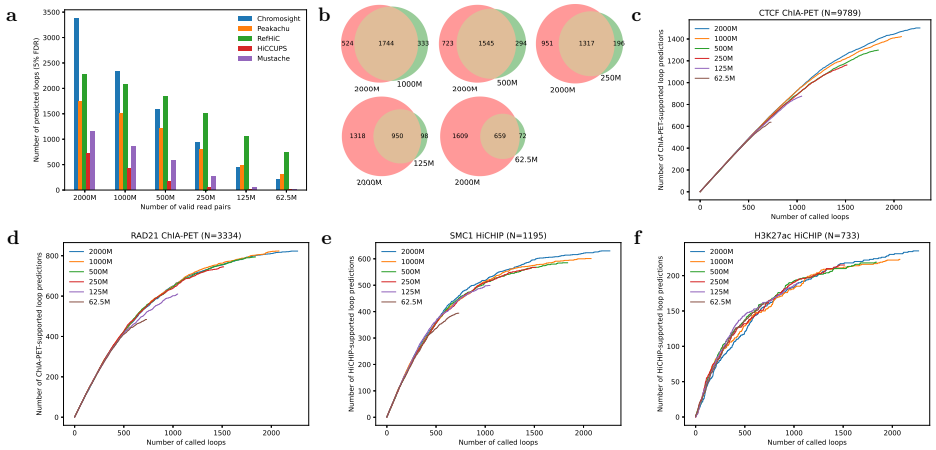


Figure 2.4: **Detection of loops at lower sequencing depths. a,** Number of loops predicted by different tools at 5% FDR, for decreasing number of valid intra-chromosomal read pairs. **b,** Venn diagram of loops predicted from Hi-C data of different sequencing depths by RefHiC. **c-f,** Number of RefHiC loop predictions supported by experimental GM12878 ChIA-PET/HiCHIP data (test chromosomes chr15-17) at different levels of sequencing coverage: CTCF ChIA-PET (c), RAD21 ChIA-PET (d), SMC1 HiCHIP (e), H3K27ac HiCHIP (f).

loops annotated from Hi-C data containing 1,000M, 500M, and 250M contact pairs overlapped those annotated from the 2,000M contact pairs data set. This percentage was even higher (90%) on low-depth Hi-C data (i.e. 125M, and 62.5M contact pairs). This shows that not only is RefHiC capable of detecting a good number of loops in low coverage data, but it also does not introduce significantly more false positives. Fig. 2.4c-f confirm that RefHiC predictions on low depth data sets maintain a very high level of accuracy when evaluated against loops mediated by CTCF, RAD21, SMC1 and H3K27ac. In short, this means that predictions made on low coverage data are nearly as specific as those made on the full data, but are simply less sensitive. At all sequencing depths, RefHiC achieved higher accuracy than alternative tools (Fig. 2.10). This superior robustness, accuracy, and reliability is attributable to the use of reference panel.

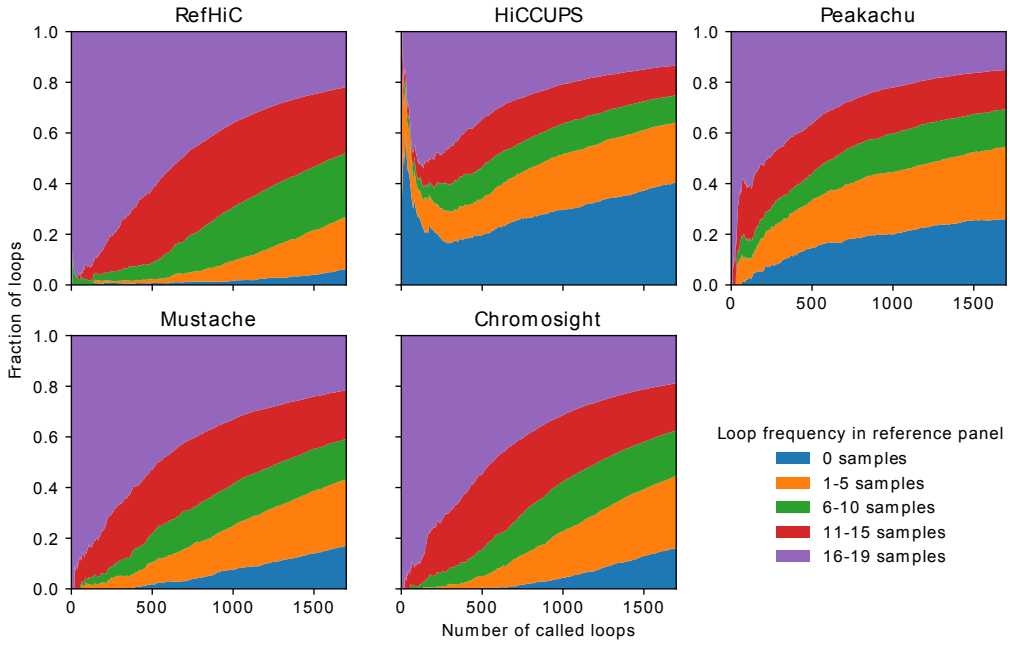


Figure 2.5: **Comparison of tools' ability to identify rare and common loops.** Loop frequency within the reference panel is assessed based on Mustache and Chromosight's predictions on individual reference Hi-C data sets.

2.3.5 RefHiC identifies both rare and common loops

Since RefHiC uses a reference panel to complement data from the study sample, one may expect that it performs best on common loops (i.e. those present in a large number of cell types from our reference panel). To determine the prevalence of each loop, we ran Mustache and Chromosight on our reference samples and merged their predictions (allowing a 2-bin shift; the two tools failed on 10 of the 29 samples as one or both detected less than 10 loops, leaving a total of 19 samples annotated). We then assessed the frequency at which loops predicted by RefHiC on GM12878 were found in the 19 reference samples. The distribution of reference panel frequencies among loops predicted by RefHiC resembled that of Peakachu, Mustache, and Chromosight (Fig. 2.5). For all these tools, the majority of highest-scoring loops were found to be present across nearly all samples, suggesting that constitutive, non-cell-type specific loops have features that make them easily predictable. Still, more than 20% of loops predicted by RefHiC are rare (found in

at most 5 of the panel data sets), and 4% are specific to GM12878, demonstrating that the use of a reference panel does not strongly bias the results in favor of common loops. Still, those proportions are slightly lower than those obtained with the three other tools, which could be explained by a combination of a weak bias toward common loops for RefHiC, and an increased false-positive rate (which usually will appear as cell-type specific loops) for the other tools. Indeed, the number of GM12878-specific loop predictions that are supported by experimental data is actually comparable across tools (Fig. 2.15). Peakachu identified more cell-type specific validated loops than other tools, but with a lower specificity than RefHiC. Among loops found to occur at least once in the panel, RefHiC gets more ChIA-PET/HiCHIP-supported predictions than alternative tools (Fig. 2.15e,g-t), except that Peakachu identified more H3K27ac HiCHIP supported loops (Fig. 2.15f). Finally, loops predicted by HiCCUPS were very different, containing more of what looks like GM12878-specific loops (i.e. loops absent from the reference panel), many of which are likely false-positives.

2.3.6 RefHiC accurately detects TADs

RefHiC is a versatile framework for topological structure annotation. Here we show that RefHiC can detect TADs once trained using as target values RobusTAD TAD boundary scores obtained on a high-coverage HiC data set (see Methods). We first compared RefHiC's performance on downsampled versions of a GM12878 Hi-C contact map to that of two established TAD boundary predictors (RobusTAD [115] and insulation score [69]). Fig. 2.6a,b and Fig. 2.21 show that at 500M valid read pairs, RefHiC and RobusTAD succeed at identifying a similar total number of CTCF-supported TAD boundaries, although RefHiC's specificity is much higher considering that its total number of predictions (at 5% FDR) is approximately 40% less than with RobusTAD. Figs. 2.16 and 2.17 show that at

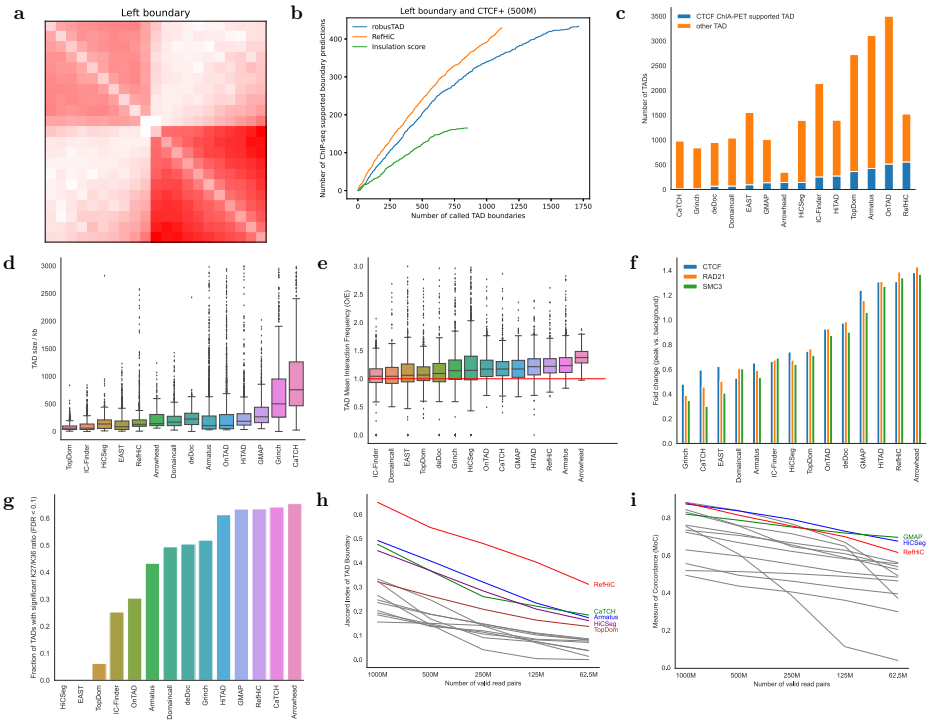


Figure 2.6: **Detection of TAD boundaries and TADs on GM12878 Hi-C data. a,** TAD boundary pileups for left boundaries predicted by RefHiC. **b,** Number of predicted left TAD boundaries supported by ChIP-seq identified CTCF binding sites (positive strand only), for RefHiC, RobusTAD, and Insulation score. **c-i,** Benchmarking RefHiC against 13 other TAD callers on TAD annotation. **c,** Number of TADs predicted by different tools, and proportion of predicted TAD boundary pairs that are supported by CTCF ChIA-PET data. Size (**d**), and mean interaction frequency (observed/expected) (**e**) of TAD predictions. The number of TADs used to generated box plots are provided in **c**. In each box, the upper edge, central line and lower edge represent the 75th, 50th and 25th percentile, respectively. Upper whiskers represent 75th percentile+1.5× interquartile range (IQR), lower whiskers represent minimum values, and dots represent samples above the 75th percentile+1.5×IQR. **f,** Enrichment of CTCF, RAD21, and SMC3 peak signals at TAD boundaries. **g,** Fraction of TADs predicted by each caller with a significant (high or low) H3K27me3/H3K36me3 log10-ratio (FDR < 0.1). Jaccard index of predicted TAD boundaries (**h**) and Concordance between TADs (**i**) predicted on high-coverage GM12878 data (2B valid read pairs) compared to those predicted on downsampled Hi-C data. Note: **a-g** are based on a downsampled GM12878 Hi-C data set that containing 500M valid read pairs).

very low coverage (125M and 62.5M valid read pairs), RefHiC achieves both higher sensitivity and specificity. In all cases, both RefHiC and RobusTAD outperform Insulation score.

We then benchmarked RefHiC against 13 TAD callers (TopDom [65], Armatus [66],

deDoc [67], Arrowhead [4], HiTAD [123], EAST [124], OnTAD [71], CaTCH [68], Grinch [7], Domaincall [11], GMAP [125], HiCSeg [126], and IC-Finder [127]) on test chromosomes 15-17. Because there is no universally accepted gold-standard TAD annotation to compared against, we evaluated various aspects of the predictions made by the different tools. We first compared the number and size of TADs identified by each tool (Fig. 2.6c,d). Although the number varies from 347 to 3,499, most tools (including RefHiC) identified 1,000-1,500 TADs, with median TAD size around 130 kb for RefHiC. TADs are domains with high levels of internal interaction, so one measure of TAD annotation quality is the average observed/expected ratio within TADs (Fig. 2.6e). RefHiC's TAD predictions are among the densest in interaction frequencies. We then calculated the enrichment for ChIP-Seq signals of structural proteins known to be associated with TAD boundaries (i.e. CTCF, RAD21, and SMC3)[72] at predicted TAD boundaries and nearby (Fig. 2.6f and Fig. 2.22). Based on this metric, RefHiC is only outperformed by Arrowhead, which identifies 3 times fewer TADs. Histone marks usually correlate with regulatory activity, and most TADs are typically enriched for either activation (H3K36me3) or repression (H3K27me3) marks, but rarely both. We calculated the ratio between H3K27me3 and H3K36me3 within each TAD prediction and counted the fraction of TAD predictions where this ratio was particularly large or small (see Methods). RefHiC is among the top three TAD callers under this metric (Fig. 2.6g), only bested by tools that predict a much smaller number of TADs (Arrowhead and CaTCH). Many TADs in mammalian genomes exhibit a strong contact between their left and right boundary loci, forming a visible TAD corner; they are often referred to as loop domains. We compared predicted TAD corners against CTCF ChIA-PET data (Fig. 2.6c). RefHiC is the best-performing tool, with 556 (36.5%) TADs corners supported by CTCF ChIA-PET data (allowing 1-bin mismatch). Finally, we evaluated the prediction reproducibility at both the boundary and full TAD levels when TAD callers are applied to Hi-C data containing different numbers of valid

read pairs. RefHiC proved much more robust than other tools at the TAD boundary prediction task (Fig. 2.6h) and better than most (but slightly worse than GMAP and HiCSeg) at the full TAD prediction task (Fig. 2.6i). This last observation is likely is due to the fact that pairing predicted TAD boundaries to obtain full TAD predictions is a step that does not currently take advantage of RefHiC's reference panel.

2.4 Discussion

Here we present RefHiC, a deep learning framework that utilizes a reference panel to guide the annotation of topological structure from a given study sample. In contrast, existing topological structure detection algorithms are study-sample based (i.e. reference-free) detectors and hence their ability to reliably detect topological structures from typical sequencing depth Hi-C data is limited. Our extensive evaluation demonstrated that RefHiC outperforms existing tools for both TAD and loop annotations, in data sets ranging from very high to very low sequencing coverage, with the most striking improvements observed in the latter case. This benefit comes at little cost in terms of RefHiC's ability to identify cell-type specific loops.

Importantly, although RefHiC is a machine-learning based model trained primarily on GM12878 Hi-C data, the same trained model is effective on different cell types, at different levels of coverage, and across human and mouse. Indeed, all results reported here for loop prediction were obtained with the same trained model, which is available in our GitHub repository. This model can be used for mammalian Hi-C data analyses without retraining. Retraining would only be needed if other types of structures are sought, or if the experimental protocol used to generate the study sampled differed significantly from the standard in situ Hi-C protocol. In such cases, RefHiC would require retraining but would still be able to take advantage of our Hi-C reference panel, i.e. the reference panel does not need to be of the same type as the study sample. However, applying RefHiC to

Hi-C data obtained from other species might be more challenging, due to the lacking of reference samples, than reference-free alternative tools.

Our method has several methodological contributions. The key innovation of RefHiC is the introduction of a Hi-C reference panel. Our attention-based framework enables RefHiC to identify and take advantage of the reference samples that exhibit similar local structures as the study sample at the locus pair of interest. This approach based on local similarity significantly outperformed an analogous approach based on global similarity (Section 2.11.5). Besides, we introduced contrastive pretraining [128] and data augmentation by downsampling Hi-C contact map techniques to train a single model capable of handling Hi-C data of different sequencing depths. We believe this training procedure can improve many machine learning applications for Hi-C data analysis [75, 17, 76, 129].

In principle, reference-based approaches such as RefHiC have the potential of becoming increasingly accurate as larger compendia of high-quality Hi-C data sets obtained from diverse cell types become available and get included in the panel. However, our analyses (Fig. 2.18) suggest that limited benefits for the analysis of GM12878 data are obtained beyond a panel consisting of 10 high-coverage data sets. However, we expect that this observation is dependent on the origin of the study sample of interest, and RefHiC's performance on study samples that are divergent from the cell types represented in the panel would certainly benefit from additional, closer reference samples.

RefHiC could potentially be improved in several directions. Expanding the reference panel could improve prediction accuracy, but this is challenging memory-wise with our current implementation. In addition to further software optimization, we will develop high-diversity panels that will aim to capture most of the structural diversity through a moderate number of Hi-C data sets. In addition, we can potentially extend RefHiC to analyze data at an even higher resolution (e.g. 1 kb), although this too would require optimizing data handling to limit the memory footprint and IO time.

Across the different sub-fields of data-driven biology, major leaps forward have taken place when researchers have developed approaches that enabled the analysis of one data set to benefit from the availability of other published data sets. RefHiC is an approach to enable this type of reference-panel based analysis of 3D genomics data. It enables high-accuracy annotation of Hi-C data sets even at moderate sequencing coverage, and boosts the accuracy of the analysis of even the most deeply sequenced data sets. RefHiC and other approaches of its kind have the potential to become an essential method for topological structure annotation from Hi-C contact maps, paving the way to further our understanding of 3D genome organization and functional implications. With the increasing availability of high-quality Hi-C data sets from diverse cell types, we anticipate that the power of RefHiC will further develop.

2.5 Methods

2.5.1 RefHiC model architecture

The RefHiC network consists of three parts (see Fig. 2.1 and Fig. 2.19): (i) an encoder, (ii) an attention module, and (iii) a task-specific head. The encoder takes an input of dimension $(2 \times w + 1) \times (2 \times w + 1) \times 2$, where w is the window size ($w = 10$ in loop annotation, $w = 20$ in TAD boundary annotation) and projects the input to a d-dimensional embedding ($d = 64$). It is built with one ReLU-activated convolution layer with kernel size three and two ReLU-activated fully connected layers with d hidden units in each layer. In forward pass, the encoder an computes embedding $\mathbf{e_s} \in \mathbb{R}^{1 \times d}$ for the study sample and $[\mathbf{e_1}, \mathbf{e_2}, \ldots, \mathbf{e_n}] \in \mathbb{R}^{n \times d}$ for the n reference samples. The attention module takes as input the embeddings for both the study and reference samples and outputs $\mathbf{a} \in \mathbb{R}^{1 \times d}$ that

contains topological structural information learned from the reference panel. The layer-normalized study sample's embedding is used as query ($\mathbf{Q} \in \mathbb{R}^{1\times d}$) against the layer-normalized reference samples' embeddings, which are used as both keys ($\mathbf{K} \in \mathbb{R}^{n\times d}$) and values ($\mathbf{V} \in \mathbb{R}^{n\times d}$). We define the attention weights $\mathbf{ff}$ =softmax($\mathbf{QK}^T$) $\in \mathbb{R}^{1\times n}$, where α_j represents the relative amount of attention paid to sample j in our reference panel when analyzing the study sample. The attention output $\mathbf{a}$ is computed as,

$$\mathbf{a} = \text{softmax}(\mathbf{QK}^T)\mathbf{V} + \text{MLP}_{\text{attn}}(\text{softmax}(\mathbf{QK}^T)\mathbf{V}) \tag{2.1}$$

where MLP_{attn} has ReLU-activated fully connected layers with two hidden layers, and each layer contains d hidden units. Finally, the head is a task-specific predictor (either for loop or for TAD boundary prediction) with 2 hidden layers containing $2d$ and d hidden units. It has one sigmoid-activated output unit for loop prediction and two tanh-activated output units for TAD boundary prediction. Both tasks use the concatenation of the study sample's embedding $\mathbf{e_s}$ and attention output $\mathbf{a}$ as input. For loop prediction, it outputs a value indicating loop probability. For TAD boundary prediction, it outputs two values corresponding to left and right boundary scores. To make predictions, we apply RefHiC to each entry in the upper triangular contact matrix to compute loop probabilities, and each entry on the main diagonal to compute TAD boundary scores.

2.5.2 Detecting loops by density-based clustering

Applied to the window centered around each bin pair (i, j) (a.k.a pixel), RefHiC produces a loop probability score $L(i, j)$. Pixels where $L(i, j) > 0.5$ are called *loop candidates*. Candidate (i, j) is called an *isolated* prediction if there are less than six candidates within a 5-bin by 5-bin square centered at (i, j). We excluded all isolated predictions as they are likely to be false positives. We then grouped the remaining candidates into clusters using a density-based clustering algorithm [130]. We first computed local density $\rho(i, j)$ for

candidate (i, j) by convolving scores with a Gaussian kernel over candidates (i', j') where $\min\{|i' - i|, |j' - j|\} \leq 5$. We then calculated $\delta(i, j)$ as the minimum Chebyshev distance between candidate (i, j) and any candidate (i', j') with higher density. For candidates (i, j) with the highest local density, we defined it as $\delta(i, j) = \delta_{\max}$, where $\delta_{\max}$ is a large constant. We used a KD-tree data structure to facilitate the fast computation of $\delta(i, j)$. We discarded candidates with δ smaller than five since they were more likely to be redundant annotations. Among the remaining candidates, we then used a target-decoy search approach to find cluster centroids by identifying candidates with high local density. Given a study sample Hi-C contact map, we created a decoy contact map by permuting interaction frequencies diagonal-wise, applied RefHiC to detect loop candidates in the decoy contact map, and calculated ρ and δ for loop candidates in the decoy contact map. We then sorted candidates predicted from the study and decoy samples based on local density (ρ) and selected the top candidates while keeping the false discovery rate (FDR) at $\alpha = 0.05$. Last, we assigned the remaining candidates to their nearest clusters and chose as a loop the highest local density candidate in each cluster.

2.5.3 Detecting TAD boundary by peak finding

RefHiC annotates right and left TAD boundaries separately. To annotate discrete right boundaries, we represented right boundary scores produced by RefHiC as sequential data and annotated boundaries by finding peaks using the find_peak function in SciPy [131]. When selecting TADs, we used the target-decoy search approach to find the height (i.e. score cutoff) parameter in find_peak. We also set the minimum distance between peaks to 5 to exclude locally redundant TAD boundaries. We applied the same steps to annotate left boundaries from left boundary scores. Like TopDom and GMAP, we annotate a region starting from a left boundary l_i and ending at a downstream right boundary r_j (r_j is on the left of or identical to l_{i+1}) as a TAD. We allow a left boundary pairs with multiple right

boundaries. This produces nested TADs.

2.5.4 Feature vector and training data

RefHiC's feature vector is defined as a tensor with two channels (observed interaction frequency and observed/expected ratio) in the shape of $2 \times (2 \times w + 1) \times (2 \times w + 1)$, corresponding to the window of size $2w + 1$ centered at the pixel of interest. w is a hyper-parameter set to $w = 10$ for loop annotation and $w = 20$ for TAD boundary annotation at 5kb resolution. We trained RefHiC with Hi-C contact maps downsampled from the combined GM12878 Hi-C contact map [4].

For loop annotation, following Salameh et al. [17], we used as gold-standard (i.e. positive training cases) a set of long-range loops identified by either ChIA-PET on CTCF [118] or RAD21 [119], as well as by HiCHIP on SMC1 [37] or H3K27ac [120]. Using multiple experimental data sets ensures a broad coverage of various types of loops. We binned interactions at 5kb resolution and removed duplicates and any contact pairs with a distance shorter than 50 kb or longer than 3 Mb, resulting in 74,855 interactions used as positive cases for loop annotation. We created the negative set by selecting non-loop pairs of different types: (i) We randomly drew 50,000 contact pairs, excluding contact pairs with Chromosight scores greater than 0, while preserving the distance distribution between positive loop anchors, (ii) to increase the representation of long range negative examples, we randomly selected 10,000 long range ($1 \sim 3Mb$) pairs, most cases of (i) and (ii) are non-loop pairs in all samples. The negative set does not contain enough data representing pairs of loci that do not form a loop in the study sample, but do in some of the reference samples. Thus, we select examples (iii) from pairs identified as loops in one or more reference samples: we applied Chromosight and Mustache with default parameters on all samples in the reference panel to annotate loops, merging annotations while excluding duplicates (allowing 1-bin mismatch). Last, we merged the loop annotations of

all reference panel samples (allowing 2-bin mismatches) and kept only annotations that (i) were present in at least 5 reference samples, but (ii) were absent (Chromosight score less than 0) in GM12878, obtaining 170,283 negative pairs. Overall, the entire set contains 74,855 unique positive and 256,609 unique negative examples.

For TAD boundary annotations, we first applied RobusTAD on the combined GM12878 Hi-C contact map and reference samples to obtain TAD boundary scores and identified boundaries. By merging TAD boundaries that were identified from all samples while excluding duplicates (allowing a 2-bin shift), we collected 48,945 loci. We then selected another 54,464 loci by picking one locus every five bins along every autosome. We define the targets for the 103,409 examples as RobusTAD scores from the combined GM12878 Hi-C data. In addition, we created another 103,409 examples at the same loci by using features from a shuffled GM12878 Hi-C contact map and the corresponding RobusTAD scores as targets. In total, there are 206,818 examples.

2.5.5 Model training and evaluation

The model was trained, evaluated, and tested on contact maps downsampled from the combined GM12878 Hi-C data. During model development, we used chr11 and chr12 for validation, chromosome 15-17 for testing, and the rest of autosomes for training. For loop prediction, the dataset contains 260,940 training, 34,174 validation, and 36,350 test examples. For TAD prediction, the dataset contains 164,458 training, 22,449 validation, and 19,911 test examples. RefHiC takes feature vectors from the study and reference samples as input in the forward pass. To reduce training computation, we sampled 10 reference samples for each example in each epoch independently. During evaluation, we used all samples in the reference panel. For both TAD and loop models, we trained models with a batch size of 1,024 for 1,000 epochs on an RTX6000 GPU and used AdamW optimizer [132] (weight_decay=0.1; learning rate=1e-3). We selected the learning rate that

yields the highest validation accuracy in our grid search and used early stopping to prevent overfitting. In the first 5 training epochs, we warmed up the learning rate from 0 to the initial learning rate (i.e. 1e-3) and then reduced the learning rate to 1e-6 in the first 95% epochs using the cosine annealing learning rate scheduler. In addition, we used dropout (rate=0.25), batch normalization, and layer normalization to regularize network training. We trained the TAD boundary model with MSE loss, and the loop model with focal loss $-(1-p_t)^{\gamma}\log(p_t)$ ($\gamma = 2$) [133]. To handle various sequencing depths in a single model, which many existing machine learning applications in Hi-C data analysis are unable to do [75, 76], we performed data augmentation by downsampling Hi-C contact maps during training. This transformation preserves topological structures in Hi-C data. However, a Hi-C contact map is too large, and downsampling on the fly is infeasible. We downsampled Hi-C training data and stored them on disk in advance. During training, we randomly selected one contact map from these downsampled contact maps for each training example in each epoch independently. This operation seamlessly worked as a data augmentation by downsampling Hi-C contact map operator during training.

2.5.6 Contrastive Pretraining

We pre-trained the encoder by supervised contrastive learning [128] using Hi-C contact maps downsampled from the combined GM12878 Hi-C data. For each training example, we defined items extracted from the downsampled contact maps at the sample locus as similar items and all Hi-C contact map submatrices in the same batch with different labels as negative items. We aimed to train the encoder such that the distances of embeddings for a training example and its similar items are as close as possible while of embeddings between a training example and its negative items are as far as possible. Following [128], we defined the loss for training instance i as cross-entropy with in-batch negatives

$$l_i = -\log \frac{e^{\text{sim}(\mathbf{h_i}, \mathbf{h_i^+})/\tau}}{\sum_{j \neq i} e^{\text{sim}(\mathbf{h_i}, \mathbf{h_j^-})/\tau}} \tag{2.2}$$

where $\mathbf{h_i}$, $\mathbf{h_i^+}$, and $\mathbf{h_j^-}$ are embeddings: $\mathbf{h_i}$ represents item i, $\mathbf{h_i^+}$ represents one of item i's similar items, $\mathbf{h_j^-}$ represents an item with a label different from i (i.e. negative item). τ is a temperature that controls training, and we set it as 1. We pre-trained the encoder for 20 epochs with the LARS algorithm [134] using Adam as a base optimizer. We set batch size to 512 and learning rate to 0.1 during training.

2.5.7 Hi-C data downsampling

We downloaded the combined Hi-C contact map (.mcool file) for GM12878 cells from 4DN Data Portal (https://data.4dnucleome.org). We downsampled the combined Hi-C contact map to train RefHiC and evaluate sequencing depths' impact on annotating topological structures. We did bilinear downsampling with the downsample function provided in FAN-C [135] from the combined Hi-C contact map to get a series of downsampled data until reaching at ∼62M valid read pairs.

2.5.8 Loop detection with Chromosight, Peakachu, Mustache and HiCCUPS

We used a variety of loop prediction tools to benchmark against RefHiC. They are executed as follows. Chromosight: We applied the program to each Hi-C contact map with parameters 'detect -p 0.2' to detect loops, sorted detected loops according to scores and selected the top loops from our test chromosomes. Peakachu: We trained different models for different sequencing depths on GM12878 Hi-C data using our training and validation

examples. To match RefHiC, we set the width parameter to 10 and other parameters as default values. We applied the trained models to Hi-C contact maps, adjusted the probability threshold in its pool function to identify loops, sorted loop annotations and included top loops from test chromosomes as its predictions. Mustache: We used the program by adjusting '-pt' and '-st' to detect at least 1700 loops on our test chromosomes, sorted and selected top loops according to FDR. HiCCUPS: We converted .mcool to .hic files at 5Kb resolution using the 'pre' function provided in Juicer [48]. We applied HiCCUPS by adjusting the '-f' parameter to detect at least 1700 loops on our test chromosomes, sorted and selected top loops according to FDR (obtained as the product of FDR for different filters) as HiCCUPS' prediction. To evaluate the performance of the recommended setting of each tool, we also applied them to annotate loops with their recommended parameters and set FDR as 5% whenever possible.

2.5.9 TAD detection with alternative tools

We used a variety of TAD callers to benchmark against RefHiC. All tools take .mcool file or files coverted from .mcool file as input. We ran TopDom, Armatus, Arrowhead, EAST, CaTCH, Domaincall (DI), GMAP, ICFinder and HiCSeg as suggested in [72]. We have updated parameters to reflect that we were analyzing data at 5kb resolution, as needed. We ran HiTAD and deDoc with their default settings. OnTAD: We set maxsz=600 to allow OnTAD to detect TADs as large as 3Mb. Grinch: following Lee and Roy [7], we detected TADs by setting the expected TAD length as 2Mb, 1Mb, and 500Kb in three runs and combined all results. Table 2.4 contains parameters that we used to execute each TAD caller.

We also compared RefHiC's boundary prediction to two boundary prediction tools. We reimplemented RobusTAD [115] in Python and used Insulation score (IS) function in cooltools [136] in our study. Both take a .mcool file as input. We used RobusTAD

to calculate TAD boundary scores and identified boundaries with default parameters. We ran IS with win=10 to detect TAD boundaries. As IS only detected insulating bins, to assign boundary orientation (i.e. left vs right), we used RobusTAD's left and right boundary scores to classify IS annotations.

2.5.10 Enrichment analysis of structural proteins and Histone-3 marks at predicted TADs

To compare the performance of TAD callers, we used an established TAD caller benchmarking scripts [72] to study Histone-3 marks and structural proteins enrichment inside TADs or at TAD boundaries. Briefly, we downloaded ChIP-Seq peak files for CTCF (ENCFF796WRU), RAD21 (ENCFF662DRZ), SMC3 (ENCFF887CRE), H3K36me3 (ENCFF171MDW), and H3K27me3 (ENCFF039JOT) from ENCODE [119]. For structural protein enrichment, we counted the average number of peaks per 5-kb intervals within the regions flanking predicted TAD boundaries ($\pm$ 500 kb). Next, we computed the fold-change as the average peaks in a narrow interval surrounding a boundary ($\pm$10kb) over the average peaks coverage at distant flanks ($\pm$400-500kb). For Histone-3 marks enrichment analysis, we split TADs into 20kb intervals, summed ChIP-Seq signals inside each interval, computed the log10-ratio of H3K27me3 and H3K36me3 signals (LR), and obtained the average LR for each TAD. We then shuffled the LR values ten times to compute an empirical p-value for within-TAD LRs and corrected the p-value with the Benjamini-Hochberg procedure to select TADs with significant preference for high or low ratios (FDR$\leq$0.1). To compare TAD partitions, following Zufferey et al. [72] , we used the Measure of Concordance (MoC), which ranges from 0 (absence of concordance) to 1 (full concordance) and is defined as follows,

$$\mathrm{MoC}(\mathbf{P},\mathbf{Q}) = \begin{cases} 1 & \text{if } N_P = N_Q = 1 \\ \frac{1}{\sqrt{N_P N_Q}-1}\left(\sum_{i=1}^{N_P}\sum_{j=1}^{N_Q}\frac{|\mathbf{F_{i,j}}|^2}{|\mathbf{P_i}||\mathbf{Q_j}|}-1\right) & \text{otherwise} \end{cases} \tag{2.3}$$

where $\mathbf{P} = \{\mathbf{P_i}\}$, and $\mathbf{Q} = \{\mathbf{Q_i}\}$ are sets of TADs including N_P and N_Q TADs, $\mathbf{F_{i,j}}$ is the overlap region between $\mathbf{P_i}$ and $\mathbf{Q_j}$, and $|\cdot|$ represents cardinality. MoC does not handle nested TADs, thus we only included TADs without any smaller TAD in this analysis.

2.5.11 Enrichment analysis of transcription factors and histone modifications at loop anchors

We downloaded ENCODE ChIP-Seq peak files for 122 TFs and 11 histone modifications in the GM12878 from the UCSC genome browser [119, 137] and calculated occupancy fold changes for each TF at loop anchors. We first created a list of unique loop anchors inferred by each tool. For each TF, we counted the number of anchors that overlapped with at least one binding site. We denoted this value as the target. For each chromosome, we randomly created 100 control sets of anchors from the whole genome excluding blacklisted regions [119]. The number of anchors in each control set equals the number of loop anchors in the target set. We then computed the expected overlaps as the mean of overlaps between each control set and the TF's binding sites. Last, we computed fold change as the ratio between the target and the expectation calculated based on control sets.

2.5.12 Hi-C reference panel

Human reference panel: We downloaded Hi-C sequencing data from the GEO repository and processed them with distiller (https://github.com/open2c/distiller-nf). Briefly, we used bwa mem [51] to map reads to hg38 with option '-SP' and processed the aligned reads with pairtools (https://github.com/open2c/pairtools) to remove duplicates and

low-quality read pairs (MAPQ<10). We then created and normalized contact matrices at 5 kb resolutions using cooler [49] and saved contact maps in .mcool files. Last, we converted these .mcool files into the .bcool file format using cool2bcool function provided in RefHiC. The .bcool format represents a Hi-C contact map as a band matrix and enables fast random access to square submatrices. Table 2.1 lists all Hi-C data sets included in the human reference panel. Mouse reference panel: We downloaded 20 Hi-C contact maps from 4DN Data Portal (https://data.4dnucleome.org) and processed them as for human. Table 2.2 lists all Hi-C data sets included in the mouse reference panel. Our distributed reference panels contain the aforementioned reference samples. In our experiments, we excluded samples that belong to the study sample's cell type from the reference panel to prevent potential data leakage.

2.5.13 RefHiC implementation

RefHiC is a Python program available at https://github.com/BlanchetteLab/RefHiC. We implemented the neural network with the PyTorch library [138], and the filtering components for TAD and loop selection with libraries including Pandas [139], SciPy, and NumPy [140]. Using RefHiC to predict loops or TAD boundary scores requires loading data from the study and reference Hi-C contact maps. To reduce memory usage, we extended the Cooler [49] library by implementing a band matrix representation for a contact map and a square function to fetch contact pairs in a given square region and used it to read Hi-C contact maps. ReHiC can make predictions on both CPU and GPU, but is much faster on the latter. RefHiC requires at least 3GB free space for saving reference panel data and at least 12GB RAM for loading reference samples during prediction. We tested RefHiC to annotate TAD boundaries and loops using 20 CPU threads and an RTX6000 GPU. RefHiC calculated TAD boundary scores for whole genome annotation at 5 kb resolution in 30 min. It is impractical and unnecessary to calculate loop scores for all

pairs of loci. RefHiC only computes loop scores at bin pairs located within 3 Mb and for which at least one read pair is observed. Thus, the loop annotation running time depends on the study contact map. For instance, it annotates Hi-C data containing 500M valid read pairs in 275 min and Hi-C data containing 250M valid read pairs in 180 min.

2.6 Data Availability

The data that support this study are available from the corresponding author upon reasonable request. All data used in this study are publicly available and their reference numbers are listed in Table 2.1, 2.2, and 2.3. Hi-C contact maps were obtained from 4DN data portal with the following accession code: 4DNFIXP4QG5B (GM12878), 4DNFI4DGNY7J (K562), 4DNFIJTOIGOI (IMR90), 4DNFILP99QJS (HCT-116), and 4DNFIDA2WGV8 (mESC). ChIP-Seq data were obtained from the ENCODE portal with the following accession code: ENCFF796WRU (GM12878 CTCF), ENCFF039JOT (GM12878 H3K27me3), ENCFF662DRZ (GM12878 RAD21), ENCFF171MDW (GM12878 H3K36me3), ENCFF887CRE (M12878 SMC3), ENCFF508CKL (mESC CTCF), ENCFF203SRF (IMR-90 CTCF), ENCFF119XFJ (K562 CTCF). The CTCF ChIA-PET for IMR-90 were obtained from ENCODE with accession code ENCFF682YFU. The CTCF ChIA-PET for mESC were obtained from ENCODE with accession code ENCFF550QMW. The RAD21 ChIA-PET for GM12878 were obtained from ENCODE with accession code ENCLB784HEF. The CTCF ChIA-PET for K562 were obtained from ENCODE with accession code ENCFF001THV. The RAD21 ChIA-PET for K562 were downloaded from the GEO repository with accession code GSM1436264. The H3k27ac HiChIP data for GM12878 were obtained from [120]. The SMC1 HiCHIP data for GM12878 were obtained from [37]. The CTCF ChIA-PET data for GM12878 were obtained from [118]. Experiment results and intermediate data generated in this study have been deposited in the zenodo repository with DOI 10.5281/zenodo.7133194.

2.7 Code availability

Software and documentation available at https://github.com/BlanchetteLab/RefHiC or at this DOI:10.5281/zenodo.7324669. All scripts required to reproduce figures and analyses are available at DOI:10.5281/zenodo.7133194.

2.8 Acknowledgements

The authors thank Dr. Yue Li, Dr. Jacek Majewski and members of M.B.'s laboratory Audrey Baguette, Zichao Yan, and Elliot Layne for useful discussions in this project, and Audrey Baguette for testing RefHiC. This work was funded by Genome Quebec/Canada and a Genome Quebec/Oncopole/IVADO grants to M.B., and FRQNT Doctoral (B2X) Research Scholarships to Y.Z..

2.9 Contributions

Y.Z. and M.B. conceived the study and designed models. Y.Z. implemented models, performed data analysis, and wrote the manuscript. M.B. supervised the project and wrote the manuscript. All authors read and approved the final paper.

2.10 Competing interests

The authors declare no competing interests.

2.11 Appendix

Supplementary Information

2.11.1 Comparison with contact map enhanced approach

Hi-C contact map enhancement has been intensively explored in recent years [76, 9, 59, 75, 78]. We can perform contact map enhancement for Hi-C contact maps and then use loop detection tools to detect loops. Here we compared RefHiC with tools that predict loops from DeepLoop [59] enhanced contact maps. DeepLoop is a deep learning approach for Hi-C contact map enhancement. It enhances or denoises loop signals from Hi-C contact maps. It outputs a list of pairs (i.e. loop clusters) with a value named LoopStrength. Although Zhang et al. [59] suggest selecting the highest LoopStrength pairs as loops, the results are still loop clusters. We called loops from DeepLoop's output by selecting loops with RefHiC's clustering method (minDelta=5 and minScore=0). Due to the absence of essential input files for running DeepLoop for contact maps aligned to hg38, we were unable to duplicate DeepLoop's result with our data (hg38), so we used DeepLoop's output (GSE167200, hg19) produced by Zhang et al. [59] for Rao's GM12878 Hi-C data. We executed RefHiC on the same data but aligned to hg38. To facilitate comparison, we lifted RefHiC's predictions over to hg19 using liftOver (https://github.com/dphansti/liftOverBedpe). Fig. 2.9 shows RefHiC identified more ChIA-PET/HiCHIP-supported loop predictions than all DeepLoop-based approaches when evaluated using CTCF ChIA-PET, H3K27ac HiCHIP, and SMC1 HiCHIP data.

2.11.2 Comparison with a baseline deep learning model

Although we demonstrated that RefHiC outperformed HiCCUPS, Mustache, Chromosight, and Peakachu, these results do not imply that these gains can necessarily be attributed to the use of a reference panel. To address this question, we built a deep learning model named Baseline. The model architecture is similar to RefHiC's but without using the reference panel as input. It consists of (i) an encoder that is identical to RefHiC's and (ii) four fully connected layers with $[d, d/2, d/2, 1]$ units. All hidden layers are ReLU activated. Like RefHiC, we used batch normalization and dropout to regularize our model. The training data and procedure are identical to RefHiC's. We applied Baseline model to several downsampled GM12878 Hi-C contact maps and compared it with all other tools. Fig. 2.10 shows that RefHiC outperformed Baseline and other tools at all sequencing depths. This result suggests RefHiC benefits from the introduction of a reference panel.

2.11.3 Additional analysis of the properties of loop predictions

To understand the specific properties of the loops predicted by each tool, we studied local interaction frequency ranks and radii of predicted loops. We found the interaction frequencies for $17-36\%$ of loops predicted by Chromosight, RefHiC, Peakachu, and Mustache were local maxima in a 3×3 region centered at loop predictions. In contrast, 75% of loops predicted by HiCCUPS were local maxima in the same region (Fig. 2.12). Loops predicted by HiCCUPS had smaller radii than alternative tools (Fig. 2.13). These two observations explained the diffuse loop center in Fig. 2.2b. Although we define loops as locally enriched significant interactions, they are not necessarily local maxima in terms of interaction frequencies. To demonstrate the validity of non-locally-optimal loops, We moved all non-locally-maximal loop predictions to locally maximal contact pairs in the surrounding 3×3 region. Fig. 2.14 shows that for all comparisons excepts the HiCCUPS/RAD21, revised predictions are worse than the original predictions.

Loop size estimation

Loops form blob-shaped patterns in Hi-C contact matrices, and we can approximately represent them as circles with a minimum radius of 0.5 bin. To estimate a predicted loop's size (radius), we utilized the scale space representation [18] in computer vision by treating Hi-C contact maps around a loop prediction as an image. Briefly, for a given loop prediction at contact bins (i, j), we extracted the 21×21 Hi-C sub-matrix centred at (i, j) and used the 'blob_dog' function in scikit-image [141] to compute the radius of the blob that covers (i, j). As not all loops could be detected by 'blob_dog', we only included loops with radii found by 'blob_dog' in our analysis.

2.11.4 Applying RefHiC to novel cell types

Although we demonstrated that RefHiC does not produce more false-positive predictions than alternative tools in analyzing cohesin-depleted Hi-C data (Fig. 2.3a), we did not benchmark RefHiC in analyzing distant study samples. To evaluate it, we performed two different experiments (i.e. method 1, and 2). We applied both methods to detect loops from test chromosomes (chromosomes 15-17) on a downsampled GM12878 Hi-C contact (500M valid contact pairs). Method 1 uses the default human reference panel and assumes that all test chromosomes correspond to chromosome 2 in the reference panel. For instance, when computing the loop score for the pair (chr15:100000-105000, chr15:180000-185000), we extracted data for (chr2:100000-105000, chr2:180000-185000) from the reference panel. In this case, all loops in our study sample are novel. Method 2 uses a new reference panel containing eight samples. This new reference panel does not contain any samples that come from the same developmental lineage as the study sample or samples such that the similarity [142, 143] between it and the study sample is larger than 0.7. Fig. 2.20 shows that RefHiC with input data configured as Method 1 performed equally well as conventional tools. At the same time, worse than RefHiC (with default reference panel) and our deep learning baseline model. RefHiC with reference panel defined as Method 2 is equally well as RefHiC (with default reference panel).

2.11.5 RefHiC outperforms a global similarity based approach

RefHiC's superior performance is achieved by its deep learning module using local information from the study sample and reference samples for topological structure annotation. It is interesting to evaluate the performance of a simple approach that uses only reference data similar to the study sample to make predictions. We first compared the study sample against each reference sample with HiCRep [142, 143] and selected 5 reference samples most similar to the study sample. We then performed Mustache and our baseline deep learning model (Section 2.11.2) to detect loops from the five reference samples. We evaluated four different alternative approaches: Mustache top-1, predict loops as loops predicted from the most similar reference sample by Mustache, Mustache top-5, predict loops as the ensemble of loops predicted from the five reference samples by Mustache, DL-Baseline top-1, predict loops as loops predicted from the most similar reference sample by our deep learning baseline, DL-Baseline top-5, predict loops as the ensemble of loops predicted from the five reference samples by our deep learning baseline. Applying both tools to the five samples identified 19,890 (Mustache), and 32,520 (DL-Baseline) unique loops, and many of them are nearby loops within 5 bins each other (i.e. a cluster of loops) within each set. Each cluster may correspond to a single true loop. We thus used RefHiC's pooling algorithm to select a represent loop for each cluster while using the loop occurrence in the five samples as loop score. We selected 2,962 (Mustache top-5) and 3,060 (DL-Basele top-5) loops at the end (a simple voting ensemble does not work as most loops only occur once among the five samples). Fig. 2.23 shows RefHiC outperformed all other approaches significantly, and some of the four proposed approaches are slightly worse than applying the corresponding tool directly to the study sample.

ID	Sample	Valid read pairs	Source
HIC00001	22Rv1 (prostate cancer cell line)	685096962	[144]
HIC00002	293TRex-Flag-BRD4-NUT-HA	1660296754	[145]
HIC00007	BLaER (lymphoblastic leukemia cell line)	406322404	[146]
HIC00041	HCT-116 (colorectal cancer cell line)	603179292	[147]
HIC00067	HeLa Kyoto cell, MboI G1 sync control	1045885928	[148]
HIC00090	HepG2 (hepatocellular carcinoma cell line)	1759654311	[119]
HIC00091	HL60/S4 (neutrophil-like Myeloid leukemia cell line)	478434139	[149]
HIC00113	Nalm6 (B cell precursor leukemia cell line)	816274711	[150]
HIC00168	WI38_RAF (WI-38hTERT/GFP-RAF1-ER)	602556180	[151]
HIC00172	Embryonic stem cell, Cardiomyocyte differentiation : hESCs (day 0)	1914642484	[152]
HIC00183	teloHAEC (endothelial cell line)	911486437	[153]
HIC00200	Naïve human embryonic stem cells	731906045	[154]
HIC00203	GM23248 (primary skin fibroblasts)	1797370277	[155]
HIC00221	MDM (monocyte-derived macrophages)	590449106	[156]
HIC00269	Astrocytes of the cerebellum primary cell	430822244	[119]
HIC00273	HAP1 (near-haploid cell line)	413436528	[157]
HIC00280	Purified human germinal center B cells	426222299	[158]
HIC00287	Liver	447028100	[159]
HIC00295	Thymus	507309033	[159]
HIC00296	H1 Embroynic Stem Cell	989388439	[160]
HIC00310	A549 00h 100 nM dexamethasone	1548684355	[119]
HIC00318	HUVEC	438880295	[161]
HIC00319	IMR90	1053932182	[161]
HIC00320	K562	880877579	[161]
HIC00321	KBM7	877658969	[161]
HIC00322	NHEK	653628335	[161]
HIC00337	Gastric tissue	426476775	[162]
HIC00343	Left Ventricle	547477074	[162]
HIC00354	Spleen	490487515	[162]
HIC00360	GM12878	1994319522	[161]

Table 2.1: **Human reference panel**

ID	Sample	Valid read pairs	Source
4DNFIKK3QG34	46C with Sox1-GFP	5370123882	[122]
4DNFIHBTUDO9	Olfactory receptor cells	2024307320	[163]
4DNFIAVHP5AV	B cell derived cell line	1869747094	[164]
4DNFIMV54HXI	Olfactory receptor cells	1585896942	[163]
4DNFI5QJNFAT	F123-CASTx129 (Tier 2)	1072601088	[165]
4DNFIS5ZK13C	CH12	1163734014	[164]
4DNFI3M6726I	Olfactory receptor cells	1202229560	[163]
4DNFIPNP9H9T	B cell derived cell line	1125298567	[164]
4DNFI3QLT3KJ	B cell derived cell line	1203634092	[164]
4DNFI6RG9TXL	ES-E14	812429482	[166]
4DNFIJLJIRKT	CH12.LX	770649212	[4]
4DNFIB8NZIAK	F123-CASTx129 with Sox2 tags and RMCE site between Sox2 and its SE	770655888	[167]
4DNFITSJBJ9G	F123-CASTx129 with Sox2 tags and RMCE site between Sox2 and its SE	759678776	[167]
4DNFIASQYF5S	CH12	915292686	[164]
4DNFI4LH8RMQ	ES-E14 with Flo/Flox deletion of Mll3 and Mll4 genes	717826154	[166]
4DNFIOPKGMBL	ES-E14	576226710	[166]
4DNFIB7RFFBB	Sertoli cell	589079877	[168]
4DNFIX524X88	ES-E14 with Flo/Flox deletion of Mll3 and Mll4 genes	693626370	[166]
4DNFI37GAU3L	ES-E14 with Flo/Flox deletion of Mll3 and Mll4 genes	579214478	[166]
4DNFIC453HVL	F123-CASTx129 (Tier 2)	514097723	[165]

Table 2.2: **Mouse reference panel**

Experiment	Training	Evaluation	Figure	Identifier
GM12878 CTCF ChIP-Seq		✓	Fig. 2.2g,h, 2.6b,f, 2.16, 2.17, 2.21b, 2.22	ENCFF796WRU
GM12878 H3K27me3 ChIP-Seq		✓	Fig. 2.6g	ENCFF039JOT
GM12878 RAD21 ChIP-Seq		✓	Fig. 2.6f, 2.22	ENCFF662DRZ
GM12878 H3K36me3 ChIP-Seq		✓	Fig. 2.6g	ENCFF171MDW
GM12878 SMC3 ChIP-Seq		✓	Fig. 2.6f, 2.22	ENCFF887CRE
IMR-90 CTCF ChIA-PET		✓	Fig. 2.3e	ENCFF682YFU
mESC CTCF ChIP-seq		✓	Fig. 2.3h	ENCFF508CKL
mESC CTCF ChIA-PET		✓	Fig. 2.3g	ENCFF550QMW
IMR-90 CTCF ChIP-Seq		✓	Fig. 2.3f	ENCFF203SRF
GM12878 RAD21 ChIA-PET	✓	✓	Fig. 2.2e, 2.4d, 2.10, 2.11, 2.14, 2.15, 2.18, 2.20, 2.23	ENCLB784HEF
GM12878 H3k27ac HiCHIP	✓	✓	Fig. 2.4f, 2.8a, 2.9-2.11, 2.14, 2.15, 2.18, 2.20, 2.23	[120]
GM12878 SMC1 HiCHIP	✓	✓	Fig. 2.2f, 2.4e, 2.9-2.11, 2.14, 2.15, 2.18, 2.20, 2.23	[37]
GM12878 CTCF ChIA-PET	✓	✓	Fig. 2.2d, 2.4c, 2.6c, 2.9-2.11, 2.14, 2.15, 2.18, 2.20, 2.23	[118]
K562 CTCF ChIA-PET		✓	Fig. 2.3b	ENCFF001THV
K562 RAD21 ChIA-PET		✓	Fig. 2.3c	GSM1436264
K562 CTCF ChIP-seq		✓	Fig. 2.3d	ENCFF119XFJ

Table 2.3: **Different datasets used in this study**

Tool	Configuration (Parameters)
CaTCH	default (resol=5000)
EAST	EAST2, default (resol=5000)
HiTAD	default (resol=5000)
GMAP	default (resol=5000)
Armatus	-m -N -r 5000 -g 0.5 -s 0.05 -n 100
deDoc	default (resol=5000)
Grinch	-e 1000000,2000000,500000 (resol=5000)
OnTAD	-maxsz 600 (resol=5000)
Arrowhead	default (resol=5000)
DomainCall	default (resol=5000)
TopDom	default (resol=5000)
ICFinder	default (resol=5000)
HiCSeg	6 TADs per 1MB region, Gaussian distribution, block-diagonal model (resol=5000)

Table 2.4: **TAD callers and parameters used in this study**

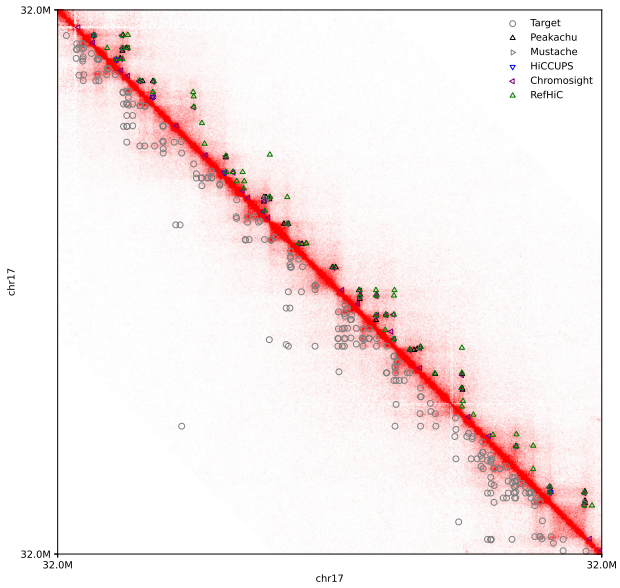


Figure 2.7: **Comparison of loops predicted by RefHiC, Chromosight, Peakachu, HiCCUPS, and Mustache on a small region of GM12878 HiC data (500M valid read pairs).** Target annotations are loops revealed by experimental data including CTCF ChIA-PET, RAD21 ChIA-PET, SMC1 HiCHIP, and H3K27ac HiCHIP.

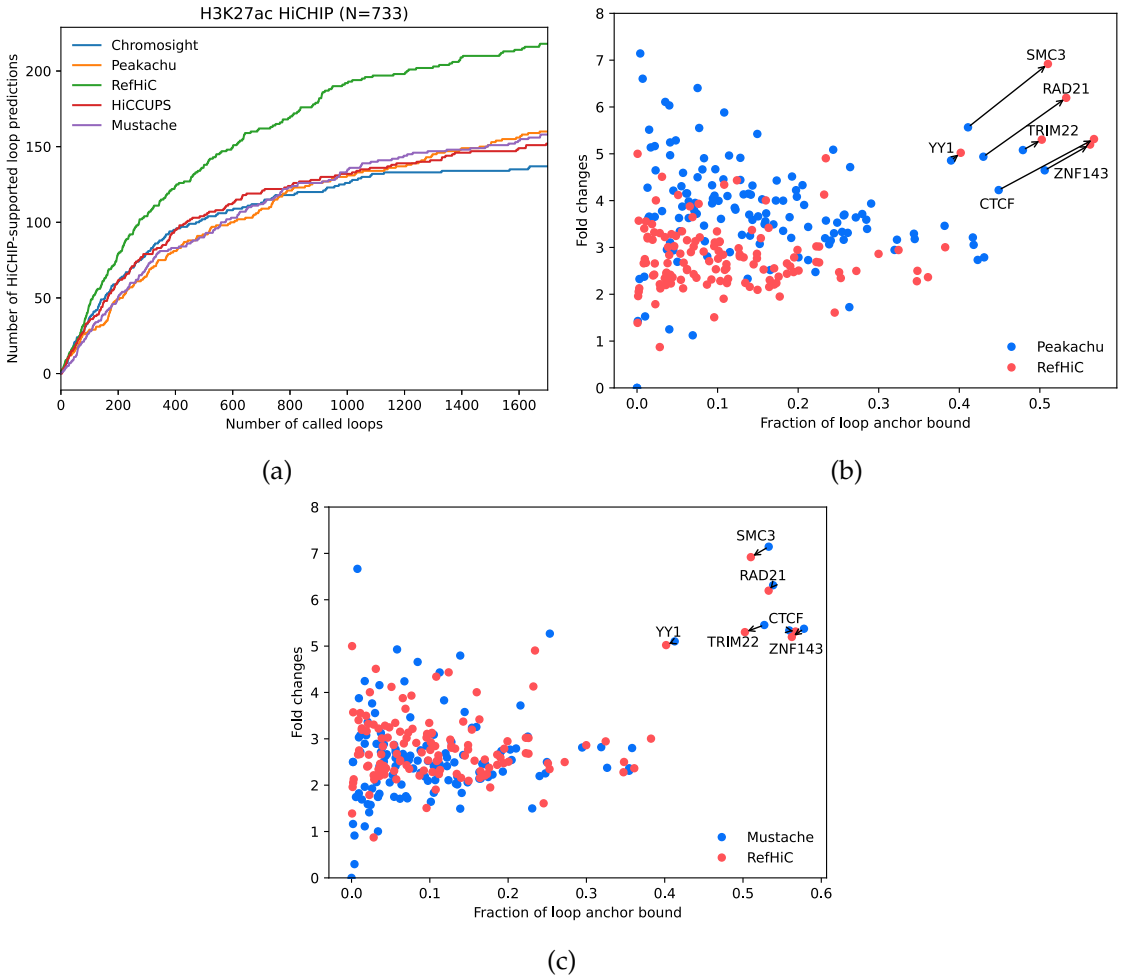
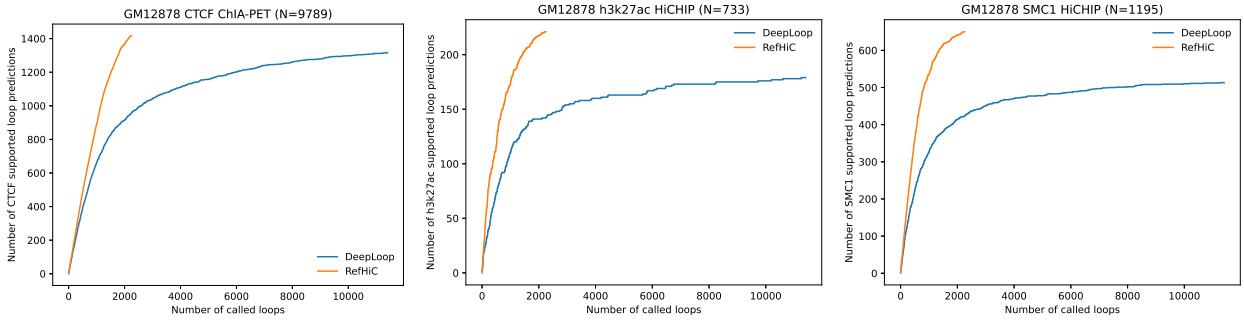


Figure 2.8: **Additional comparison of RefHiC, Chromosight, Peakachu, HiCCUPS, and Mustache on GM12878 HiC data (500M valid read pairs). a,** Same as Fig. 2.2d, but compared against H3K27ac HiCHIP data. **b,** Same as Fig. 2.2i, but for RefHiC's and Peakachu's predictions. Though most transcription factors are more strongly enriched at Peakachu's loop predictions than at RefHiC's loop predictions, the TFs involved in loop formulations are more strongly enriched at RefHiC loop predictions. **c**, Same as c, but for RefHiC's and Mustache's prediction. The TF are enriched similarly at both types of loop predictions.



(a) (b) (c)

Figure 2.9: **Comparison of loops predicted by RefHiC and DeepLoop on GM12878 HiC data (2000M valid read pairs).** Number of ChIA-PET/HiCHIP-supported loop predictions, among predictions made by RefHiC and DeepLoop for test chromosomes 15-17 compared against CTCF ChIA-PET (a), H3K27ac HiCHIP (b), and SMC1 HiCHIP (c). RefHiC's loop predictions matches all data better than predictions made by DeepLoop.

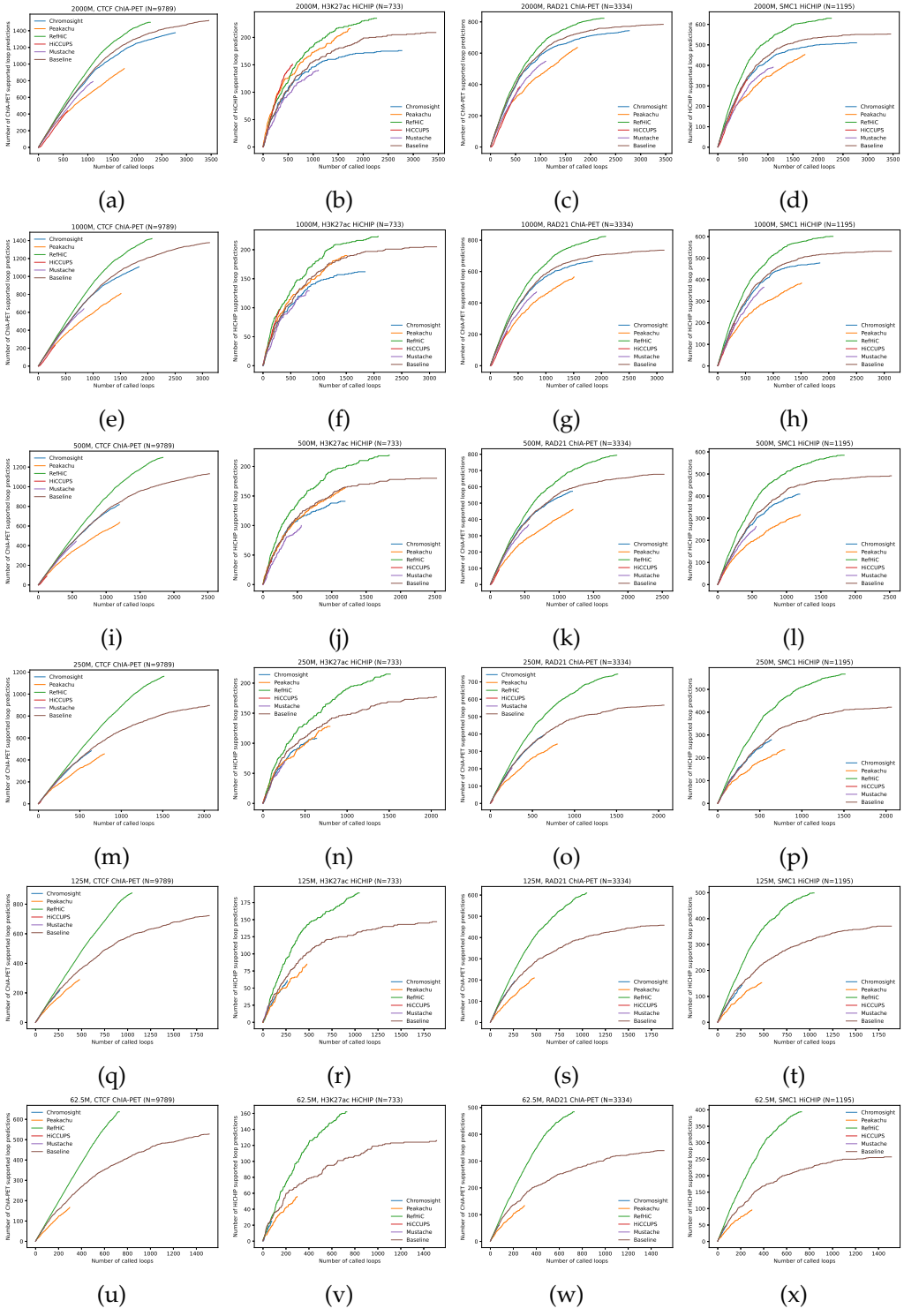


Figure 2.10: **Comparison of RefHiC, Chromosight, Peakachu, HiCCUPS, Mustache, and Baseline (Section 2.11.2) on GM12878 HiC data of lower sequencing depths.** Number of ChIA-PET/HiCHIP-supported loop predictions, among the top predictions made by RefHiC and other tools, for test chromosomes chromosomes 15-17, compared against CTCF or RAD21 ChIA-PET, as well as H3K27ac or SMC1 HiCHIP. The Baseline model, which does not use a reference panel, does not perform as well as RefHiC.

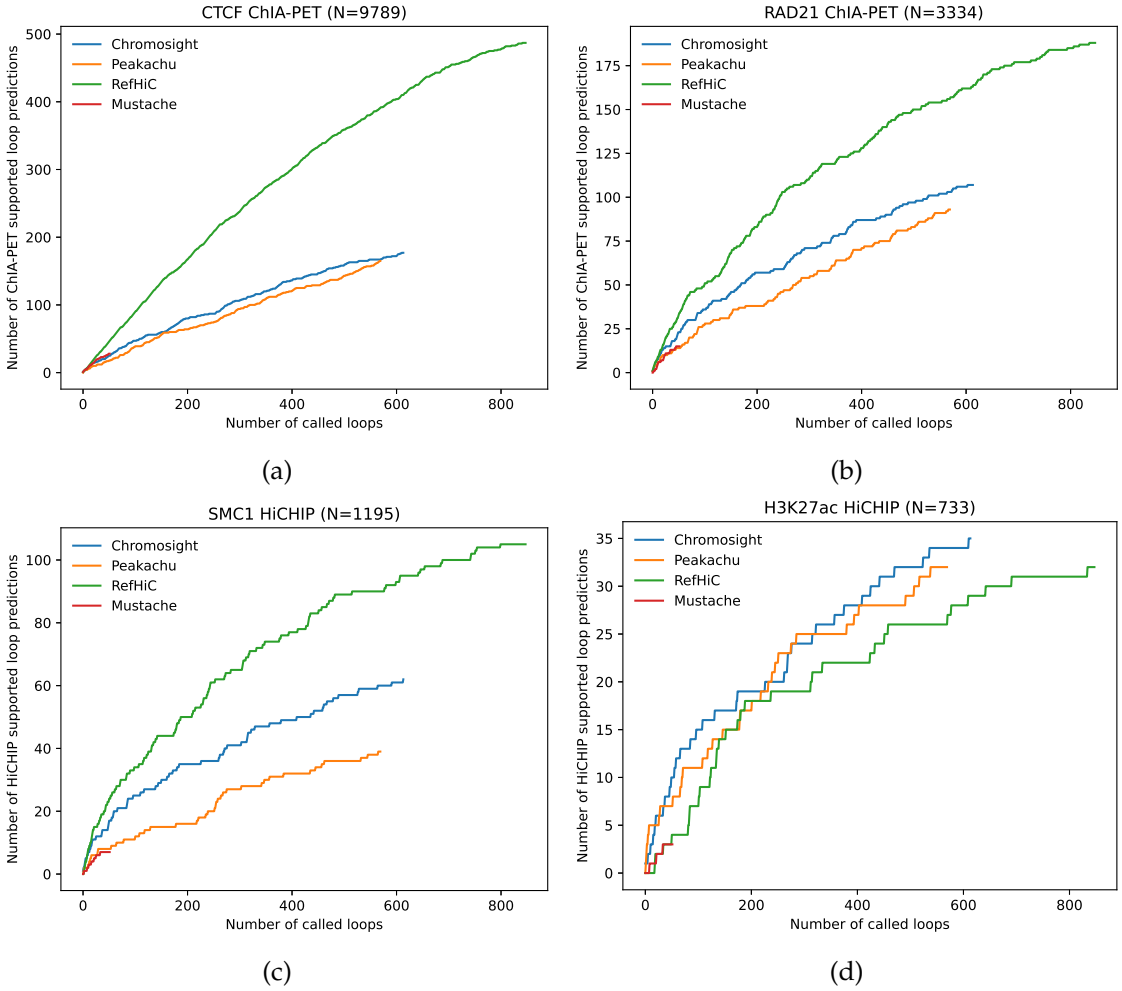


Figure 2.11: **Comparison of unique loops predicted by RefHiC, Chromosight, Peakachu, HiC-CUPS, and Mustache on GM12878 HiC data (500M valid read pairs).** Number of ChIA-PET/HiCHIP-supported loop predictions, among the top predictions made by RefHiC and other tools, for test chromosomes chr15-17, compared against CTCF ChIA-PET (a), RAD21 ChIA-PET (b), SMC1 HiCHIP (c), and H3K27ac HiCHIP (d). Only loops uniquely predicted by one of the four tools are being considered.

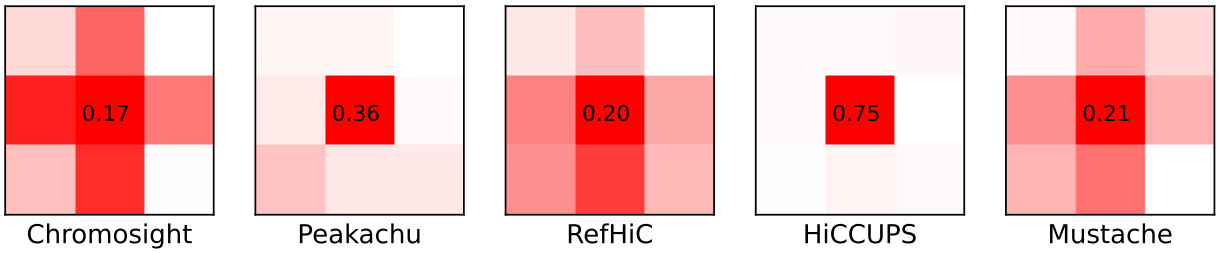


Figure 2.12: **For each prediction tool, the heatmap shows the fraction of predicted loops where the locally maximum interaction frequency is observed at the site of the predicted loop itself (central pixel) or at one of the 8 neighboring pixels.** HiCCUPS tends to predict local maximas as loops. In contrast, only 17-36% of predicted loops produced by other tools, including RefHiC, are local maxima in terms of read counts.

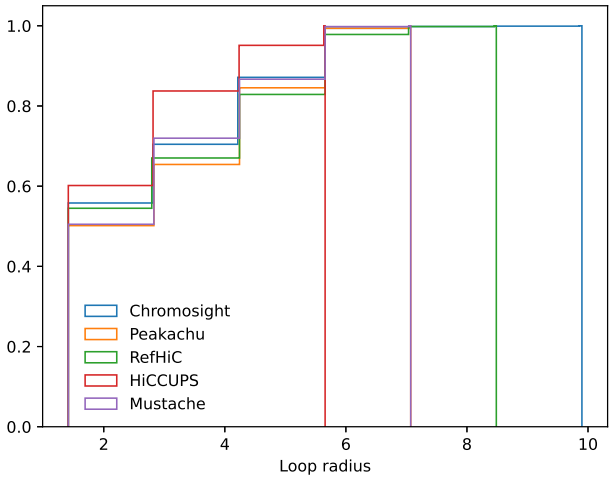
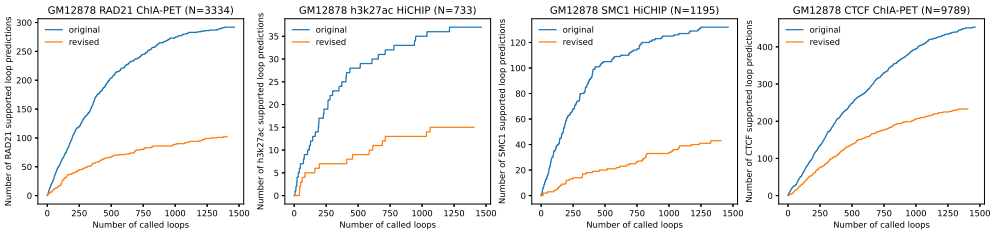
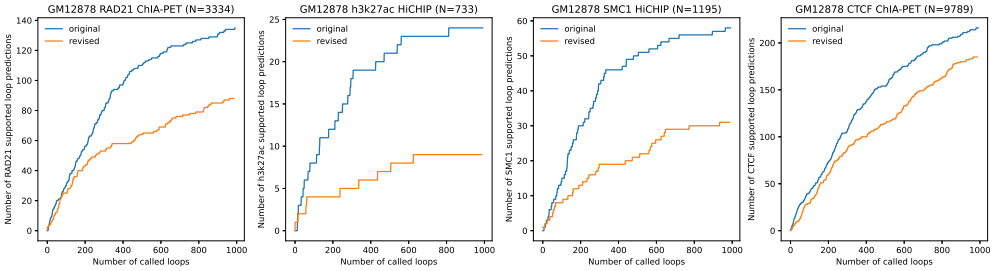


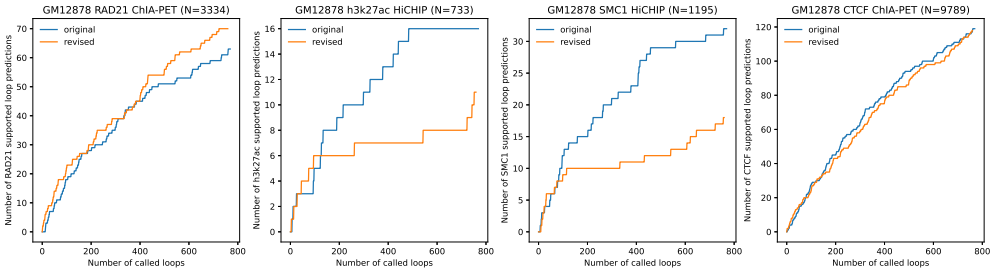
Figure 2.13: **Radius of loops predicted by RefHiC and other tools.** The cumulative distribution of loop radius shows that HiCCUPS predicted more narrow loops than other tools, including RefHiC. In contrast, the distribution of loop sizes in Chromosight's, Peakachu's, RefHiC's, and Mustache's predictions are similar.



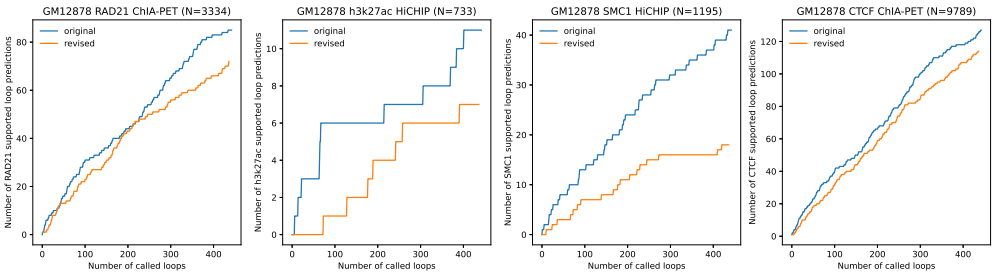
(a) RefHiC



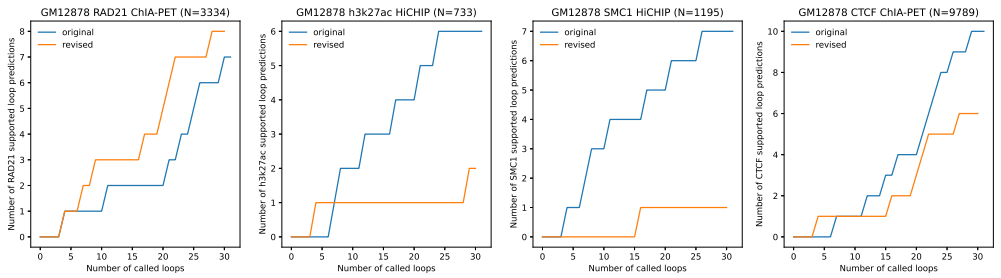
(b) Chromosight



(c) Peakachu



(d) Mustasche



(e) HiCCUPS

Figure 2.14: **Comparison of the performance of different prediction tools, compared to a modified version (labeled "revised") where each loop prediction is "corrected" to the local maximum (in terms of read count) in the 3×3 sub-matrix around each original prediction.**

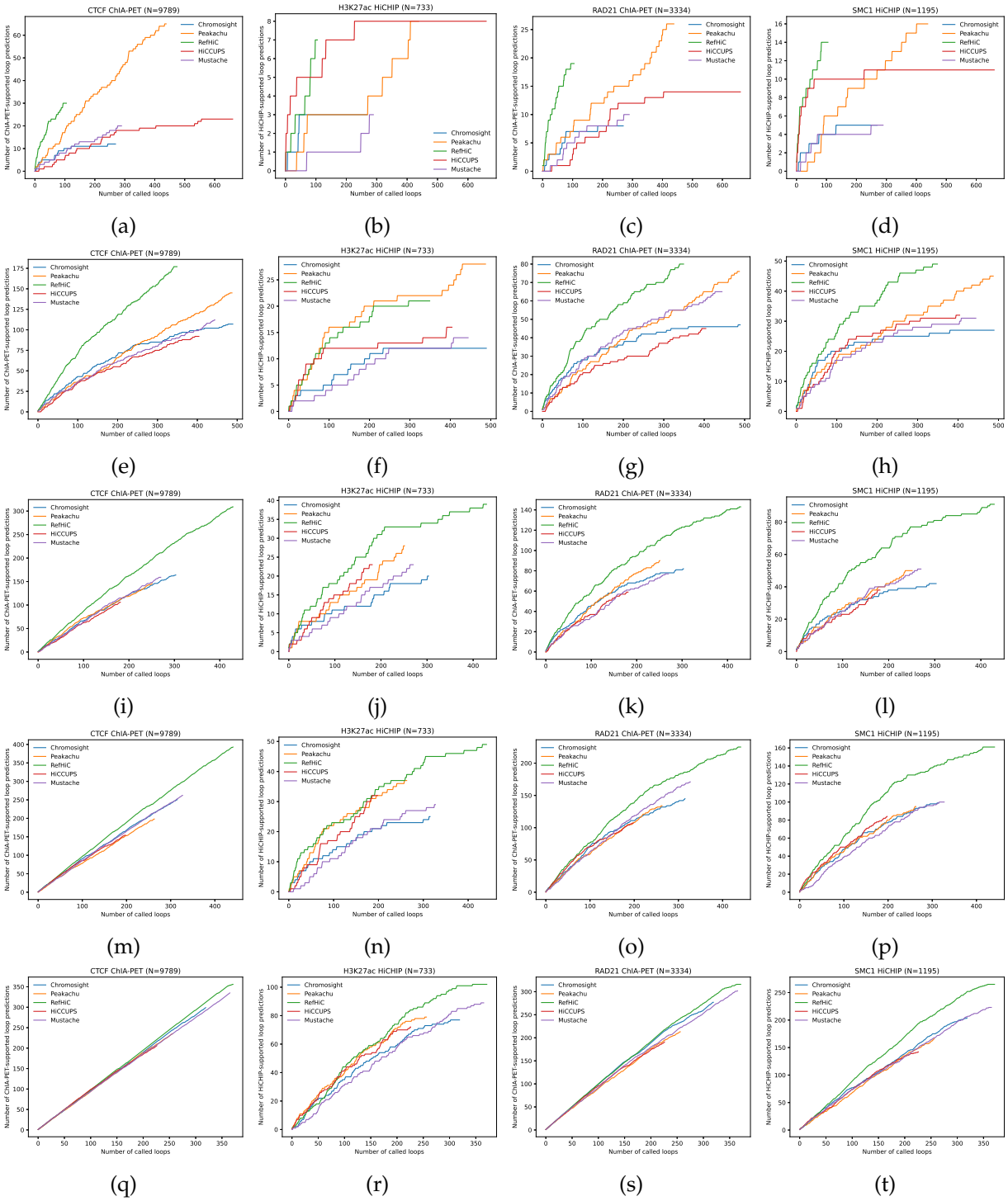


Figure 2.15: **Comparison of tools' ability to identify rare and common loops.** Number of ChIA-PET/HiCHIP-supported loop predictions, among the top predictions of loop frequencies 0 (a-d), 1-5 (e-h), 6-10 (i-l), 11-15 (m-p), and 15-19 (q-t), made by RefHiC and other tools, for test chromosomes chr15, chr16, and chr17, compared against CTCF or RAD21 ChIA-PET, as well as H3K27ac or SMC1 HiCHIP. RefHiC's loop predictions matches those experimental data better than predictions made by other tools or equally well for both rare and common loops.

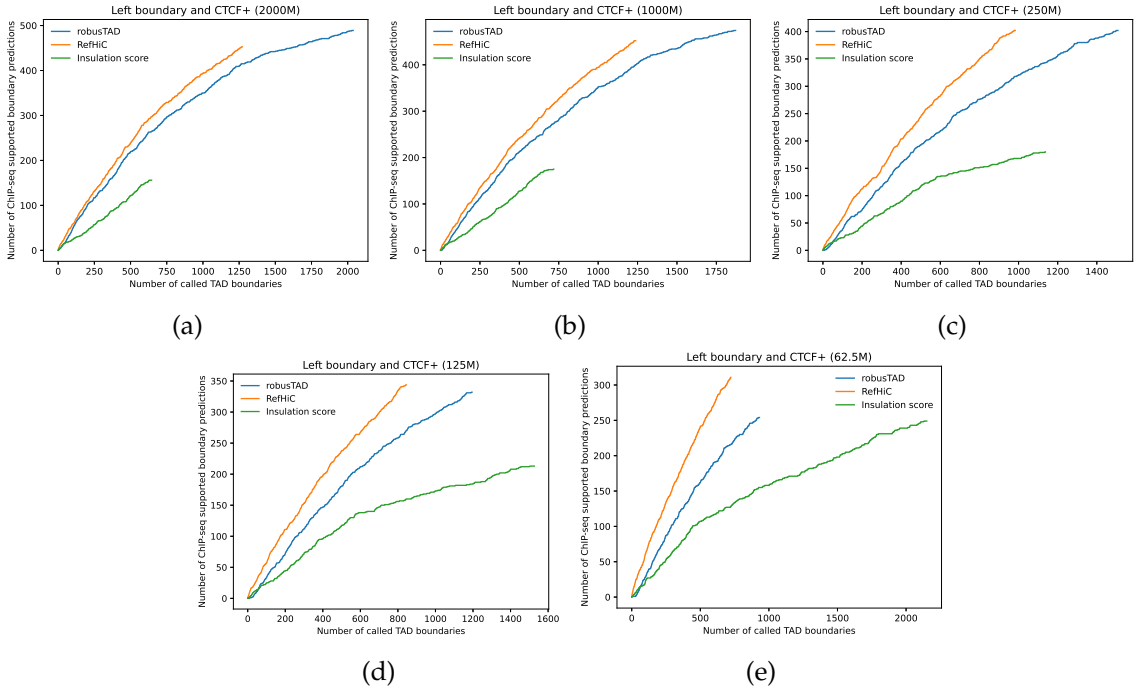


Figure 2.16: **Comparison of RefHiC, robusTAD, and Insulation score on detecting left TAD boundaries from GM12878 HiC data at lower sequencing depths.** Number of predicted TAD boundaries for Hi-C data containing 2,000M (a), 1,000M (b), 250M (c), 125M (d), and 62.5M (e) vaild read pairs supported by CTCF ChIP-seq data (test chromosomes chr15, chr16, and chr17).

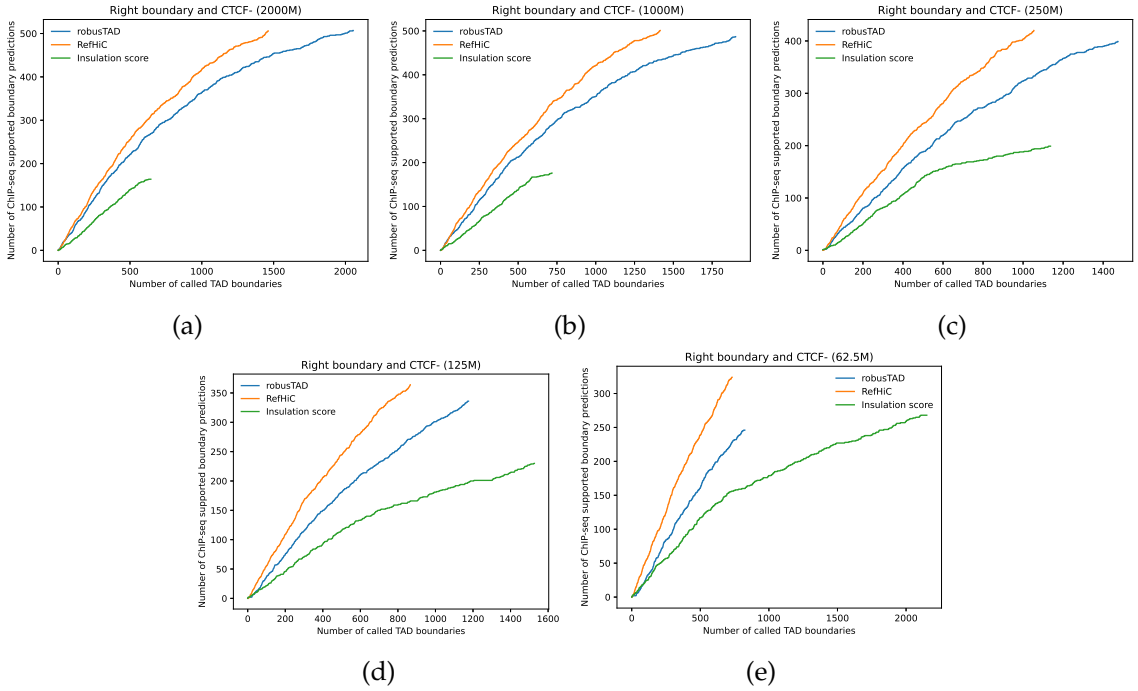


Figure 2.17: **Comparison of RefHiC, robusTAD, and Insulation score on detecting right TAD boundaries from GM12878 HiC data at lower sequencing depths.** Number of predicted TAD boundaries for Hi-C data containing 2,000M (a), 1,000M (b), 250M (c), 125M (d), and 62.5M (e) vaild read pairs supported by CTCF ChIP-seq data (test chromosomes chr15, chr16, and chr17).

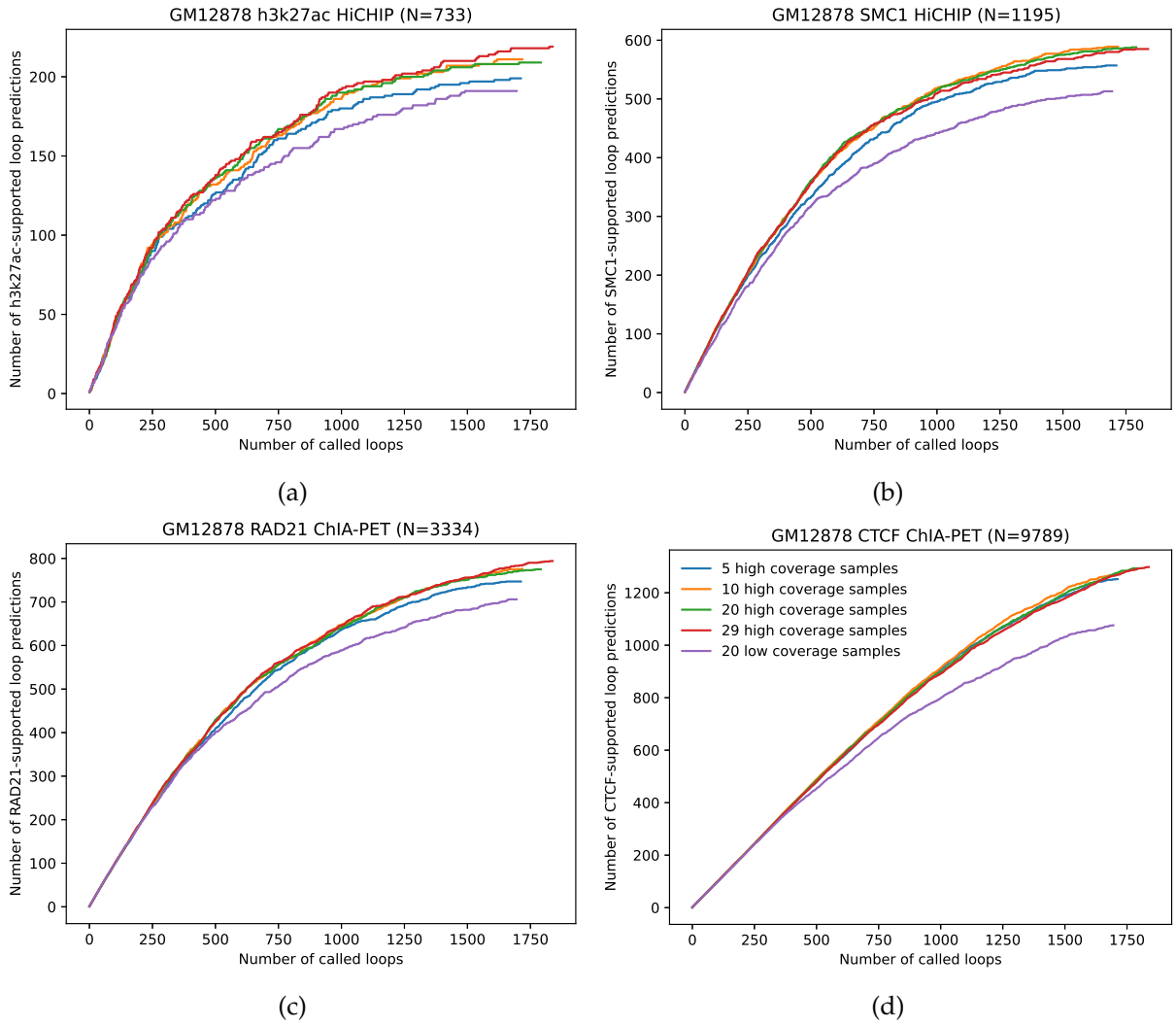


Figure 2.18: **RefHiC Detects loops using reference panel with different samples for GM12878 Hi-C data (500M valid read pairs).** Number of ChIA-PET/HiCHIP-supported loop predictions among the top predictions made by RefHiC using reference panels containing 5, 10, 20, 20 low coverage, and 29 Hi-C samples for test chromosomes (chr15, chr16, and chr17), compared against H3K27ac HiCHIP (a), SMC1 HiCHIP (b), RAD21 ChIA-PET (c), and CTCF ChIA-PET (d) data.

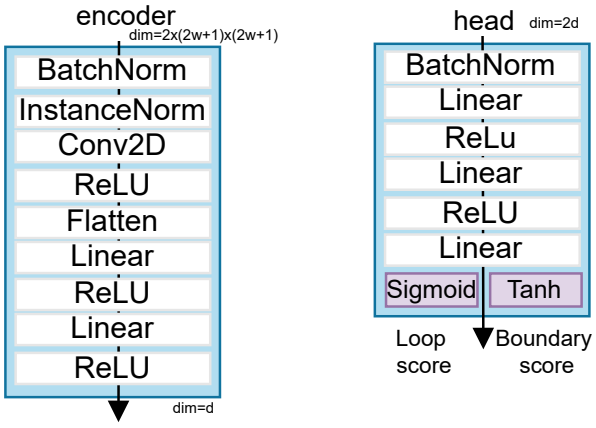


Figure 2.19: **Details of the encoder and head modules.** The task-specific head has different output activation functions for TAD boundary (Tanh) and loop (Sigmoid). w is a hyperparameter for square size, and d is embedding dimension.

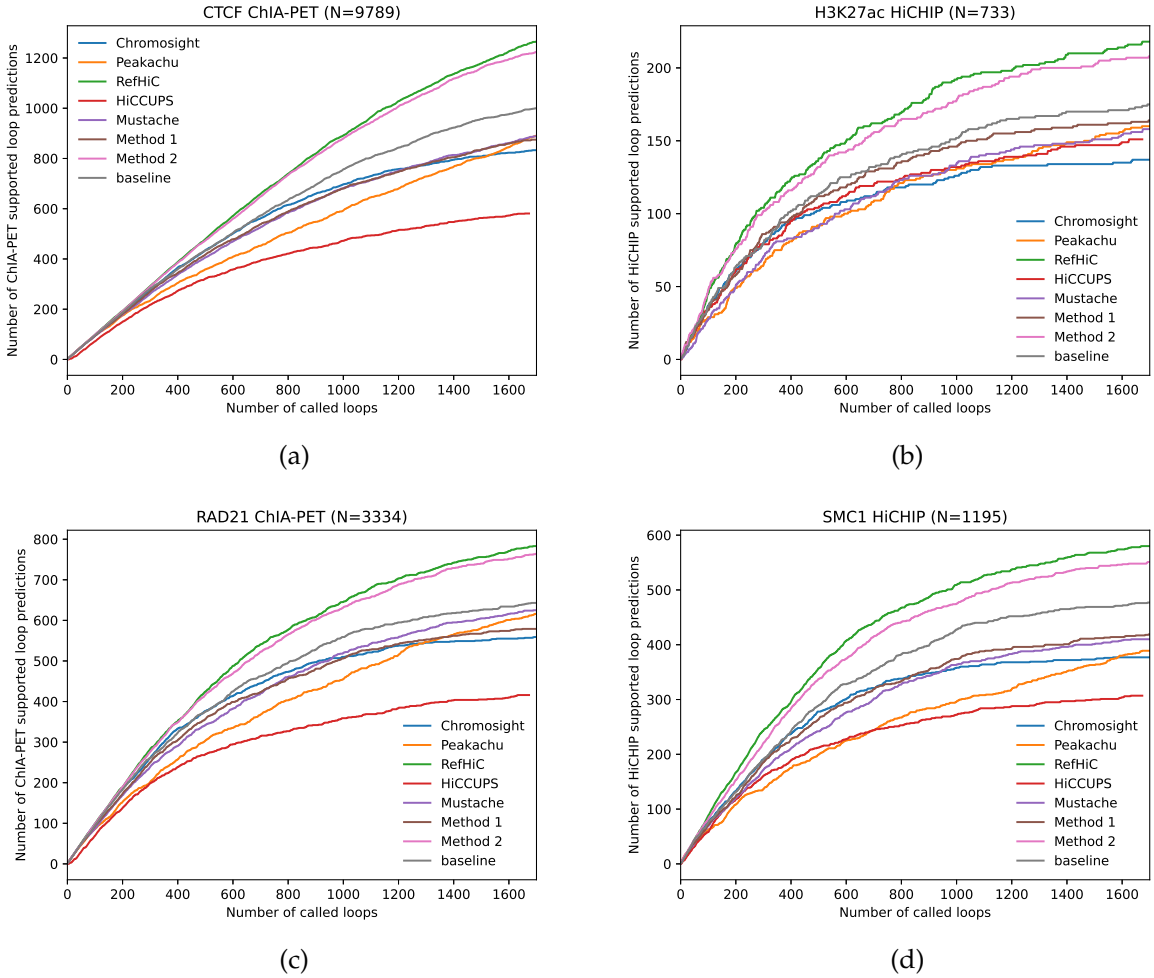
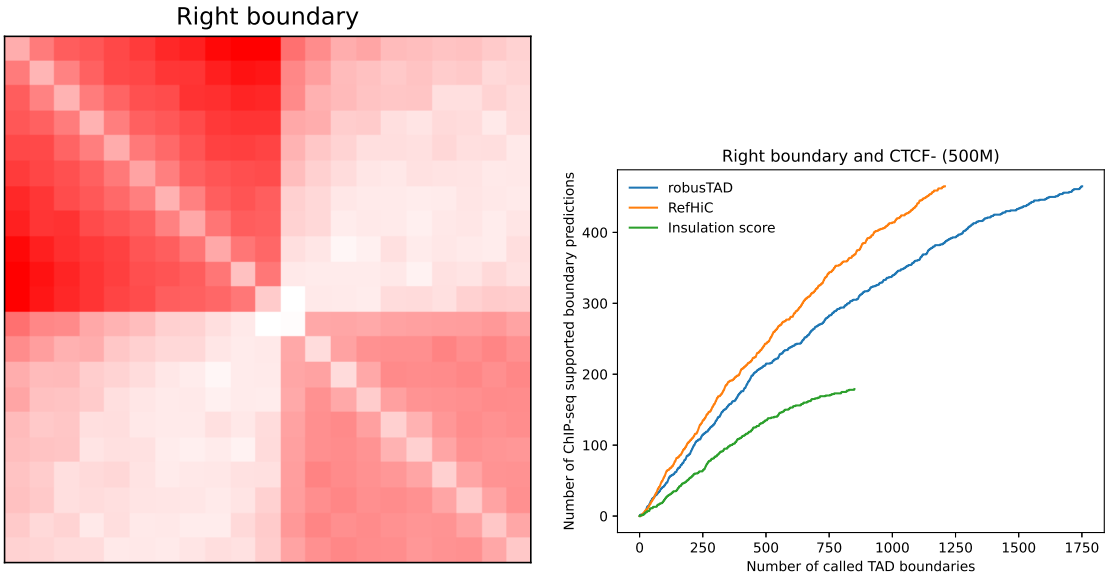


Figure 2.20: **Evaluating RefHiC's ability to identify loops from novel cell types by using GM12878 Hi-C data (500M valid read pairs) as input.** Number of ChIA-PET/HiCHIP-supported loop predictions, among the top predictions made by RefHiC and other tools, for test chromosomes chromosomes 15-17, compared against CTCF (a) or RAD21 (c) ChIA-PET, as well as H3K27ac (b) or SMC1 (d) HiCHIP. The Baseline model is explained in Section 2.11.2. Method 1 and 2 as explained in Section 2.11.4 are used to evaluate RefHiC's ability to identify loops from novel cell types.



(a) (b)

Figure 2.21: **Detection of TAD boundaries on GM12878 Hi-C data (500 valid read pairs). a,** TAD boundary pileups for right (c) boundaries predicted by RefHiC. **b** Number of RefHiC, RobusTAD, and Insulation score predicted right TAD boundaries supported by ChIP-seq identified CTCF binding sites on the reverse (CTCF-) strand.

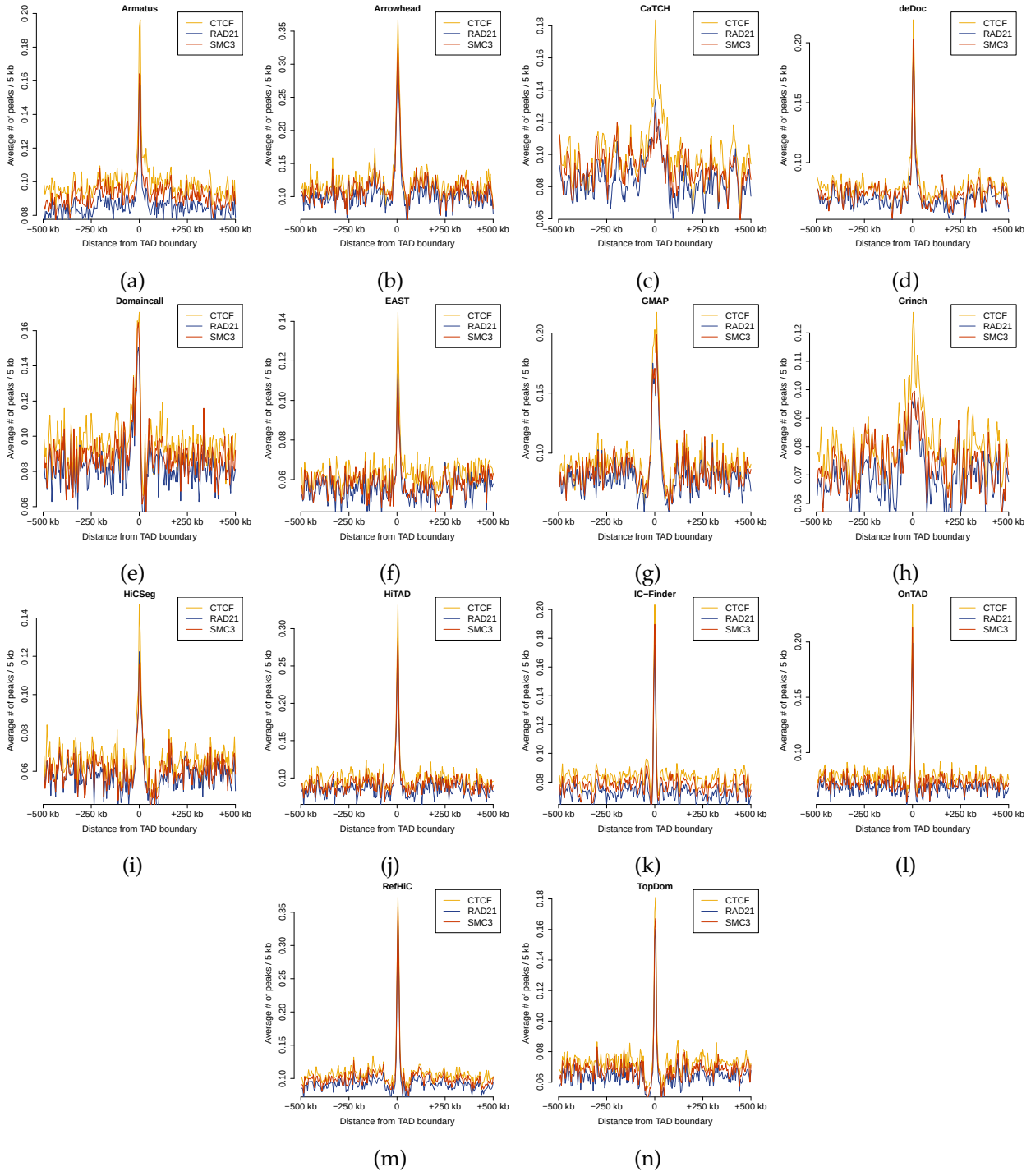


Figure 2.22: **ChIP-seq peak signals for CTCF, RAD21, and SMC3 around TAD boundaries annotated by each tool.**

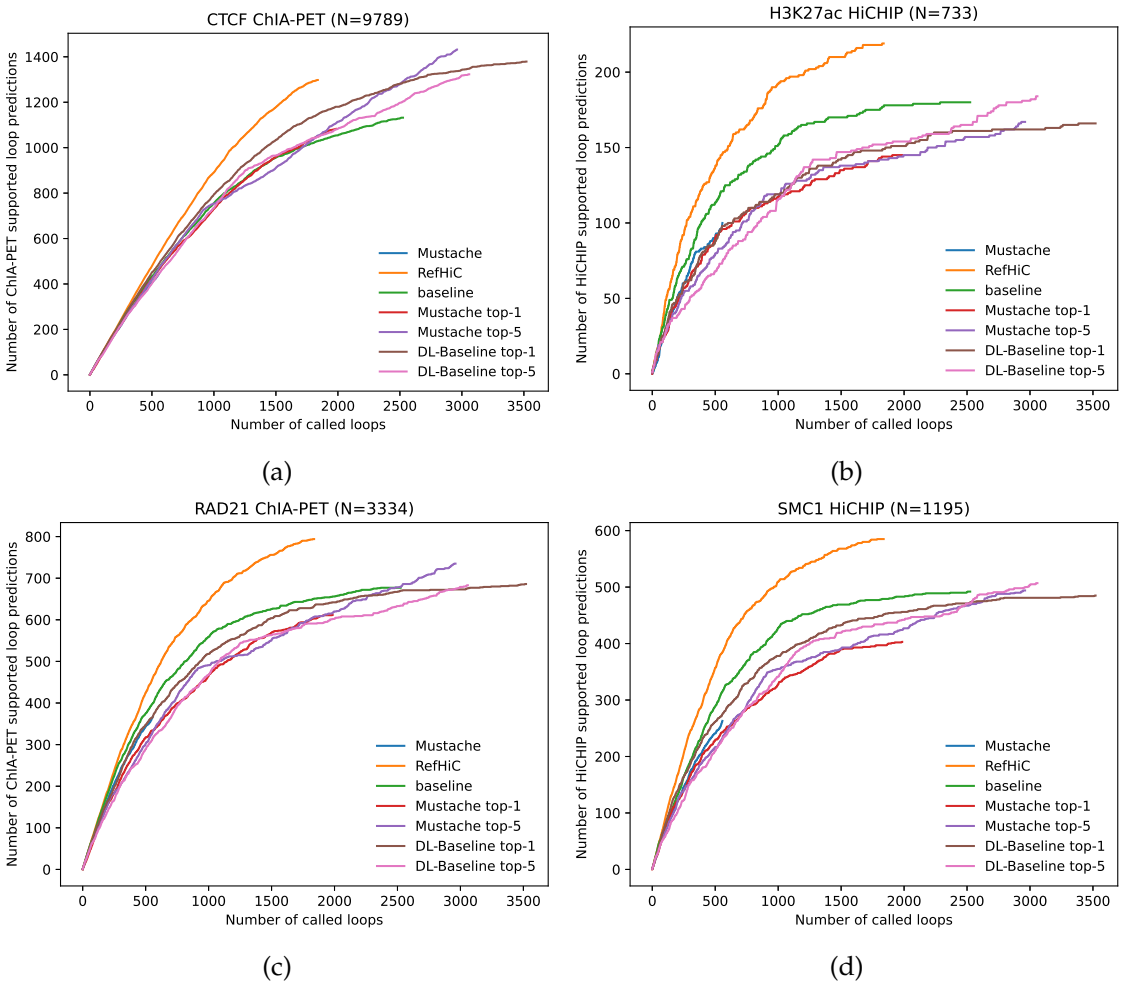


Figure 2.23: **Comparison of loops predicted by RefHiC, baseline, Mustache, and four reference only approaches described in Section 2.11.5 on GM12878 HiC data (500M valid read pairs).** Number of ChIA-PET/HiCHIP-supported loop predictions, among predictions made by each tool for test chromosomes 15-17 compared against CTCF ChIA-PET (a), H3K27ac HiCHIP (b), RAD21 ChIA-PET (c), and SMC1 HiCHIP (d). RefHiC's loop predictions matches all data better than predictions made by alternative methods.

Chapter 3

Reference panel guided super-resolution inference of Hi-C data

Yanlin Zhang, and Mathieu Blanchette *

School of Computer Science, McGill University, Montréal, Québec, H3A 0E9, Canada

Preface

In the previous chapter, we developed RefHiC to annotate chromatin loops and topologically associating domains from a Hi-C contact map of interest. We demonstrated that the introduction of a panel of reference samples improves TADs and loops annotations. Although we can adapt RefHiC to detect other spatial features such as architectural stripes or significant interactions, solving these one by one is not straightforward and tedious. In this chapter, we focus on solving a more fundamental and challenging problem, contact map enhancement, using the reference panel enabled framework. Contact map enhancement aims at predicting a super-resolution contact map that is equivalent to a very high coverage Hi-C contact map from the coverage input. Existing tools are exclusively study sample based and perform poorly at enhancing low-coverage contact maps. Here, we developed RefHiC-SR, a reference panel enabled deep learning application, for

contact map enhancement. Compared to existing approaches, we show that RefHiC-SR improves the accuracy of contact map enhancement, and leads to more accurate topological structure annotation. There are various off-the-shelf tools for topological structure annotation, structure inference, differential analysis, etc. They perform well in handling high-coverage Hi-C data. We believe integrating RefHiC-SR into the stack of computational 3D genomics will allow researchers to fully benefit from the reference panel enabled framework in this field.

The rest of this chapter is the entire text from the following article:
Zhang, Y., & Blanchette, M. (2023). Reference panel-guided super-resolution inference of Hi-C data. Bioinformatics, 39(Supplement_1), i386-i393.

3.1 Abstract

Motivation: Accurately assessing contacts between DNA fragments inside the nucleus with Hi-C experiment is crucial for understanding the role of 3D genome organization in gene regulation. This challenging task is due in part to the high sequencing depth of Hi-C libraries required to support high resolution analyses. Most existing Hi-C data are collected with limited sequencing coverage, leading to poor chromatin interaction frequency estimation. Current computational approaches to enhance Hi-C signals focus on the analysis of individual Hi-C data sets of interest, without taking advantage of the facts that (i) several hundred Hi-C contact maps are publicly available, and (ii) the vast majority of local spatial organizations are conserved across multiple cell types.

Results: Here, we present RefHiC-SR, an attention-based deep learning framework that uses a reference panel of Hi-C datasets to facilitate the enhancement of Hi-C data resolution of a given study sample. We compare RefHiC-SR against tools that do not use reference samples and find that RefHiC-SR outperforms other programs across different cell types, and sequencing depths. It also enables high accuracy mapping of structures such as loops and topologically associating domains.

Availability: https://github.com/BlanchetteLab/RefHiC

Contact: blanchem@cs.mcgill.ca

3.2 Introduction

Technologies such as Hi-C [3], and micro-C [5] capture spatial contacts between DNA fragments in genomes, enabling the inference of various aspects of 3D genome organization. These approaches have revealed a hierarchical spatial organization of topological structures of the genome inside nuclei and spatial patterns such as topologically associating domains (TADs), loops, and compartments. These structures are of vital importance

to gene regulation and are dynamic within cells [113]. Identifying these spatial patterns, especially at high resolution, requires the availability of high-coverage Hi-C sequencing data. The investigation of fine-scale structures would even require ultra-high coverage contact maps [5].

While Hi-C and its variants remain the most popular approaches to map chromatin contacts on a genome-wide scale, the analysis of the data they produce is challenging, in large part due to the moderate sequencing depth (typically 200-500 Million valid read pairs) compared to the size of the contact frequency matrices that need to be estimated. The majority of TAD and loop annotation tools are primarily optimized for high-coverage data and may not provide satisfactory results when applied to typical low- to medium-coverage data, though tools like Grinch [7] have been proposed to analyze low-coverage data. To close the gap, many efforts have been undertaken to perform *in silico* enhancement of Hi-C contact maps [78, 76, 9, 75]. Given a low-coverage Hi-C data set, contact map enhancement tools seek to predict a dense, high-resolution version of the contact map aiming to reproduce the map that would be obtained through very deep coverage sequencing of the same library. Super-resolution enhancement could in theory enable high-resolution analysis of low-coverage Hi-C data, e.g. through the application of third-party analysis tools to enhanced maps.

Most existing contact map enhancement tools are deep learning (DL) approaches and are inspired by super-resolution algorithms in image processing. As a high-resolution (5 kb per bin) contact map for a single human chromosome contains 10,000-50,000 bins, existing applications usually split contact maps into non-overlapping blocks and enhance each block iteratively. HiCPlus [75] was the first DL-based tool proposed for this type of tasks. It is a convolutional neural network (CNN) that contains one hidden layer and is trained from low-coverage and high-coverage contact map pairs (respectively the input and target values) by minimizing the mean square error (MSE) loss. Later, Liu et

al. proposed a deeper CNN with residual connections – HiCNN [76] and trained it following the strategy used in HiCPlus. Similar to super-resolution analysis in computer vision, the MSE loss leads both models to produce blurry predictions [9]. To alleviate the issue of over-smoothness, more recent approaches utilize generative adversarial (GAN) frameworks in model training. For example, HiCGAN [9] is a CNN built upon a generator containing five residual blocks and a discriminator containing three residual blocks. The generator is trained to produce enhanced contact maps from downsampled contact maps, and the discriminator is trained to distinguish high-coverage contact maps from enhanced contact maps. Liu et al. trained HiCGAN with GAN loss and used the generator for prediction. DeepHiC [8] furthers model performance by introducing additional terms to the loss function (i.e., MSE, perceptual loss, and total variance) into training. In contrast, conventional tools [78, 79] usually treat contact map enhancement as imputation. They enhance Hi-C signals by fitting a Markov Random Field or performing random walk on a Hi-C graph.

Although deep learning models have achieved significant successes in contact map enhancement, there is still room for improvement, particularly in the enhancement of very low-coverage contact maps. First, most existing tools are trained on data containing 250M valid read pairs (typically a 16-fold downsampled version of a very high-coverage Hi-C data set produced for human GM12878 cells data by Rao et al. [4]), and can only be effectively used to enhance contact maps containing 200-300M valid read pairs. In addition, similar to super-resolution analysis in computer vision, contact map enhancement is an ill-posed problem as a single low-coverage contact map may correspond to multiple potential high-coverage contact maps. While existing tools can infer high-fidelity predictions, these may not necessarily be correct predictions, especially in sparse regions, potentially leading to false-positives in downstream annotation tasks.

To address the issue of ill-posedness in single-image super-resolution, computer vision researchers have introduced additional images to assist with the prediction task. For example, some studies have created databases of image patches and used them to improve prediction accuracy [169, 170]. In recent years, incorporating external data has become a popular research direction and has been shown to lead to better models with fewer parameters [171]. Within Hi-C data analysis, our recent approach – RefHiC [172] achieves superior performance in annotating topological structures (loops and TADs) from a study sample while using a reference panel of other Hi-C data sets as complement. In reference-based image super-resolution, the reference database is assumed to contain a diverse set of images, regardless of their relationship to the test image. In 3D genome analysis, the conformation of a small region in one cell type may be observed in another cell types [172]. Therefore, to improve the resolution of a small region of a Hi-C contact map, we use contact maps of the same region as a reference.

Here, we introduce RefHiC-SR, a model for enhancing Hi-C contact maps. While RefHiC [172], our model for topological structure annotation, is limited in its ability to learn features from large patches required in a super-resolution task, RefHiC-SR overcomes these limitations by redesigning the encoder as a modified U-net architecture [112], and introducing a multiscale attention mechanism. This novel model allows RefHiC-SR to handle large patches in Hi-C matrices while still benefiting from a reference panel.

3.3 Materials and Methods

3.3.1 RefHiC-SR model architecture

The RefHiC-SR network follows the U-net architecture [112] (Fig. 3.1), originally introduced for image segmentation, to enhance the expressiveness of latent features produced

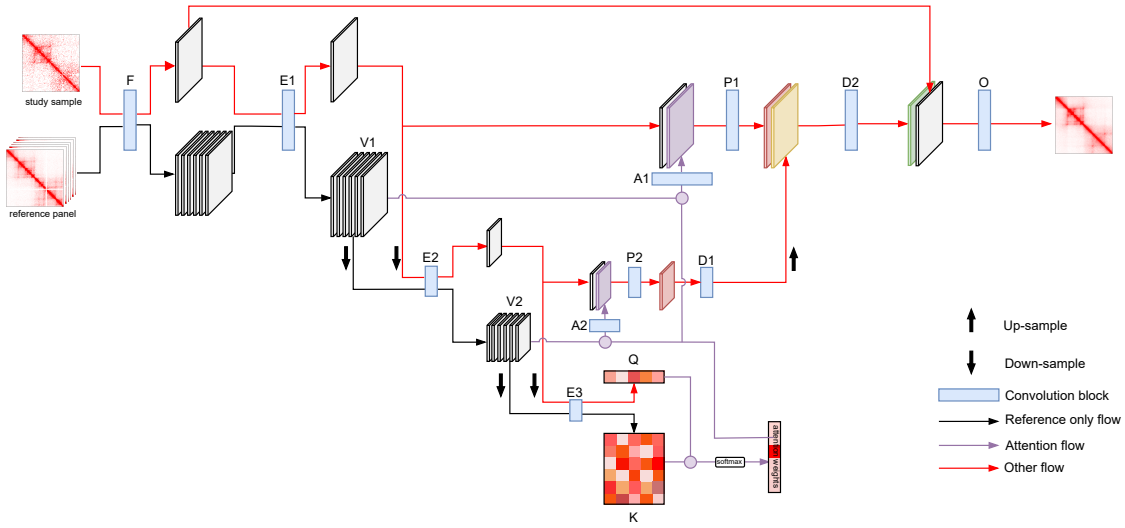


Figure 3.1: **RefHiC-SR architecture.** Overview of the RefHiC-SR neural network for enhancing Hi-C contact maps.

by encoding blocks and enable effective handling of large patches (i.e, 200×200). It contains (i) a low-level feature extraction block (F) that transforms a Hi-C matrix to multi-channel features, (ii) an output block (O) that transforms multi-channel features to an enhanced Hi-C matrix, (iii) multiscale encoding blocks (E1, E2, and E3) that transform low-level features to high-level features at different scales (i.e. keys, values, and query in attention equations 1 and 2), (iv) multiscale decoding blocks (D1 and D2) that transform features at different scales, and (v) attention convolutional blocks (A1 and A2) and projecting layers (P1 and P2) that increasing the model complexity and reduces hidden feature dimensions. To inject information from reference samples into the U-net computation graph, in the forward pass, blocks F and E1-E3 compute multiscale embeddings for the study sample and for the n reference samples. We denote parts of these embeddings as V1 and V2 (values), K (keys), and Q (query). We then compute combined multiscale representations of all reference samples from these embeddings with an attention mechanism. Last, we replace skip connections in U-net [112] with a concatenation of the study sample's embedding and a transformed attention output at the same scale.

F takes an input of dimension $w \times w$, where w is the window size ($w = 200$ at 5 kb

resolution) and projects the input to a $w \times w \times d$ embedding ($d = 24$). It is built with one ReLU-activated convolution layer with d 9×9 filters. E1, E2, and E3 are three consecutive encoding blocks linked by a max pooling operator with a 2×2 kernel and a stride of 2 (i.e. downsampling by 50%). E1, E2, and E3 extract multi-scale features from the input contact maps. E1 is built with two ReLU-activated convolution layers with d 3×3 filters and a dropout layer with rate=0.2 between convolution layers. It takes an input of dimension $w \times w \times d$ and produces an output of the same dimension. E2 is built with the same layers as E1, but it takes an input of dimension $\frac{w}{2} \times \frac{w}{2} \times d$ and produces an output of the same dimension. E3 starts with a batch normalization layer and ends with a flatten layer. It contains three convolution layers: The first two contain d 3×3 filters, and the last contains one 3×3 filter. The first convolution layer in E3 is followed by a dropout layer with rate 0.2 and a max pooling operator with a 2×2 kernel and a stride of 2. We did not use batch normalization in blocks F, E1, and E2 as we observed it introduces artifacts in enhanced contact maps. E3 takes as input the downsampled output of E2 and produces embedding of dimension $1 \times (\frac{w}{8})^2$. The attention module (i.e., purple module) takes as query (Q) and keys (K) the outputs of E3, uses as values (V1 and V2) the outputs of E1 and E2 for the n reference samples. We define the attention weights $\mathbf{ff}$ =softmax$(\mathbf{QK}^T) \in \mathbb{R}^{1 \times n}$, where α_j represents the relative amount of attention paid to sample j in our reference panel when analyzing the study sample. The attention output $\mathbf{a_1}$ and $\mathbf{a_2}$ at two levels are computed as,

$$\mathbf{a_1} = \text{softmax}(\mathbf{QK}^T)\mathbf{V_1} + \text{A1}(\text{softmax}(\mathbf{QK}^T)\mathbf{V_1}) \tag{3.1}$$

$$\mathbf{a_2} = \text{softmax}(\mathbf{QK}^T)\mathbf{V_2} + \text{A2}(\text{softmax}(\mathbf{QK}^T)\mathbf{V_2}) \tag{3.2}$$

where A1 and A2 are convolution blocks for attention outputs configured similarly to E1 and E2 but preceded by layer normalization. P1 and P2 are built with one ReLU-activated convolution layer with one 3×3 filter. They project the concatenation of study sample embeddings (produced by E1 and E2) and attention embeddings (i.e., $\mathbf{a_1}$ and $\mathbf{a_2}$)

to embeddings with d channels. D1 and D2 are built similarly to E2 and E1. D1 takes as input the output of P2; meanwhile, D2 takes as input the concatenation of the output of P1 and the upsampled output of D1. O is built with two ReLU-activated convolution layers with 3×3 filters. It projects the concatenation of the output of F for the study sample and the output of D2 to an enhanced contact map.

3.3.2 Hi-C data and preprocessing

RefHiC-SR's input for an individual sample (i.e., study or reference samples) is defined as a matrix in the shape of $w \times w$, corresponding to the region of interest with a window of size w. w is a hyperparameter set to $w = 200$ at 5kb resolution. We trained RefHiC-SR with ICE-normalized Hi-C contact maps. RefHiC-SR can also take raw data as input, but using raw data directly can lead to worse prediction due to systematic bias. For model training, we used Hi-C data downsampled from the combined GM12878 Hi-C contact map [4]. The length of topologically associating domains and the distance between chromatin loop anchors are usually within 3Mb. Thus, we restricted our analysis to contact pairs separated by at most 3Mb. For inference of an entire chromosome, we will first split a contact map into partially overlapping squares $X_{i,j}$ with width w indicated by top-left corner (i, j) and step $w - 20$ where $j \geq i$. We then apply the trained model to enhance each square. The width of the predicted squares is also w. Last, we extract a $(w - 20) \times (w - 20)$ matrix by trimming each side to address discontinuity between adjacency matrices. The full-chromosome super-resolution contact map is obtained by tiling the super-resolution sub-matrices.

3.3.3 Model training

We trained, evaluated, and tested RefHiC-SR on contact maps downsampled from the combined GM12878 Hi-C data. We used chr11 and chr12 for validation, chromosome 15-17 for testing, and the rest of autosomes for training. After preparing the input data as mentioned above, we collected 6,918, 798, and 813 200×200 blocks for training, validation, and testing. RefHiC-SR takes sub-matrices from the study and reference samples as input in the forward pass. To reduce training computation, we sampled 10 reference samples for each example in each epoch independently. During evaluation, we used all samples in the reference panel. We trained models with a batch size of 46 for 2,000 epochs on an RTX6000 GPU and used AdamW optimizer [132] (weight_decay=0.1; learning rate=1e-3). We also used early stopping to prevent overfitting. In the first 5 training epochs, we warmed up the learning rate from 0 to the initial learning rate (i.e. 1e-3) and then reduced the learning rate to 1e-6 in the first 95% epochs using the cosine annealing learning rate scheduler. Following RefHiC [172], we performed data augmentation by downsampling Hi-C contact maps during training. This transformation preserves topological structures in Hi-C data. Briefly, we downsampled Hi-C training data and stored them on disk in advance. During training, we randomly selected one contact map from these downsampled contact maps for each training example in each epoch independently. We used L1 loss to train RefHiC-SR. It is simple and less prone to be over-smooth.

3.3.4 Contrastive Pretraining

We pre-trained low-level feature extraction block (F) and encoding blocks (E1-E3) by supervised contrastive learning [128] using Hi-C contact maps downsampled from the combined GM12878 Hi-C data. For each training example, we defined items extracted from the downsampled contact maps at the same region as similar items and all Hi-C contact map submatrices in the same batch at other regions as negative items. We aimed to train

these layers such that the distances of embeddings produced by E3 for a training example and its similar items are as close as possible while of embeddings between a training example and its negative items are as far as possible. Following [128], we defined the loss for training instance i as cross-entropy with in-batch negatives

$$l_i = -\log \frac{e^{\text{sim}(\mathbf{h_i}, \mathbf{h_i^+})/\tau}}{e^{\text{sim}(\mathbf{h_i}, \mathbf{h_i^+})/\tau} + \sum_{j \neq i} e^{\text{sim}(\mathbf{h_i}, \mathbf{h_j^-})/\tau}} \tag{3.3}$$

where $\mathbf{h_i}$, $\mathbf{h_i^+}$, and $\mathbf{h_j^-}$ are embeddings: $\mathbf{h_i}$ represents item i, $\mathbf{h_i^+}$ represents one of item i's similar items, $\mathbf{h_j^-}$ represents an item with a label different from i (i.e. negative item). τ is a temperature that controls training, and we set it as 1. We pre-trained the encoder for 20 epochs with the LARS using Adam as a base optimizer. We set batch size to 46 and learning rate to 1e-3 during training.

3.3.5 Evalution metrics

We extensively compared the performance of RefHiC-SR with alternative tools using different metrics, including mean squared error (MSE), mean absolute error (MAE, a.k.a. L1), Pearson correlation coefficient (PCC), Spearman rank correlation coefficient (SRCC), and widely used metric in super-resolution image analysis, including structural similarity index measure (SSIM) score and peak signal to noise ratio (PSNR) score [8] for each of the 200×200 sub-matrix predicted by each tool.

We compared super-resolution to high-resolution Hi-C contact maps with HiCRep [143]. HiCRep measures the reproducibility of two Hi-C experiments by computing a stratified correlation coefficient (PCC) for two contact maps. Its score ranges from -1 to 1 where a high value indicates high reproducibility. In addition to using PCC to compute HiCRep scores, we also computed HiCRep scores with SRCC.

3.3.6 Hi-C data downsampling and Hi-C reference panel

We used the original and 6-level downsampled data of the combined Hi-C contact map for GM12878 cells obtained from Rao et al. [4] to train RefHiC-SR. The reference panel that contains 30 human Hi-C contact map are used to train and evaluate RefHiC-SR. We excluded samples that belong to the study sample's cell type from the reference panel to prevent potential data leakage.

3.3.7 Hyperparameter tuning

We first evaluate the performance of different convolution blocks in RefHiC-SR by adjusting convolution layer numbers in each block and adding residual connections. We configured most blocks with two convolution layers. Following [75, 8], we used 3×3 filters in all internal convolution layers, but tested different filter sizes (i.e., 3×3, 5×5, 9×9 and 13×13) for the first and last convolution layers. We compared validation errors and determined the optimal filter size as 9×9 for both layers. We also compared RefHiC-SR trained with MSE and L1 loss. We observed the model trained with MSE loss overly smooth predictions.

3.3.8 Contact map enhancement with alternative tools

We re-trained HiCPlus, HiCCNN, and DeepHiC with the same data as we used to train RefHiC-SR. Following previous work [75, 76, 8], we trained each model by splitting Hi-C contact maps into 40×40 blocks. To train HiCPlus and HiCNN, we adjusted learning rate and used early stopping to prevent overfitting and set other hyperparameters as default. We trained HiCPlus [75] for 30,000 epochs with a learning rate of $1e$-3 and a batch size of 256. We trained HiCNN [76] and DeepHiC [8] for 1000 epochs with a learning rate of $1e$-4 and a batch size of 256. The maximum training epochs we used are much larger than their original setting, and training losses indicated all models were converged. DeepHiC's

discriminator is too strong to provide gradients to the generator with its original training procedure in our experiment. We changed the discriminative loss weight to 0.0001 and updated the discriminator every ten epochs. Once trained, we applied each model to 200×200 overlapped blocks to enhance a whole contact map. Same as RefHiC-SR, we cropped the prediction into a matrix of 180 × 180.

3.4 Results

We introduce RefHiC-SR, a reference panel-informed deep learning approach for enhancing Hi-C contact maps. Similar to RefHiC [172], it utilizes a reference panel containing 30 high-quality Hi-C data sets from multiple human cell types (Table 3.3) and employs an attention mechanism to determine which reference samples are most relevant for a given $w \times w$ region of the contact map of the study sample. The enhanced contact map at a given region is then inferred based on a combination of the study sample and the attention-weighted reference samples. RefHiC-SR takes as input a typical ICE normalized [173, 49] moderate-coverage (sparse) Hi-C contact map and outputs a high-coverage (dense) contact map prediction. Both input and output contact maps are at high resolution (i.e., 5 kb bins). The resulting prediction is referred to as enhanced or super-resolution contact maps. We divided the human autosomes into a training set (chromosomes 1-10, 13, 14, and 18-22), a validation set (chromosomes 11 and 12), and a test set (chromosomes 15-17). All results reported here pertain only to the three test chromosomes. Although we trained RefHiC-SR on GM12878 cells, the model learned is not cell-type specific. We will demonstrate in a later section that we can use the same model to enhance Hi-C data of other cells.

RefHiC-SR's neural network takes as input a matrix of 200 bins by 200 bins, and outputs a super-resolution matrix of the same dimension. When applying a trained

model to full-chromosome contact map enhancement, we extract from the output matrix a 180×180 matrix by trimming each side to address discontinuity between adjacency matrices. The full-chromosome super-resolution contact map is obtained by tiling the super-resolution sub-matrices. Other super-resolution tools were applied in the same manner.

3.4.1 RefHiC-SR accurately enhances low-coverage contact maps

	PCC			SRCC		
	chr15	chr16	chr17	chr15	chr16	chr17
RefHiC-SR	**0.898±0.001**	**0.865±0.001**	**0.877±0.003**	**0.845±3e-4**	**0.845±3e-4**	0.824±0.001
DeepHiC	0.884±6e-4	0.842±0.001	0.863±0.004	0.833±5e04	0.830±4e-4	0.820±9e-4
HiCNN	0.888±6e-4	0.847±0.001	0.867±0.003	0.844±4e-4	0.844±3e-4	**0.830±8e-4**
HiCPlus	0.865±4e-4	0.823±0.001	0.845±0.003	0.807±5e-4	0.804±5e-4	0.792±0.001
Low coverage (input)	0.643±0.001	0.632±0.002	0.661±0.002	0.559±5e-4	0.564±4e-4	0.558±6e-4

Table 3.1: HiCRep scores between high-coverage and $\frac{1}{16}$ downsampled (low coverage)/enhanced contact maps of GM12878 cells. HiCRep scores are computed with PCC and SRCC metrics. We computed the standard deviations by repeating the analysis 5 times on data downsampled with different random seeds.

We first assessed the contact map enhancement performance of RefHiC-SR, in comparison to three approaches: HiCPlus, HiCNN, and DeepHiC on test chromosomes 15-17 of GM12878 cells. Each model represents one of the three types of deep learning models in contact map enhancement (i.e., shallow model, deep model, and GAN), with DeepHiC featured as the state-of-the-art model in several studies. We used as input a 5-kb resolution Hi-C dataset produced from 250M valid read pairs, obtained by downsampling a Hi-C data set for human GM12878 cells [4]. This is equivalent to a $\frac{1}{16}$ downsampling that most existing tools were trained and evaluated at. As existing models are trained from data at 10kb resolution and with different normalization approaches, it is impractical to benchmark trained models on the same data at 5kb. Thus, we retrained HiCPlus, HiCNN, and DeepHiC with the same set of training data as we used to train RefHiC-SR (see Methods). The accuracy of the enhanced contact maps is assessed by comparing it to the full-coverage contact map, using seven metrics: (i) Mean-Squared Errors (MSE), (ii)

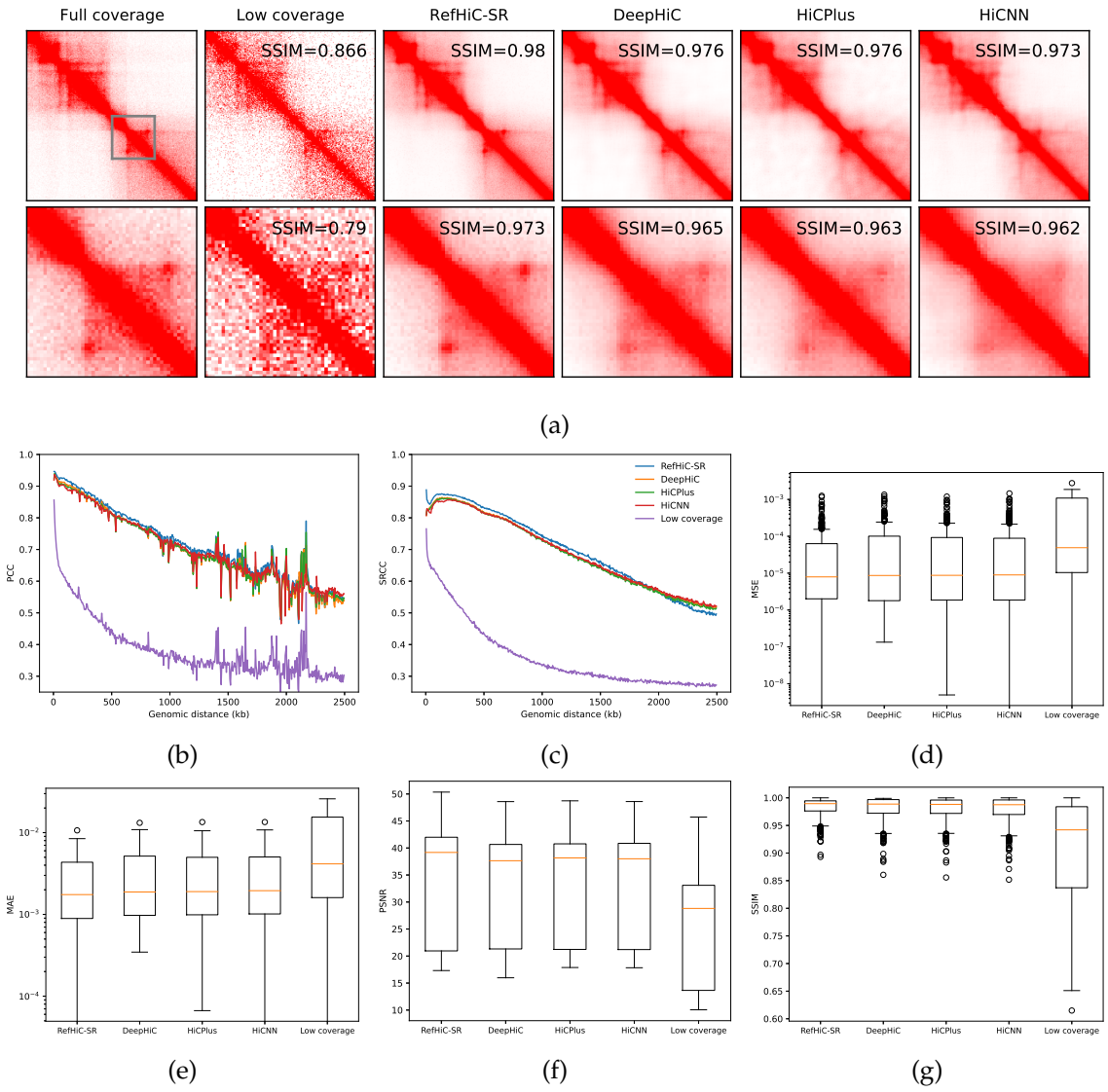
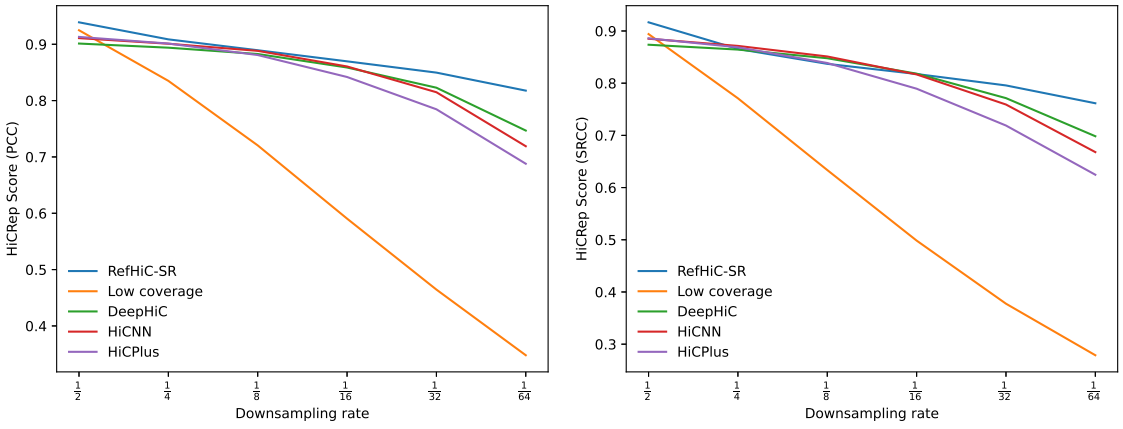


Figure 3.2: Comparison of RefHiC-SR and other tools on GM12878 Hi-C data (250M valid read pairs, test chromosomes 15-17). a. Examples of full coverage, low coverage, and enhanced contact maps on a 1 Mb genomic region (chr17:5000000-6000000) and a zoom in portion. Diagonal-wise PCC (b) and SRCC (c). Boxplots of MSE (d), MAE (e), PSNR (f), and SSIM (g) between full coverage and enhanced contact maps.

Mean-Absolute Error (MAE), (iii) Peak Signal-to-Noise Ratio (PSNR) [9], (iv) the Structural Similarity Index Measure (SSIM) [8], (v) the diagonal-wise Pearson Correlation Coefficient (PCC), (vi) the diagonal-wise Spearman rank correlation coefficient (SRCC), and (vii) the HiCRep score [142] for Hi-C data comparison.

Fig. 3.2a and 3.5 illustrate the full-coverage (target), low-coverage (input), and enhanced contact maps and their differences on a typical 1-Mb genomic region (chr17:5000000-6000000). We observed that all enhanced contact maps better match the full coverage contact map than the low coverage contact map does, with a slightly advantage for RefHiC-SR. RefHiC-SR and DeepHiC are better capturing fine-scale structures such as loops. We then compared the prediction quality on test chromosomes 15-17. The diagonal-wise PCC and SRCC between the enhanced and full-coverage contact maps (Fig. 3.2b,c, 3.14), show that RefHiC-SR is comparable to or outperforms existing tools across all distance ranges. We then compared RefHiC-SR with existing tools at individual 180×180 submatrices. The distributions of MSE (Fig. 3.2d), MAE (Fig. 3.2e), PSNR (Fig. 3.2f), and SSIM (Fig. 3.2g) show that all tools achieve similar performance, with a slight advantage for RefHiC-SR. Last, we compared the similarity of super-resolution and full-coverage contact maps at the whole-chromosome level with HiCRep [142]. Table 3.1 shows HiCRep scores for test chromosomes. It indicates that RefHiC-SR is among the best across test chromosomes.

3.4.2 RefHiC-SR is robust to sequencing depths



(a) (b)

Figure 3.3: Average HiCRep scores from test chromosomes 15-17 from the GM12878 cell line across downsampling ratios $\frac{1}{2}, \frac{1}{4} \ldots \frac{1}{64}$. HiCRep scores are computed with PCC (a) and SRCC (b) metrics.

To benchmark RefHiC-SR's ability to enhance contact maps from Hi-C data at different sequencing depths, we produced downsampled versions (i.e, $\frac{1}{2}$,$\frac{1}{4}$,...,$\frac{1}{64}$, where $\frac{1}{64}$ = 62.5M valid read pairs) of the same GM12878 contact map [4] and applied RefHiC-SR and other tools to enhance contact map resolutions for test chromosomes. We evaluated the accuracy of enhanced contact maps by comparing them against the full-coverage contact maps using HiCRep. Although lower sequencing depths led to less accurate enhancement for all tools (Fig. 3.3), RefHiC-SR was most robust to low sequencing depths, clearly outperforming other tools at very low coverage ($\frac{1}{32}$=125M and $\frac{1}{64}$=62.5M). For Hi-C data containing less than 250M valid read pairs, RefHiC-SR can produce predictions comparable to the second best tool using only half of the read pairs. To study the performance of each tool at the most extreme case (i.e., $\frac{1}{64}$ downsampled data), we repeated the battery of tests originally performed at $\frac{1}{16}$ = 250M downsampled data (Fig. 3.6, 3.14). We observed RefHiC-SR performed the best on all metrics.

3.4.3 RefHiC-SR performs well across cell types

	IMR90			K562		
	chr15	chr16	chr17	chr15	chr16	chr17
RefHiC-SR	**0.868**	**0.850**	**0.878**	**0.813**	**0.817**	**0.825**
DeepHiC	0.858	0.839	0.866	0.799	0.810	0.812
HiCNN	0.861	0.842	0.870	0.802	0.804	0.811
HiCPlus	0.858	0.840	0.866	0.812	0.817	0.824
Low coverage (input)	0.716	0.697	0.722	0.788	0.787	0.806

Table 3.2: HiCRep scores between high-coverage and low-coverage/enhanced contact maps of IMR-90 and K562 cells. HiCRep scores are computed with PCC metrics.

Next, we aimed to assess the performance of RefHiC-SR and other tools, which were all trained on Hi-C data obtained from GM12878 cells, on data from other cell types. We applied each model to the enhancement of Hi-C data from IMR90 and K562 [4] cell lines (test chromosomes 15-17 only). We used HiCRep, MAE, MSE, PSNR, and SSIM to evaluate model performance. Table 3.2 shows that RefHiC-SR outperformed other tools by achieving the highest HiCRep scores in both cells. Fig. 3.12 and 3.13 show that

RefHiC-SR outperformed or was comparable to other tools in both cells as evaluated by super-resolution image analysis metrics.

3.4.4 RefHiC-SR enables improved loop and TAD boundary annotation

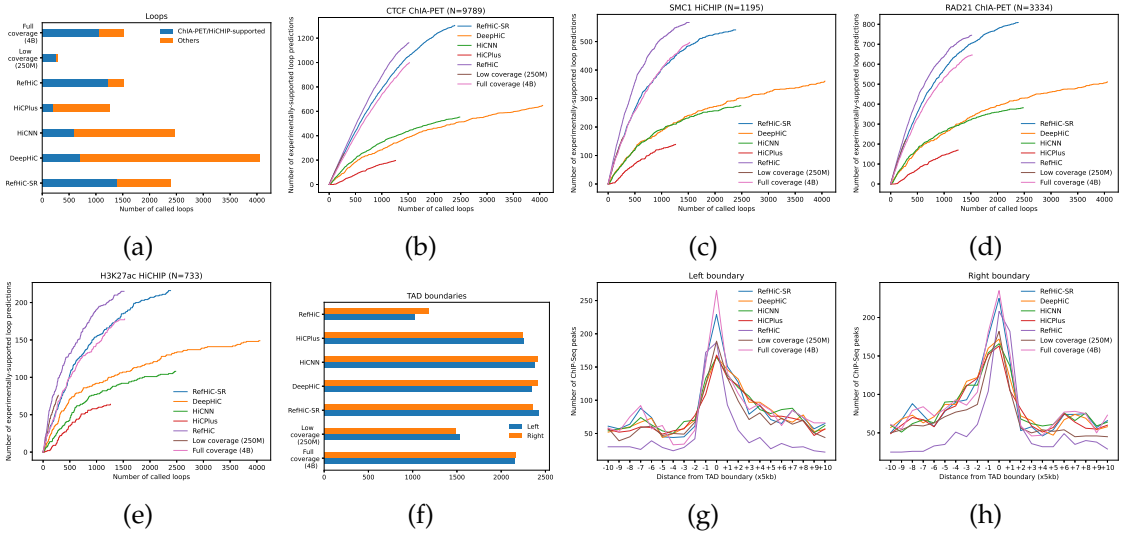


Figure 3.4: Comparison of loops and TADs annotated from low coverage, full coverage, and enhanced contact maps. (a) Number of loop annotations. (b-e) Number of ChIA-PET/HiCHIP-supported loop predictions compared against CTCF ChIA-PET (b), SMC1 HiCHIP (c), RAD21 ChIA-PET (d), and H3K27ac HiCHIP (e). Occupancy of ChIP-seq identified CTCF binding site as a function of distance to left (g) and right (h) boundary annotations.

RefHiC-SR and other resolution enhancement tools are meant to ease downstream analyses, such as TAD and loop annotation, by imputing missing signals in contact maps. Here we show that RefHiC-SR facilitates annotating TAD and loop with off-the-shelf annotation tools, without introducing many false positives.

We first assessed the ability of RefHiC-SR to produce enhanced maps that enable high-accuracy loop prediction. We applied Mustache [18] to annotate loops from the original full-coverage GM12878 contact map, the $\frac{1}{16}$ downsampled contact map, and enhanced contact maps produced by different tools. For each analysis, we set the same Mustache 5%

FDR cutoff, keeping other parameters as default, and sorted loops by Mustache-reported FDR. We also included predictions made by RefHiC [172] for comparison. The number of predicted loops is quite different among different inputs, with DeepHiC leading to the largest number of predictions (Fig. 3.4a). We then evaluated predicted loops by comparing them to loops identified by loop-targeting experimental data (ChIA-PET on CTCF [118] and RAD21 [119], and HiCHIP on SMC1 [37] and H3K27ac [120]), allowing up to a 5 kb shift (Fig. 3.4b-e). When applied to enhanced contact maps, Mustache produced 1,245 CTCF-supported loops, 761 RAD21-supported loops, 512 SMC1-supported loops, and 197 H3K27ac-supported loops from RefHiC-SR enhanced contact maps. These numbers exceed those obtained on enhanced maps produced by other tools by 25%-550%, and even those obtained on the full-coverage data itself. The accuracy of loop annotated from RefHiC-SR enhanced contact maps is comparable to annotating loops from the full-coverage data. In contrast, contact maps enhanced by alternative tools introduced a large number of false positive loop predictions. The combination of RefHiC-SR and Mustache was only beat by RefHiC slightly. We next evaluated the extent to which RefHiC-SR facilitates loop annotation from Hi-C contact maps containing different numbers of valid read pairs. Unexpectedly, coverage reduction leads to an increase in the number of loops being predicted on most tools' enhanced maps (including RefHiC-SR enhanced, Fig. 3.7). Fig. 3.8 shows that as the coverage drops from 256M to 62.5M valid read pairs, the number of experimentally-supported predicted loops remains similar when Mustache is applied to RefHiC-SR enhanced contact maps, but drops significantly using enhanced maps produced by other super-resolution tools.

To evaluate RefHiC-SR's usefulness for facilitating TAD annotation, we used RobusTAD [115] to annotate TAD boundaries from the same set of contact maps as above. We also included predictions made by RefHiC [172] for comparison. RefHiC made the least

predictions. The number of predicted TAD boundaries is similar among full-coverage Hi-C data and super-resolution inputs (Fig. 3.4f). Fig. 3.4g,h show that RobusTAD identified a similar total number of CTCF-supported TAD boundaries from RefHiC-enhanced and full-coverage contact maps. In contrast, contact maps enhanced by alternative tools lead to fewer CTCF-supported TAD boundaries. Annotating TAD boundaries from contact maps enhanced from low-coverage data shows that boundary annotation is robust to sequencing coverage (Fig. 3.9,3.10). Fig. 3.10 shows that at very low coverage, RefHiC-SR can still help to identify a large number of CTCF-supported TAD boundaries.

We then repeated the analysis on Hi-C datasets for K562 and IMR-90 cells. Fig. 3.16 and 3.17 show RefHiC-SR outperformed alternative methods in both loop and TAD annotations.

3.4.5 RefHiC-SR implementation

RefHiC-SR is a Python program available at https://github.com/BlanchetteLab/RefHiC with scripts to reproduce our experiments. We implemented the neural network with the PyTorch library [138]. RefHiC-SR can run on either a CPU or GPU, but it performs three times faster on GPU. RefHiC-SR requires at least 3GB of storage space for saving reference panel data and at least 15GB RAM for loading reference samples during prediction. RefHiC-SR is efficient and can process the longest human chromosome within three minutes when run on a GPU.

3.5 Discussion and Conclusion

Here we present RefHiC-SR, a deep learning framework that utilizes a reference panel to facilitate enhancing Hi-C data resolution for a given study sample. In contrast, existing contact map enhancement algorithms are exclusively study-sample based, and hence

their ability to reliably enhance contact maps from typical sequencing depth Hi-C data is limited. Our extensive evaluation demonstrated that RefHiC-SR outperforms existing tools in data sets ranging from very high to very low sequencing coverage, with the most striking improvements observed in the latter case. RefHiC-SR also outperformed a simple global similarity based baseline (Section 3.9.2), indicating the necessity of designing this model to incorporate the reference panel. Although RefHiC-SR is a machine-learning model trained primarily on GM12878 Hi-C data, the same trained model is effective on different cell types, and at different levels of coverage. Comparison between RefHiC-SR and a Baseline model similar to RefHiC-SR but lacking a reference panel shows RefHiC-SR's superior to the introduction of a reference panel (Section 3.9.1). The super-resolution contact maps predicted by RefHiC-SR are ready for downstream analysis and do not introduce a significant number of false positives. In contrast, other enhancement tools often introduce false positive annotations in downstream analysis.

Across the different sub-fields of data-driven biology, researchers have developed many reference panel enabled approaches to aid the analysis of a study sample. RefHiC-SR is the first approach to enable this type of reference panel based analysis of 3D contact map enhancement. In addition, RefHiC-SR is the only contact map enhancement model that is robust to low sequencing coverage. We believe RefHiC-SR has the potential to become an essential method for enhancing Hi-C contact maps, paving the way to further our understanding of 3D genome organization and functional implications at a finer scale.

3.6 Acknowledgements

This work was funded by Genome Quebec/Canada and a Genome Quebec/Oncopole/IVADO grants to M.B., and FRQNT Doctoral (B2X) Research Scholarships to Y.Z..

3.7 Contributions

Y.Z. and M.B. conceived the study. Y.Z. performed analysis. M.B. supervised the project. Y.Z. and M.B. wrote the manuscript. All authors read and approved the final paper.

3.8 Competing interests

The authors declare no competing interests.

3.9 Appendix

Supplementary Information

3.9.1 Comparison with a U-Net Baseline

We built a U-Net model named Baseline to demonstrate the superior performance achieved by RefHiC-SR is attributed to the use of a reference panel. This Baseline model is similar to RefHiC-SR but does not use a reference panel. We followed RefHiC-SR's training procedure to train Baseline. Fig. 3.8,3.10,3.11 show that RefHiC-SR outperformed Baseline in enhancing contact maps containing different numbers of valid read pairs. These results suggest that RefHiC-SR benefits from the introduction of a reference panel.

3.9.2 RefHiC-SR outperforms a simple top-K averaging baseline

Although RefHiC-SR has demonstrated outstanding performance when evaluated by various methods, it remains uncertain whether this superiority is attributable to the proposed local similarity-based U-Net model. It is interesting to investigate the performance of predicting as the average of the study sample and top-K most similar Hi-C reference samples. To establish a baseline for comparison, we employed a method, **Baseline (top 5)**, based on the top 5 most similar reference samples, which were identified by comparing the study sample against each reference sample using HiCRep. Subsequently, we calculated the average interaction frequency of the study sample and the selected reference samples to make our prediction. Fig. 3.15 shows that Baseline (top 5) is much worse than RefHiC-SR.

ID	Sample	Valid read pairs	Source
HIC00001	22Rv1 (prostate cancer cell line)	685096962	[144]
HIC00002	293TRex-Flag-BRD4-NUT-HA	1660296754	[145]
HIC00007	BLaER (lymphoblastic leukemia cell line)	406322404	[146]
HIC00041	HCT-116 (colorectal cancer cell line)	603179292	[147]
HIC00067	HeLa Kyoto cell, MboI G1 sync control	1045885928	[148]
HIC00090	HepG2 (hepatocellular carcinoma cell line)	1759654311	[119]
HIC00091	HL60/S4 (neutrophil-like Myeloid leukemia cell line)	478434139	[149]
HIC00113	Nalm6 (B cell precursor leukemia cell line)	816274711	[150]
HIC00168	WI38_RAF (WI-38hTERT/GFP-RAF1-ER)	602556180	[151]
HIC00172	Embryonic stem cell, Cardiomyocyte differentiation : hESCs (day 0)	1914642484	[152]
HIC00183	teloHAEC (endothelial cell line)	911486437	[153]
HIC00200	Naïve human embryonic stem cells	731906045	[154]
HIC00203	GM23248 (primary skin fibroblasts)	1797370277	[155]
HIC00221	MDM (monocyte-derived macrophages)	590449106	[156]
HIC00269	Astrocytes of the cerebellum primary cell	430822244	[119]
HIC00273	HAP1 (near-haploid cell line)	413436528	[157]
HIC00280	Purified human germinal center B cells	426222299	[158]
HIC00287	Liver	447028100	[159]
HIC00295	Thymus	507309033	[159]
HIC00296	H1 Embroynic Stem Cell	989388439	[160]
HIC00310	A549 00h 100 nM dexamethasone	1548684355	[119]
HIC00318	HUVEC	438880295	[161]
HIC00319	IMR90	1053932182	[161]
HIC00320	K562	880877579	[161]
HIC00321	KBM7	877658969	[161]
HIC00322	NHEK	653628335	[161]
HIC00337	Gastric tissue	426476775	[162]
HIC00343	Left Ventricle	547477074	[162]
HIC00354	Spleen	490487515	[162]
HIC00360	GM12878	1994319522	[161]

Table 3.3: **Human reference panel**

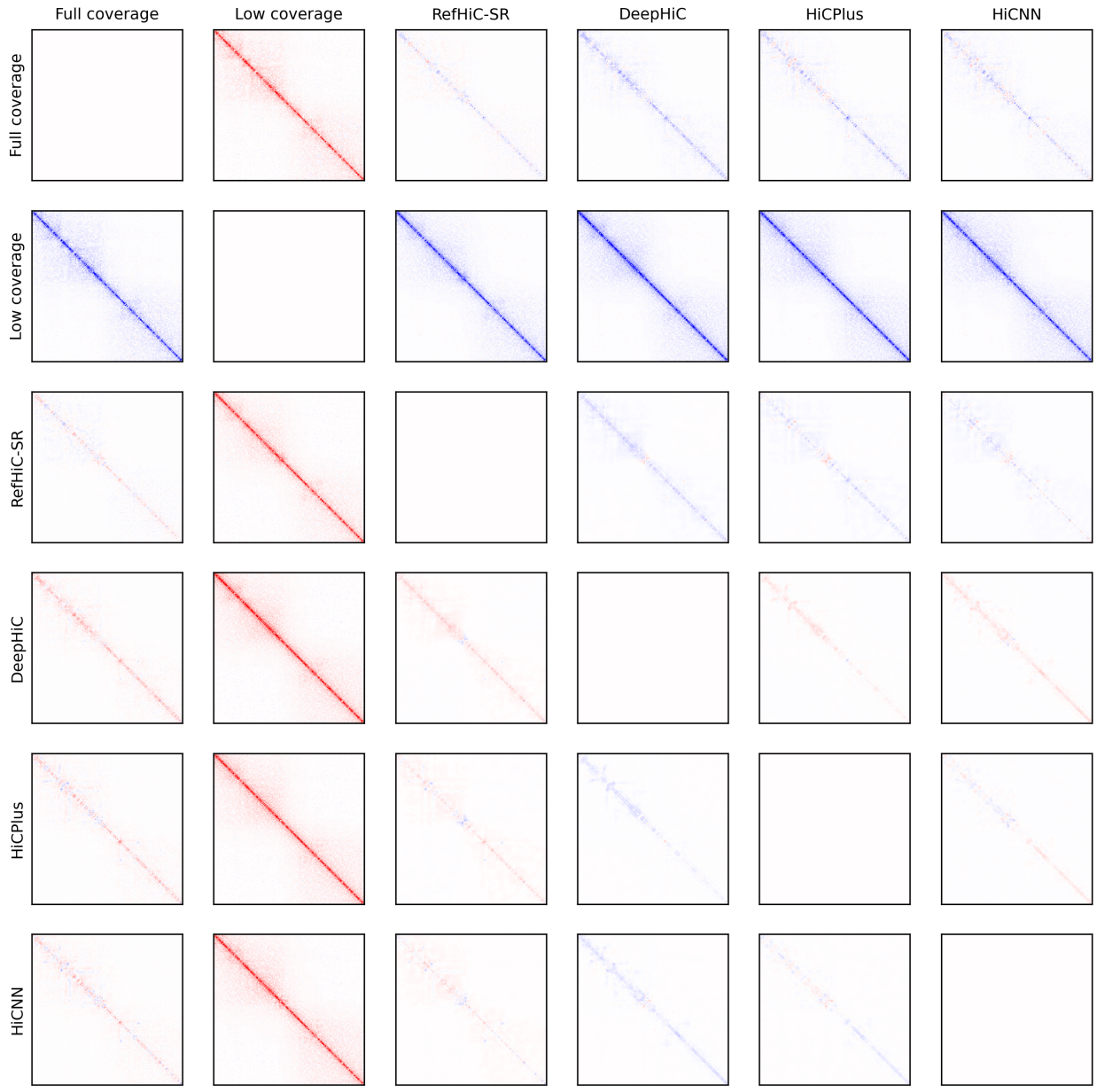


Figure 3.5: Pairwise difference among low-coverage, full-coverage, and enhanced contact maps on a 1 Mb genomic region (chr17:5000000-6000000). We clipped values to the range of [-0.5,0.5] for better visualization.

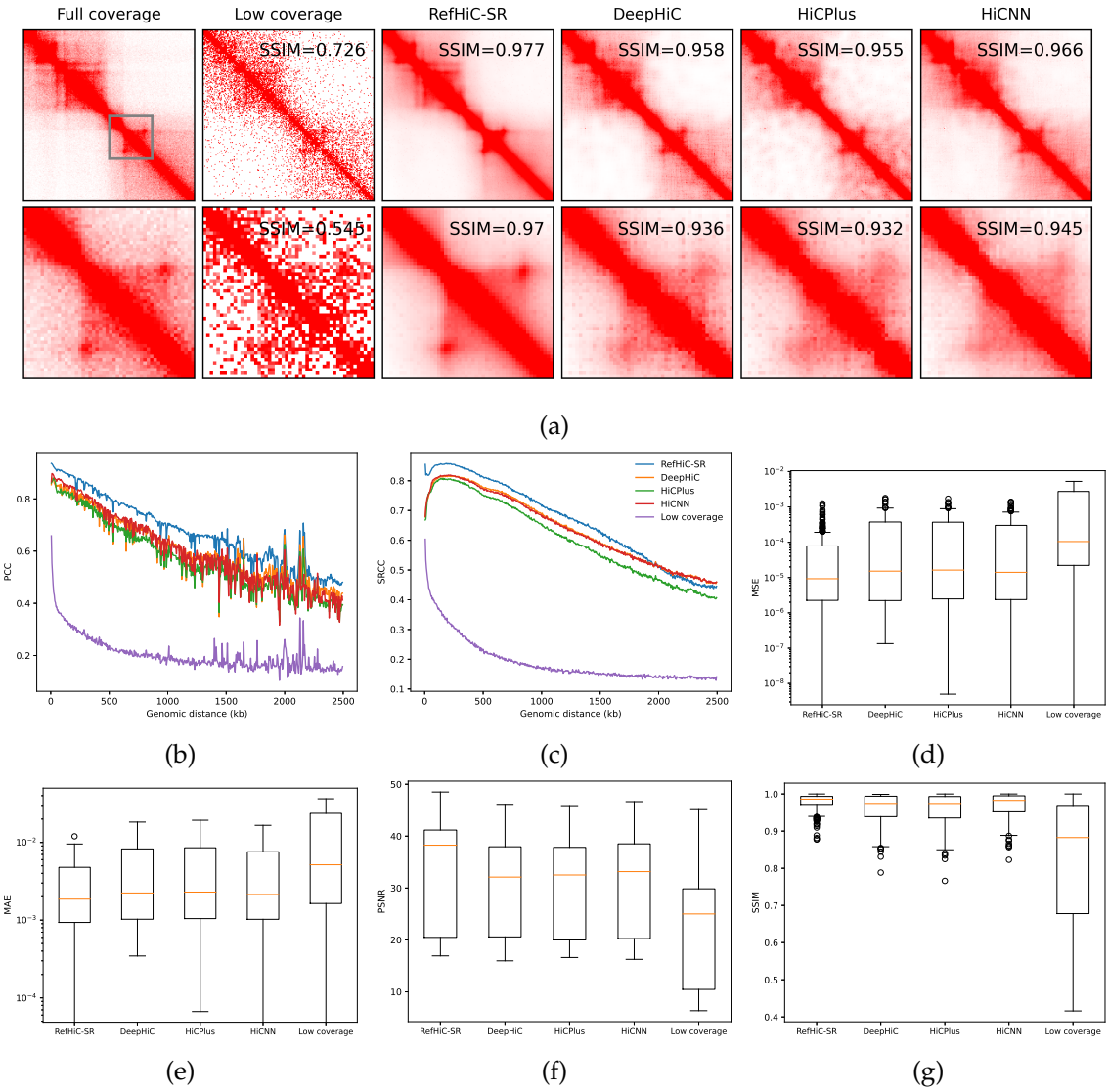


Figure 3.6: Comparison of RefHiC-SR and other tools on GM12878 Hi-C data (62.5M valid read pairs, test chromosomes 15-17). a. Examples of low-coverage, full-coverage and enhanced contact maps on a 1 Mb genomic region (chr17:5000000-6000000). Diagonal-wise PCC (b) and SRCC (c). Boxplots of MSE (d), MAE (e), PSNR (f), and SSIM (g) between full-coverage and enhanced contact maps.

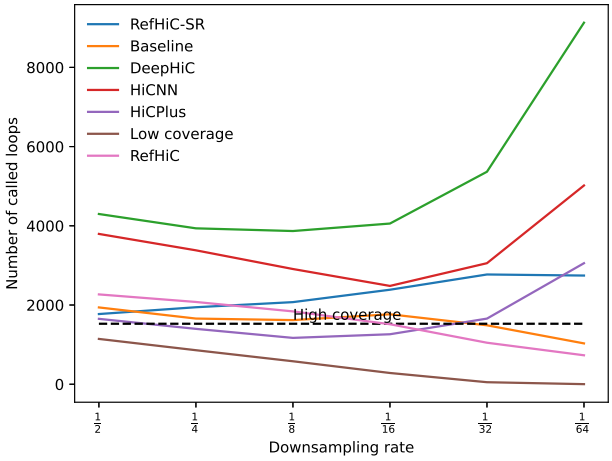


Figure 3.7: Number of called loops predicted from data across different downsampling rates.

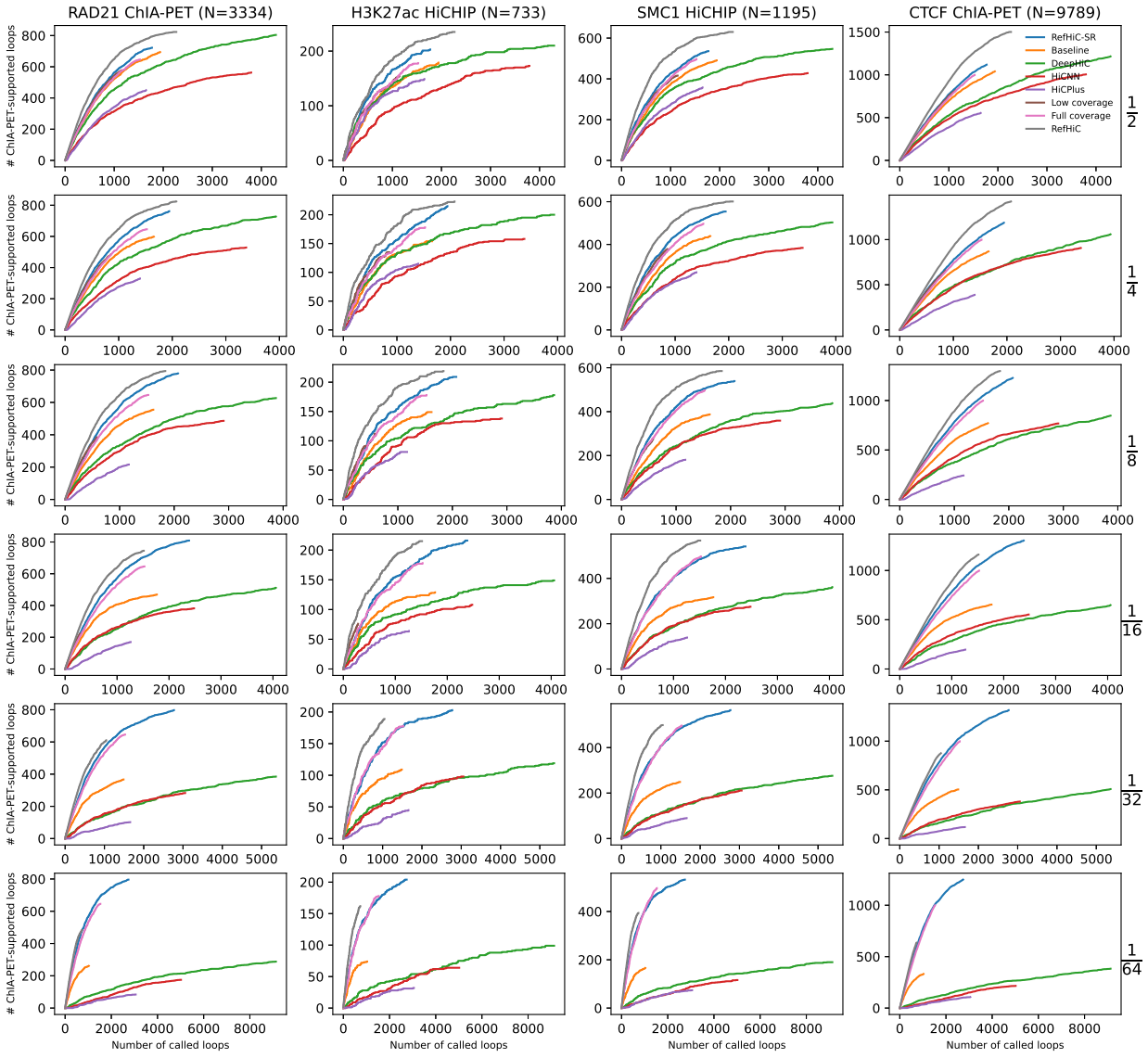
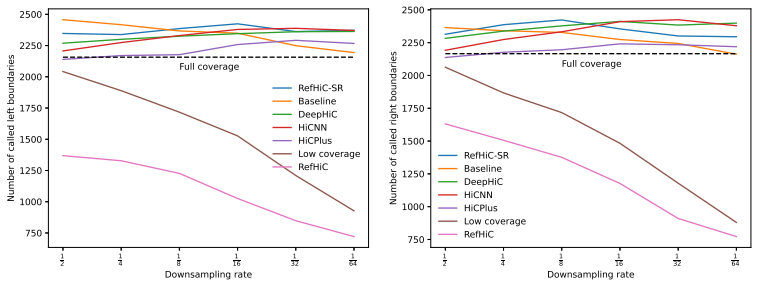
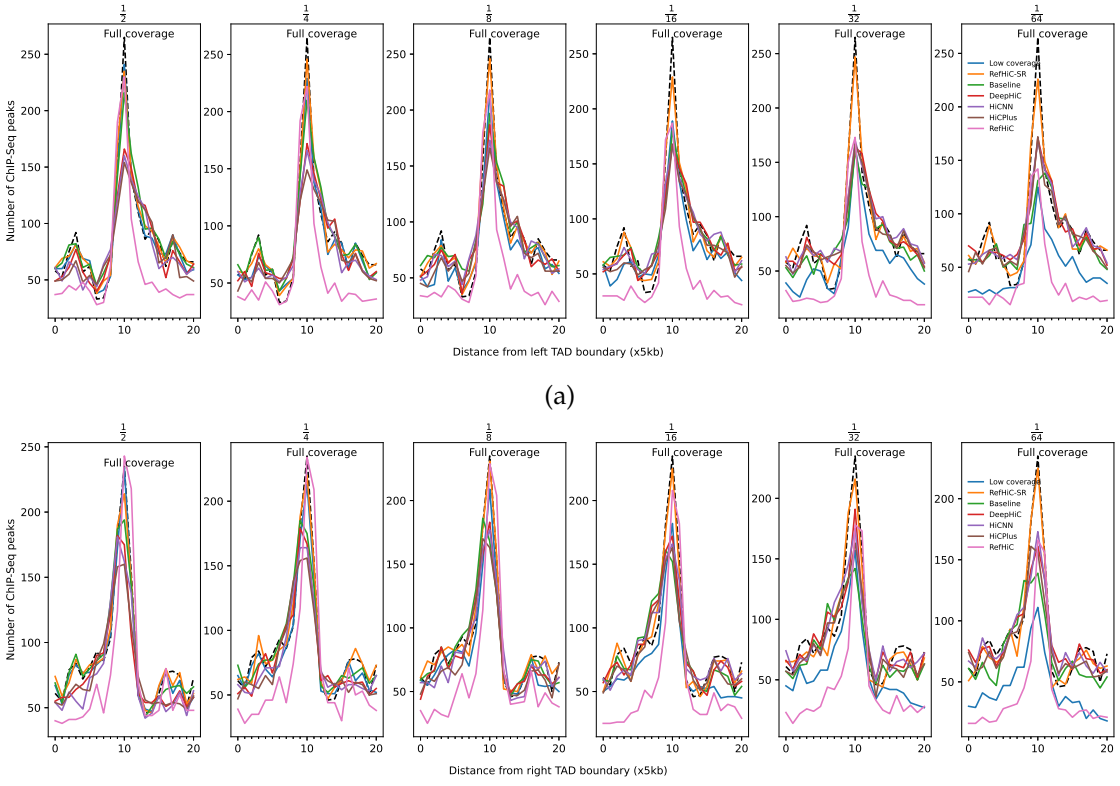


Figure 3.8: Comparison of loops annotated from low-coverage, full-coverage, and enhanced contact maps. This figure is similar to Fig. 3.4b-e but for contact maps at different downsampling rates.



(a) (b)

Figure 3.9: Number of called left (a) and right (b) boundaries predicted from data across different downsampling rates.



(a)

(b)

Figure 3.10: Comparison of TADs annotated from low-coverage, full-coverage, and enhanced contact maps. This figure is similar to Fig. 3.4g,h but for contact maps at different downsampling rates.

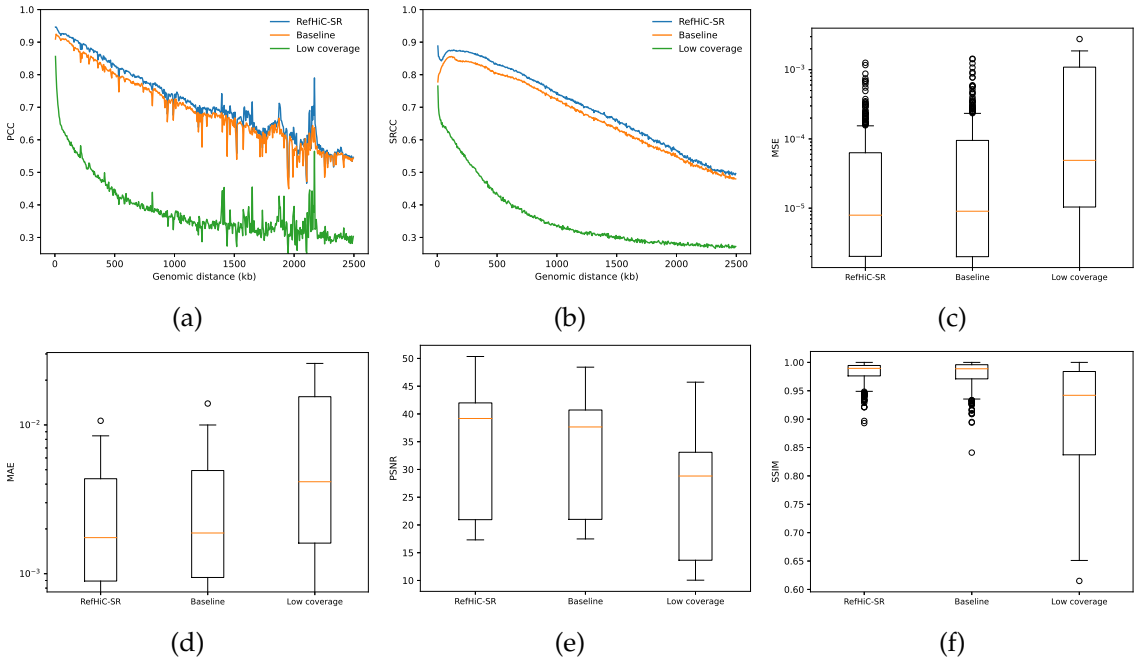


Figure 3.11: Comparison of RefHiC-SR and Baseline on GM12878 Hi-C data (250M valid read pairs, test chromosomes 15-17). These figures are similar to Fig. 3.2 but compared RefHiC-SR against Baseline (Section 3.9.1).

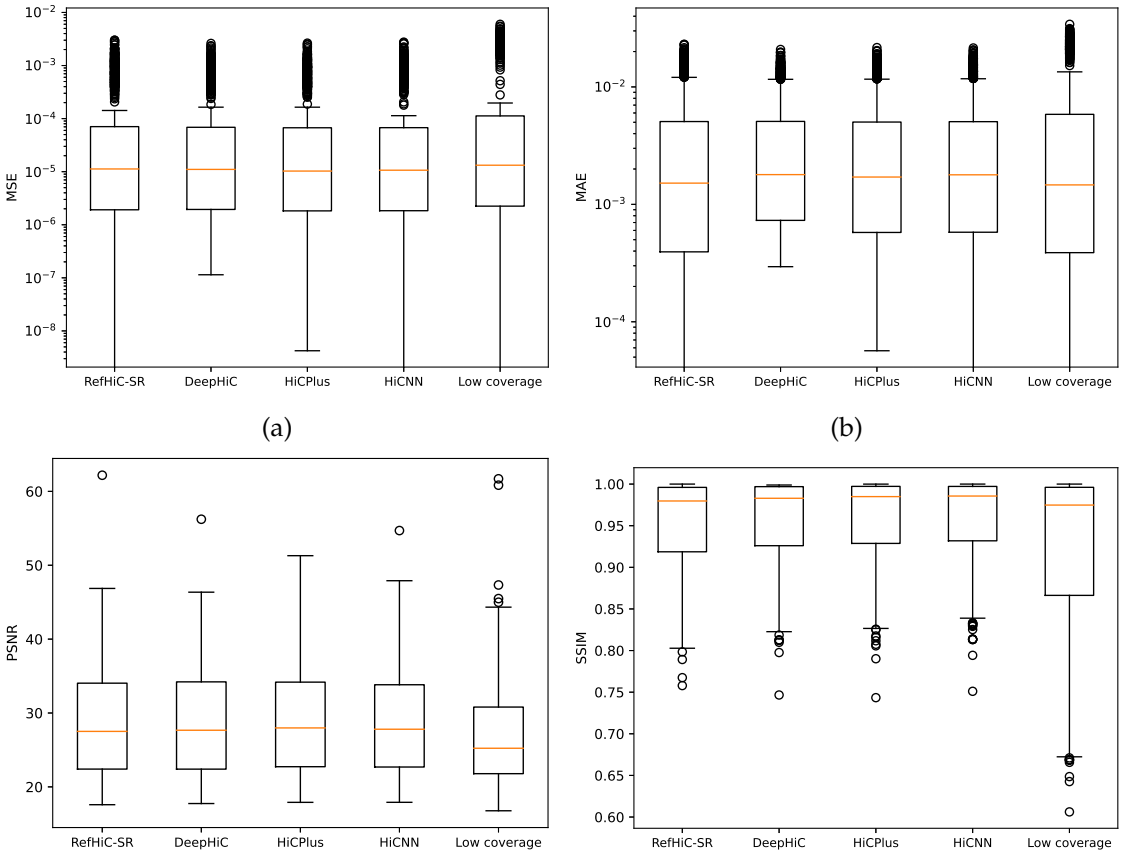


Figure 3.12: Comparison of RefHiC-SR and other tools on IMR-90 Hi-C data (test chromosomes 15-17). Boxplots of MSE (a), MAE (b), PSNR (c), and SSIM (d) between full-coverage and enhanced contact maps.

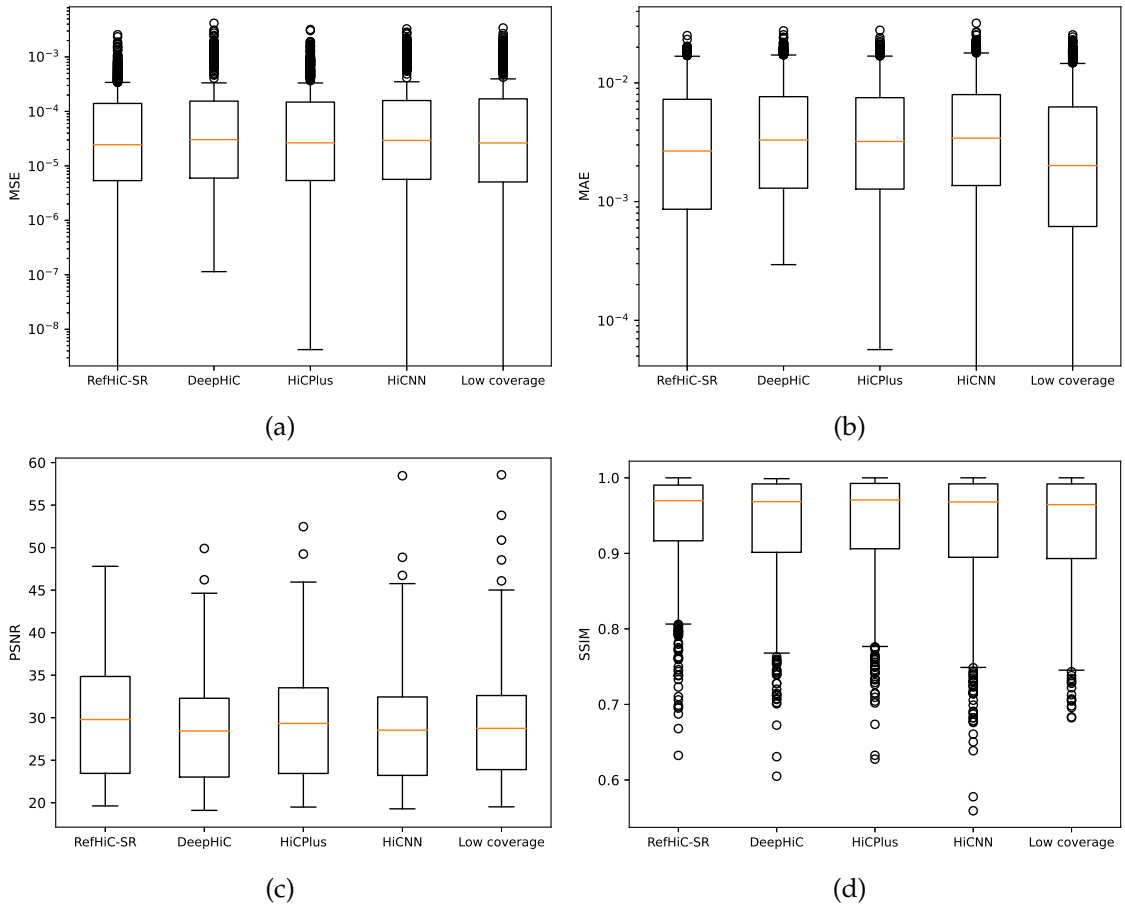
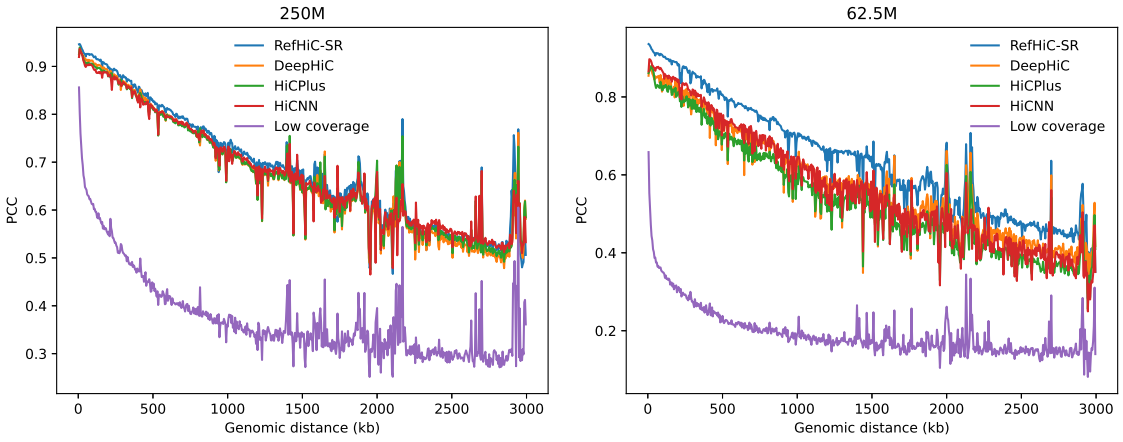
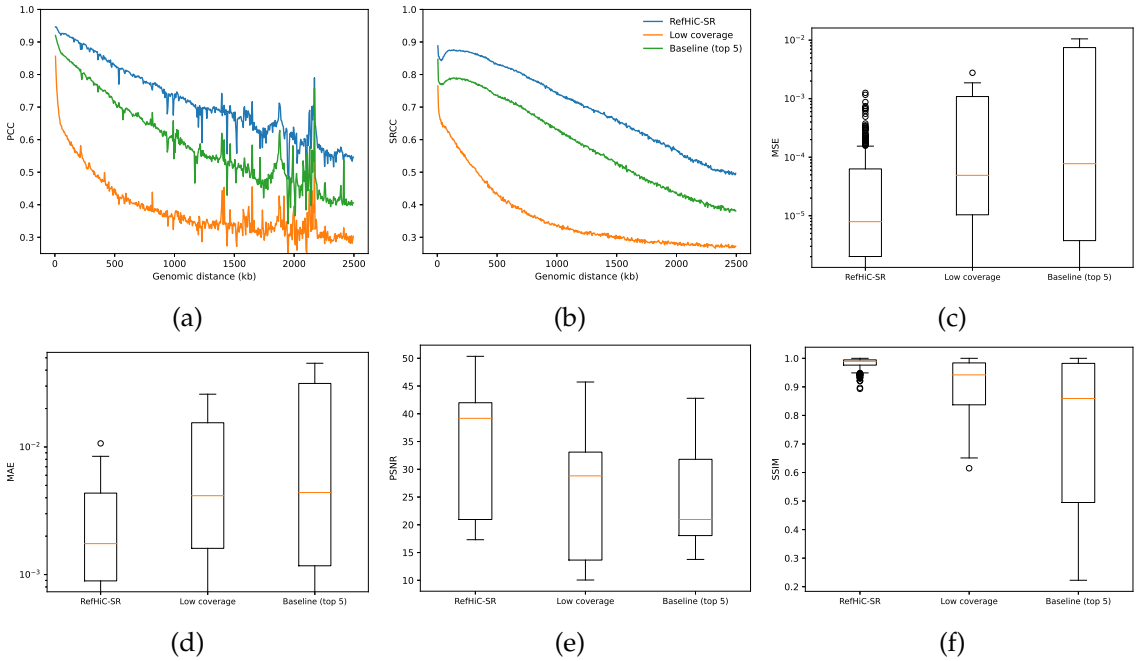


Figure 3.13: Comparison of RefHiC-SR and other tools on K562 Hi-C data (test chromosomes 15-17). Boxplots of MSE (a), MAE (b), PSNR (c), and SSIM (d) between full-coverage and enhanced contact maps.



(a) (b)

Figure 3.14: Comparison of RefHiC-SR and other tools on GM12878 Hi-C data (test chromosomes 15-17, within 3Mb distance). Diagonal-wise PCC between the full coverage contact map and a contact map enhanced from a Hi-C dataset that containing 250M valid read pairs (a), and 62.5M valid read pairs (b). These are the same figures as Fig. 3.2b and 3.6b, but including more long range interactions.



(a) (b) (c)

(d) (e) (f)

Figure 3.15: Comparison of RefHiC and Baseline (top 5) described in Section 3.9.2 on GM12878 HiC data (500M valid read pairs). Diagonal-wise PCC (a) and SRCC (b). Boxplots of MSE (c), MAE (d), PSNR (e), and SSIM (f) between full coverage and enhanced contact maps.

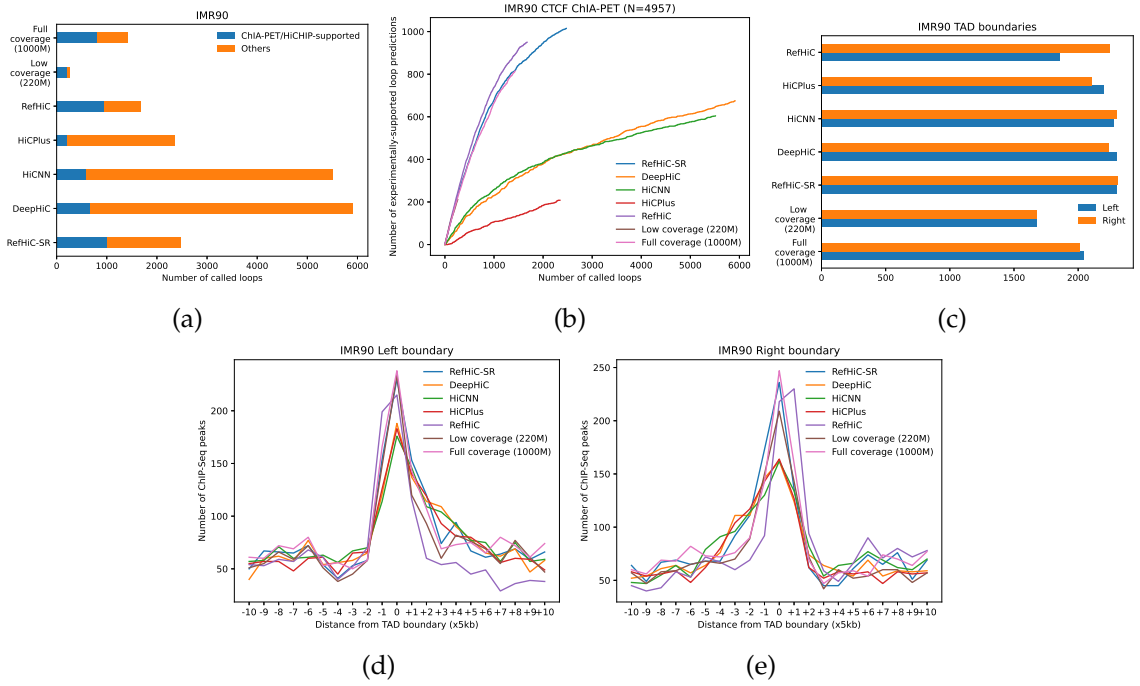


Figure 3.16: Comparison of loops and TADs annotated from low coverage, full coverage, and enhanced contact maps for human IMR-90 cells. (a) Number of loop annotations. (b) Loop predictions compared against CTCF ChIA-PET. (c) Number of TAD boundary annotations. Occupancy of ChIP-seq identified CTCF binding site as a function of distance to left (d) and right (e) boundary annotations.

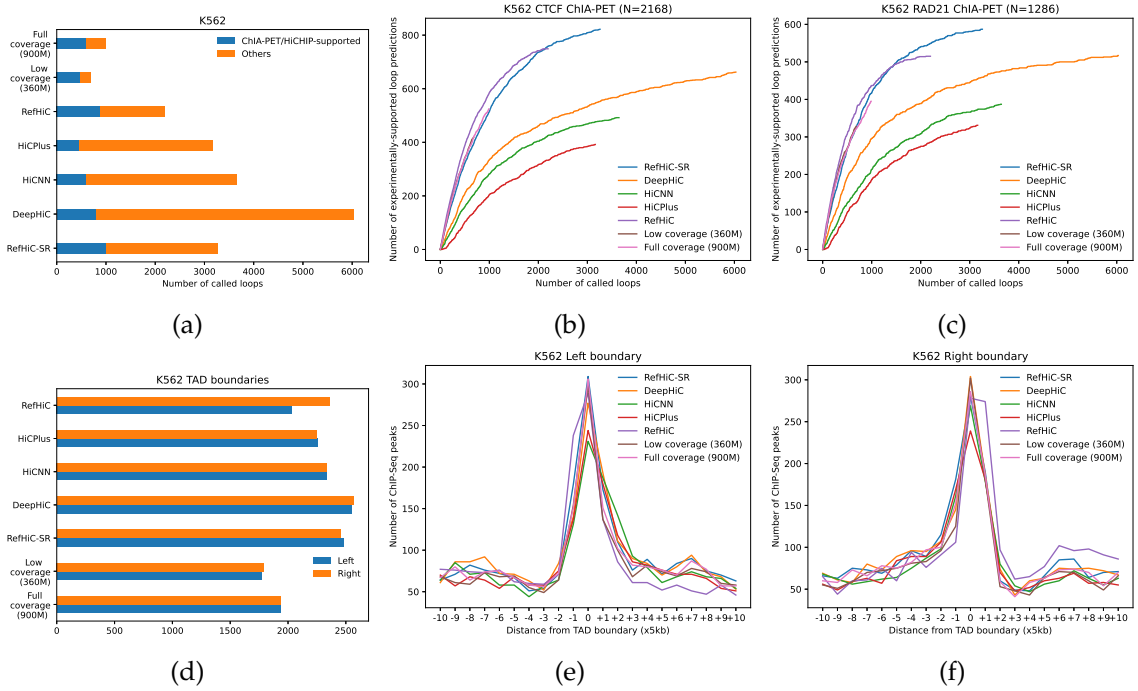


Figure 3.17: Comparison of loops and TADs annotated from low coverage, full coverage, and enhanced contact maps for human K562 cells. (a) Number of loop annotations. Loop predictions compared against CTCF ChIA-PET (b) and RAD21 ChIA-PET (c). (d) Number of TAD boundary annotations. Occupancy of ChIP-seq identified CTCF binding site as a function of distance to left (e) and right (f) boundary annotations.

Chapter 4

RobusTAD: Reference panel based annotation of nested topologically associating domains

Yanlin Zhang, Rola Dali, and Mathieu Blanchette *

School of Computer Science, McGill University, Montréal, Québec, H3A 0E9, Canada

Preface

Emerging evidence shows that chromatin is hierarchically organized, with topologically associating domains (TADs) playing a crucial role within this hierarchy. Consequently, TADs themselves are organized hierarchically. However, the development of tools for detecting such TAD hierarchies is still in its infancy. Existing tools often struggle to precisely identify TAD boundaries and/or hierarchies, particularly when dealing with low-coverage Hi-C data. Although we demonstrate that the RefHiC introduced in Chapter 2 significantly improved TAD annotation, it cannot produce TAD hierarchies.

In this chapter, we introduced RobusTAD [1] to infer TAD hierarchies from Hi-C data.

[1] The RobusTAD that we used in Chapter 3 is an old version [115] which only support annotating boundary and does not use a reference panel.

We demonstrated its effectiveness in accurately detecting TAD hierarchies using both low and high coverage Hi-C data. We attribute this success to the incorporation of an extensive reference panel comprising additional Hi-C contact maps. As elucidated in the previous chapters, our approach capitalizes on the conservation of local structures (including contacts, loops, and domain boundaries) among the study sample and our reference panel, benefiting the analysis of these features in samples of interest. However, it is important to note that TAD hierarchies are global structures and are specific to individual cells. To harness the benefits of introducing a panel of reference sample in detecting TAD hierarchies, we divide the annotation of TAD hierarchies into two steps: (i) Annotation of potential domain boundaries based on Hi-C contact maps. (ii) Pairing of left and right domain boundaries to form TAD hierarchies. To facilitate domain boundary annotation, we introduce a reference panel of samples. We perform stratified rank-sum tests to score domains and boundaries in both steps. The pairing of left and right boundaries are achieved through dynamic programming. Our contributions in this chapter are twofold. Firstly, we have substantially enhanced the accuracy of domain boundary annotation and the inference of TAD hierarchies. Moreover, we highlight the versatility of our reference panel-enabled methodology, showcasing its potential integration into both deep learning and conventional applications.

The rest of this chapter is the entire text from the following article:
Zhang, Y., Dali, R & Blanchette, M. (2023). RobusTAD: Reference panel based annotation of nested topologically associating domains. Submitted to Genome Biology

4.1 Abstract

Topologically associating domains (TADs) are fundamental units of 3D genomes and play essential roles in gene regulation. Hi-C data suggests a hierarchical organization of TADs. Accurately annotating nested TADs from Hi-C data remains challenging, both in terms of the precise identification of boundaries and the correct inference of hierarchies. While domain boundary is relatively well conserved across cells, few approaches have taken advantage of this fact. Here, we present RobusTAD to annotate TAD hierarchies. It incorporates additional Hi-C data to refine boundaries annotated from the study sample. RobusTAD outperforms existing tools at boundary and domain annotation across several benchmarking tasks.

4.2 Background

The hierarchical organization of mammalian chromosomes within the nucleus has been increasingly identified as an essential factor for cellular functions such as gene regulation, cell fate determination, and evolution [1]. Though the multi-scale chromosomal folding revealed by chromosome conformation capture techniques - such as Hi-C - are frequently studied [3, 4, 5, 11], understanding its structural and functional roles is still in its infancy. Moreover, identifying spatial elements such as TADs and loops from Hi-C data is challenging, particularly due to the relatively low resolution permitted by Hi-C datasets of typical sequencing depth (200-500M valid read pairs) [114, 63].

TADs are self-interacting regions along the chromosome, manifesting as squares along the diagonal of Hi-C contact maps [6]. Researchers used to perform TAD annotations from Hi-C datasets at low resolution (i.e., 40 kb), and define TADs as megabase-scale structural elements [11]. Later, researchers observed that TADs can be much smaller (i.e., tens to hundreds kilobases) in mammalian chromosomes by investigating high-coverage

and high-resolution Hi-C contact maps [4]. This observation led researchers to detect TADs at high resolutions [64, 6, 1]. In addition, the study of sequence-encoded factors, such as CTCF (CCCTC-binding factor), that influence TAD formation also requires researchers to annotate TADs at much higher resolutions. Meanwhile, TAD-within-TAD (or subTAD) organization also gathered significant attention in recent years [6].

Many computational tools for TAD annotation have been proposed [65, 66, 67, 4, 123, 124, 71, 68, 7, 11, 125, 126, 127, 174, 69, 70]. These approaches can be classified as either one-dimensional (1D) score-based and or matrix-based approaches. Score-based approaches, such as TopDom [65], Insulation Score (IS) [69], OnTAD [71], etc. assign each locus a score representing the strength of a potential TAD boundary and subsequently detect TAD boundaries by identifying local optima among the list of scores. Matrix-bases approaches directly utilize two-dimensional (2D) data instead of transforming it into a 1D statistic. For example, Arrowhead [4] transforms the Hi-C contact map into an arrowhead-shaped feature map and subsequently identifies TADs by searching for corners in the transformed matrix. Chen and colleagues formulated TAD annotation as graph segmentation [70], viewing a Hi-C contact map as an adjacency matrix to model a chromosome as a graph and identified TADs through graph Laplacians.

Despite the numerous TAD annotation tools available, the identification of TAD hierarchy and the precise location of TAD boundaries at high resolution remains challenging. Most TAD annotation tools are designed for high-coverage contact maps, or operate at low-resolution. As reviewed in previous studies [63, 72, 73, 74], these tools are not robust to resolutions and sequencing coverages. Additionally, the predictions of TADs and domain boundaries show limited agreement across tools. Finally, most existing algorithms only use the contact map of the sample of interest to annotate TADs. Their performance is thus limited by the sequencing depth of the Hi-C data from that study. Most Hi-C data sets produced to date are in the range of 200M to 500M valid read pairs, with only a slow

increase over time. We anticipate that this situation remains until sequencing costs drop dramatically.

The issue of insufficient data coverage also exists in many other biological data analysis tasks. Researchers often address this issue through introducing data from samples other than the sample under study [88, 175]. For example, though a SNP array typically covers only a few hundreds of thousands of loci, it is routine to infer unobserved genotypes through imputation with a reference panel containing a larger spectrum of genotyped variants [80, 83]. Similarly, homology modelling for protein structure prediction exploits a database of known structures [88]. Conversely, the vast amount of published Hi-C datasets are seldom employed to annotate TAD. We hypothesize that since TADs are determined by genome sequence, epigenetics, and cellular dynamics, the vast majority of TADs present in a given cell-type/condition are also present in multiple other samples. Given the large number of Hi-C datasets in public depositories (e.g. [176]), the stage is set to develop TAD annotation approaches that better utilize existing Hi-C data sets. Recently, we introduced RefHiC [172], a reference panel enabled approach for TADs and chromatin loops annotation. Although we demonstrated that the introduction of a Hi-C reference panel enables RefHiC to significantly outperform alternatives in TAD annotation, RefHiC suffers from several limitations: (i) RefHiC does not predict TAD hierarchies; (ii) the computationally intensive process of projecting Hi-C samples onto the latent space impedes including more samples into the reference panel.

Here, we introduce RobusTAD. RobusTAD is a TAD annotation algorithm that provides accurate and robust TAD annotation at high resolution. It improves TAD boundary annotation by leveraging publicly available Hi-C data and achieves superior performance by exploiting locally matched chromosome conformations (LMCC). Following TAD boundary annotation, it uses non-parametric tests and a dynamic programming algorithm to obtain the optimal nested TAD structure. RobusTAD outperforms existing

TAD callers in a variety of contexts. We further demonstrated that RobusTAD is robust to low sequencing coverage and can produce high-resolution TAD annotations from Hi-C data of typical sequencing depth (250-300M reads). Finally, we show that applying RobusTAD to predict TAD at high resolution facilitates dissecting TADs according to transcription factor binding site profiles around TAD boundaries and consequently probe TAD formation.

4.3 Results

4.3.1 Overview of RobusTAD

RobusTAD takes a normalized Hi-C contact matrix as input and calls TADs in three steps (Fig. 4.1): (i) Low-accuracy TAD boundary identification based on the study sample; (ii) Refinement of TAD boundary locations based on locally-matched chromosome conformations from a reference panel of Hi-C data sets; (iii) Pairing of left and right boundaries into an optimal nested domain hierarchy.

Study sample based boundary identification is based on seeking local maxima in a vector of 1D nonparametric TAD boundary scores. RobusTAD assigns separate left and right TAD boundary scores to each locus by performing genomic distance stratified rank-sum test between upstream/downstream inter- and intra-domain interactions.

RobusTAD refines boundary calls made on the study sample by utilizing selected Hi-C samples from a reference panel. For a given candidate TAD boundary at position p, we define locally matched chromosome conformations LMCC(p) as a collection of Hi-C samples in which a predicted TAD boundary occurs within 25 kb (i.e. five 5-kb bins) of p. It then computes refined boundary scores for the 50 kb region by combining the boundary scores from LMCC(p) and the study sample itself; the position that reaches the maximum score is the final high-resolution boundary prediction.

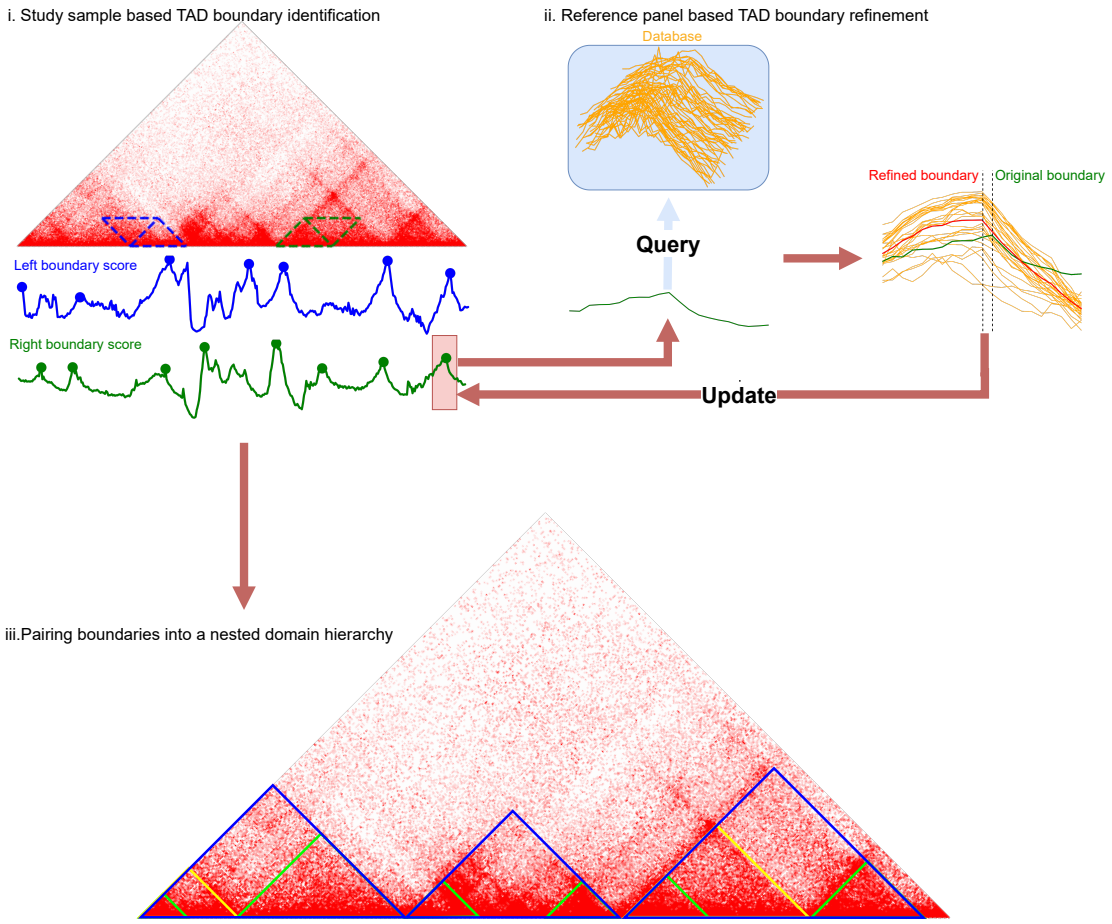


Figure 4.1: **Overview of the RobusTAD algorithm.** RobusTAD detects TAD boundaries in three steps (i, ii, and iii). First, approximate left and right TAD boundaries are identified based on the study sample. Second, RobusTAD identifies locally matched chromosomal conformations (LMCCs) from a panel of reference data sets, and uses those LMCCs to refine the position of each TAD boundary. Finally (step iii), refined left and right boundaries are paired to form an optimal nested TAD hierarchy.

In the third step, RobusTAD assembles a hierarchy of nested domains by pairing left and right boundary candidates using a dynamic programming algorithm, inspired by the Nussinov algorithm for RNA secondary structure prediction [177], that maximizes the full chromosomal TAD score (i.e. the sum of TADs scores). RobusTAD computes the TAD score with the distance-stratified rank-sum statistic of interactions between intra- and inter-domain in both upstream and downstream. To avoid the TAD score inflation

caused by the presence of sub-TADs within a given region, lower-level domains previously identified during the execution of the algorithm are excluded from a region's TAD score calculation. The algorithm is guaranteed to produce the globally optimal nested TAD hierarchy.

4.3.2 Comparison with existing TAD callers

We compared the performance of RobusTAD to 14 other TAD callers: TopDom [65], Armatus [66], deDoc [67], Arrowhead [4], HiTAD [123], EAST [124], OnTAD [71], CaTCH [68], Grinch [7], Domaincall [11], GMAP [125], HiCSeg [126], RefHiC [172] and IC-Finder [127]. We tried to include hierarchical TAD callers TADtree [178], TADpole [73] and SuperTAD [179] in our benchmarking, but they could not complete within a week of running time and hence we had to exclude them. Since RobusTAD and some other TAD callers detect nested TADs, we define TADs that do not contain any smaller TADs as level 0 TADs, and TADs that contain one or multiple smaller TADs as level 1+ TADs. We performed the benchmark evaluation experiments proposed by Zufferey et al. [72] on chromosomes 15-17 of Hi-C data for human GM12878 cells [4], down-sampled to 250 Million valid read pairs. We conducted all studies at 5kb resolution and employed iterative normalized [173, 49] Hi-C contact maps as input. The running time varies significantly among tools in annotating TADs from all three chromosomes. OnTAD and callers that do not produce nested TADs are able to annotate the three chromosomes within 20 minutes. RobusTAD took a total of 1.5 hours to annotate the three chromosomes. Most of other nested TAD callers also require a similar amount of time.

We first compared the number and size of TADs identified by each tool. Interestingly, the number of TADs varies greatly, with Arrowhead identifying less than 500 TADs and OnTAD identifying around 3,800. Most of the tools, including RobusTAD identified $1,000 - 2,000$ TADs (Fig. 4.2a). Tools that identify more TADs naturally produce smaller

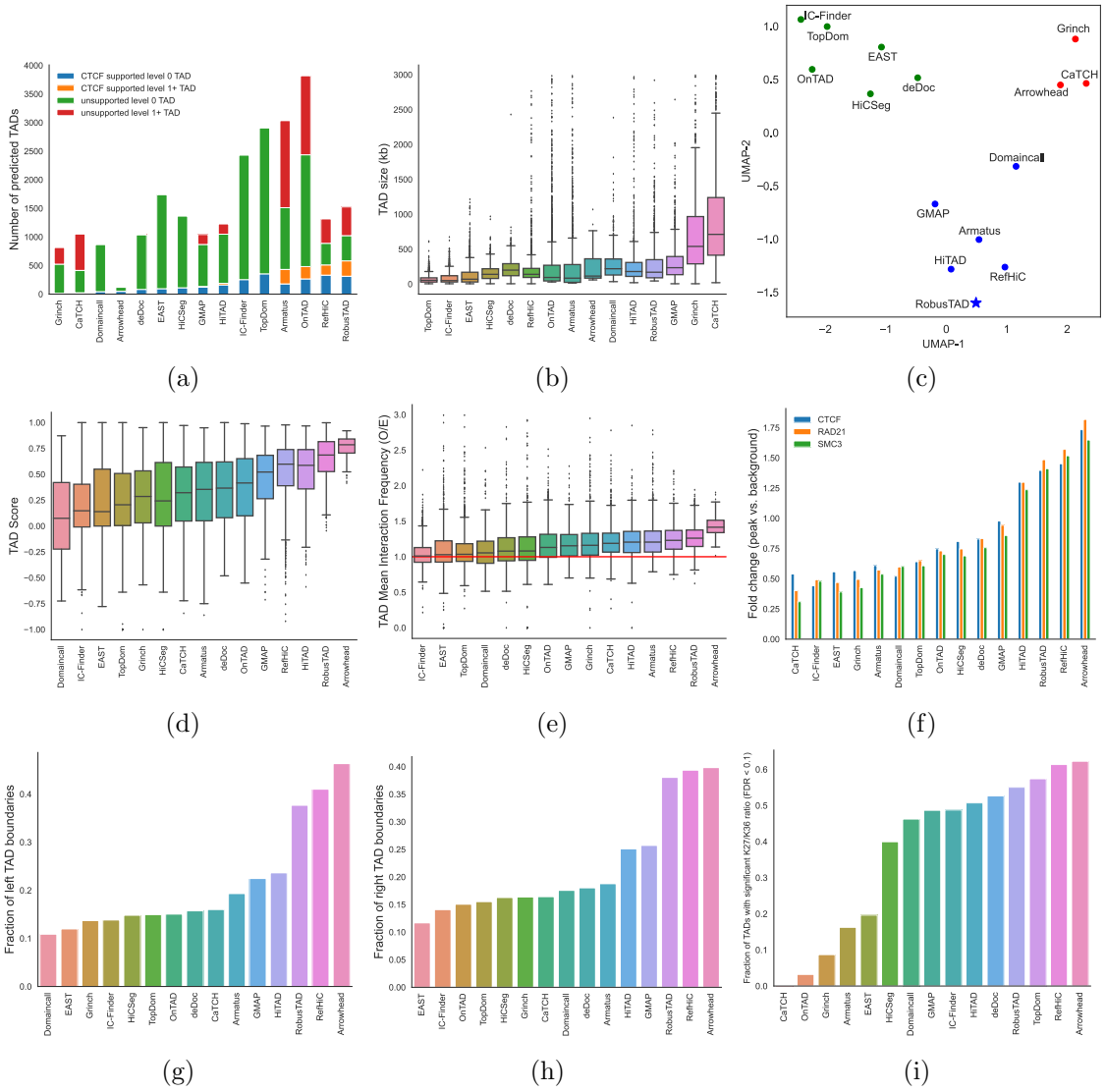


Figure 4.2: **Comparison of RobusTAD, and 14 other TAD callers on a GM12878 Hi-C data set of 250M valid read pairs. a,** Number of TADs predicted by different tools, and proportion of predicted TAD boundary pairs that are supported by CTCF ChIA-PET data. **b,** Size distribution of predicted TADs. **c,** U-MAP analysis performed on the Pearson's correlation matrix of the matrix of pairwise MoC between TADs identified by all callers. Comparison of the quality of TADs predicted by different tools using RobusTAD's TAD score (**d**), and TAD mean interaction frequency (observed/expected) (**e**). **f,** Fold change of structural protein peak signals at TAD boundaries for CTCF, RAD21, and SMC3. Number of left (**g**) and right (**h**) boundaries that contain at least one CTCF ChIP-seq peak. **i,** Fraction of TADs with significant log10 ratio between H3K27me3 and H3K36me3. Note: Panels **c,f-i** are generated with benchmarking code created by Zufferey et al. [72].

TADs (Fig. 4.2b). RobusTAD identified TADs of a wide range of sizes, with a median size of 170 kb. UMAP embedding [180] on Measure of Concordance (MoC) [72] values

among pairwise callers identifies three major caller groups. Among all callers, RobusTAD, RefHiC, Domaincall, GMAP, Armatus and HiTAD form a cluster with an average within-cluster MoC of 0.47 (Fig. 4.2c).

We then examined the quality of the TAD annotations produced by each tool. Fig. 4.7 shows an example genomic region (chr16:10.2 Mb – 12 Mb), with TADs annotated by RobusTAD and other TAD callers. TADs lack ground-truth annotation, so it is impossible to calculate the accuracy of TAD predictions. Thus, we used three metrics to evaluate each predicted TAD's quality. i) RobusTAD's TAD score (see Methods, on the full coverage data), ii) mean interaction frequency (Observed over Expected, on the full coverage data) inside a TAD, and (iii) agreement with CTCF ChIA-PET data. The TAD score measures the enrichment of interaction frequencies inside a TAD by using its neighbouring regions as the background. It ranges from -1 to 1; positive values indicate higher interactions within the TAD than across its boundaries. RobusTAD ranks second based on mean TAD score (Fig. 4.2d); only Arrowhead, a tool that predicts approximately 5 times fewer TADs, reaches a higher mean TAD score. Similar results are obtained when assessing predicted TADs based on average observed/expected ratios (Fig. 4.2e).

Loop TADs in mammalian genomes are TADs that exhibit a strong contact between their boundary loci [6]. We assessed TAD annotations by comparing predicted TAD boundary pairs with CTCF ChIA-PET data (allowing up to one 5-kb bin mismatch). Fig. 4.2a shows that 582 (38%) TAD predictions made by RobusTAD match ChIA-PET data. This is both the largest number and the largest proportion of supported TAD predictions across all tools.

We also studied the performance of TAD annotations at the level of individual TAD boundaries. We first calculated the enrichment for ChIP-Seq signals of structural proteins (CTCF, RAD21, and SMC3) associated with predicted TAD boundaries (Fig. 4.2f and 4.8).

TAD boundaries predicted by most tools are enriched for these architectural proteins. RobusTAD ranked $3^{rd}$for the mean fold-change enrichment of the three structural proteins. Fig. 4.2g,h compare left and right boundaries to CTCF ChIP-seq data (allowing 1-bin mismatch) separately; RobusTAD ranked 3^{rd} for both left and right boundary predictions, slightly outperformed by Arrowhead, and the other reference panel enabled tool, RefHiC. Histone marks usually correlate with regulatory activity, and TADs are typically consistently enriched for either activating (H3K36me3) or repressive (H3K27me3) marks. We calculated the ratio between H3K27me3 and H3K36me3 within each TAD prediction and counted the fraction of TAD predictions where this ratio was particularly large or small (see Methods). Fig. 4.2i shows RobusTAD ranked 4^{th}, slightly outperformed by TopDom, RefHiC, and ArrowHead.

We then conducted a visual examination of TAD predictions made by all tools. We performed this analysis using rescaled pileup plots generated by Coolpup.py [181], with the '--local' and '--rescale' options. Fig. 4.3, and 4.9 show most tools, including RobusTAD, identified TADs as square regions with increased interaction frequency, dot-corners, and well-defined domain borders. In contrast, tools such as Domaincall, deDoc, and Armatus yielded TAD predictions with less distinct domain borders. TAD predicted by Grinch and HiCSeg are less enriched for Hi-C contacts. In addition, we observed clear vertical and horizontal stripes with increased interaction frequencies at the boundaries of TADs predicted by RobusTAD, HiCTAD, RefHiC, TopDom, and GMAP. The two stripes indicate these TAD callers can identify both TADs and sub-TADs.

Taken collectively, these results suggest that RobusTAD is the most accurate TAD caller, as it is the only tool ranking among the top four TAD callers in all accuracy metrics.

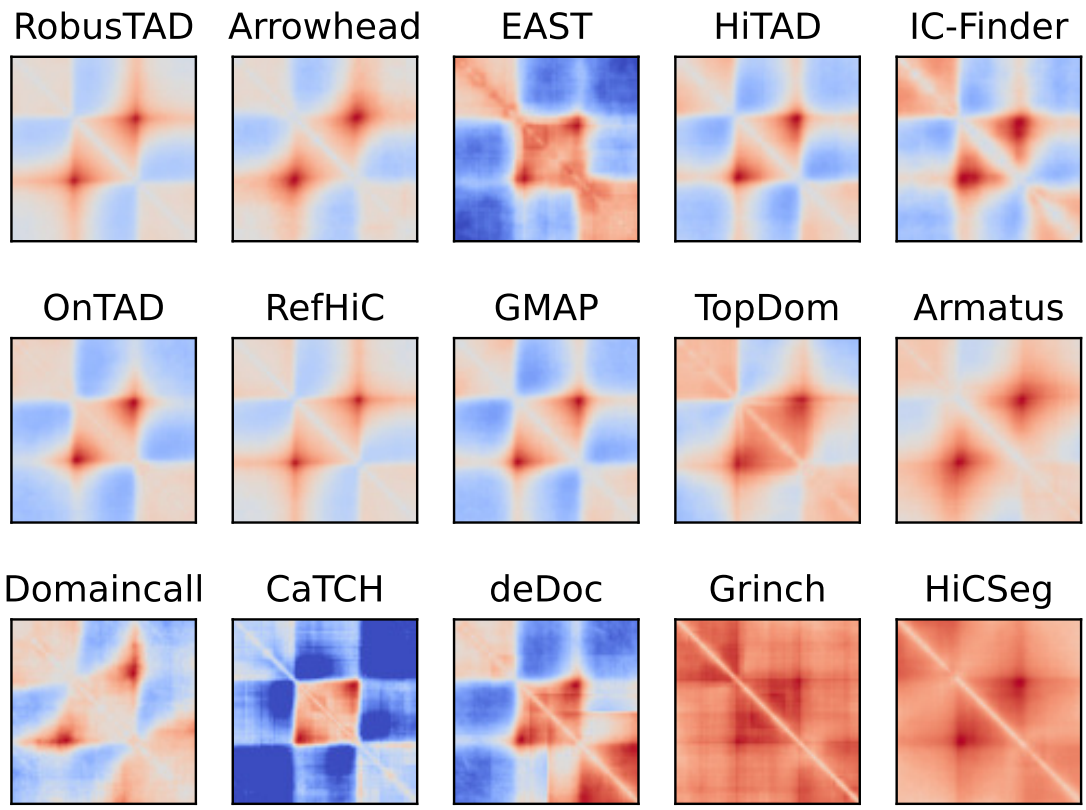


Figure 4.3: **Visual comparison of TADs predicted by RobusTAD and 14 other tools from GM12878 Hi-C data.** The rescaled pileup plots aggregate areas around TAD predictions in the full-coverage Hi-C contact map. TAD predictions were annotated against a downsampled Hi-C contact map containing 250M valid read pairs.

4.3.3 RobusTAD is robust to low sequencing coverage Hi-C data

TAD annotation is typically sensitive to sequencing depth. Many TAD callers do not perform well when the sequencing depth is low, and boundaries detected from contact maps of differing sequencing depths have been reported to lack reproducibility [114, 63]. We evaluated how RobusTAD and other tools performed on Hi-C contact maps of varying sequencing depths (generated from the combined Hi-C data set for GM12878 [4]), from data set of 4B valid read pairs down to downsampled versions with as few as 62.5M valid read pairs. As illustrated in fig. 4.4a, different tools react differently to reduced coverage:

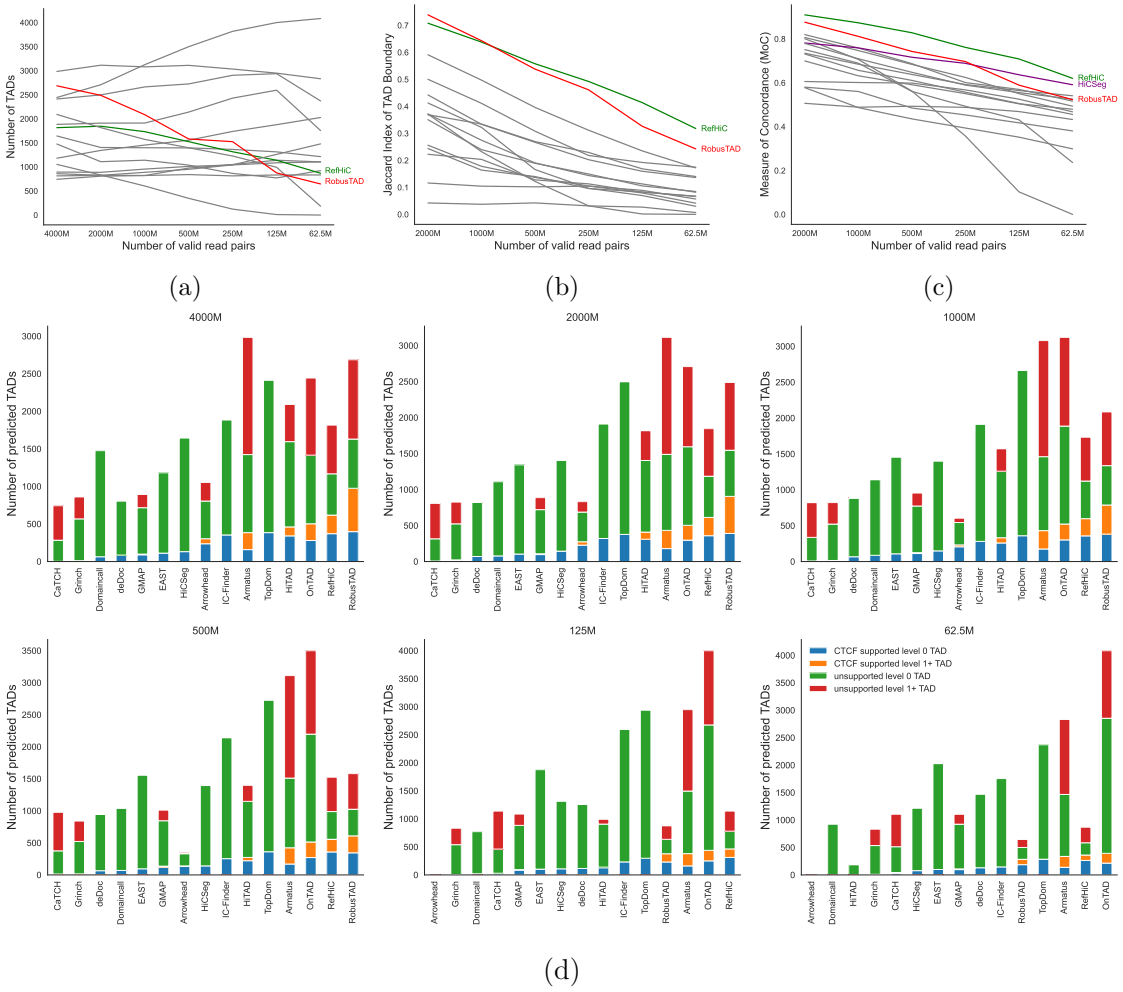


Figure 4.4: **Comparison of RobusTAD, and other 14 TAD callers on dowmsampled GM12878 Hi-C data. a,** Number of TAD predicted from downsampled Hi-C data. Jaccard index of predicted TAD boundaries (**b**) and Concordance between TADs (**c**) predicted on full data (4B valid read pairs) compared to those predicted on the downsampled Hi-C data. **d** Number of TADs predicted from downsampled data, and proportion of predicted TAD boundary pairs that are supported by CTCF ChIA-PET data.

some (including RobusTAD) conservatively reduce their predictions, while others are unaffected or even increase their number of predictions. These results suggest that tools like RobusTAD can mitigate false-positive identifications effectively.

We then assessed the tool's robustness by measuring the similarity between the predictions made on the highest coverage data set to those made on downsampled data, both at the levels of TAD boundaries (fig. 4.4b, using the Jaccard index) and TADs (fig. 4.4c,

using the Measure of Concordance). RobusTAD and RefHiC, the two reference panel based approaches, exhibit the highest levels of consistency at TAD boundaries level. Following Zufferey et al., we used the Measure of Concordance (MoC) [72] to compare two sets of TAD predictions. MoC does not handle overlapping TADs, thus we only included TADs that do not include any smaller TADs in this analysis. Fig. 4.4c shows RefHiC, HiCSeg, and RobusTAD outperformed other tools at most levels of coverage. Last, we evaluated TAD and domain boundary prediction accuracy by comparing predictions to CTCF ChIA-PET (fig. 4.4d) and CTCF ChIP-Seq (Fig. 4.10,4.11) data. For contact maps containing more than 250M valid read pairs, RobusTAD performed the best. On data containing fewer valid contact pairs, RobusTAD is only slightly outmatched by RefHiC. Although OnTAD, Armatus, and TopDom also identified more CTCF supported TADs from data containing few valid read pairs, they were less accurate than our tools, as they identified many more TADs that were not supported by CTCF data.

4.3.4 RobusTAD performs well across cell types

Here, we demonstrate that RobusTAD performs well across cell types. We applied RobusTAD and five other TAD callers (GMAP, HiTAD, Arrowhead, OnTAD, and RefHiC), to annotate TADs from Hi-C contact maps derived from IMR-90 and K562 cell lines [4]. The rescaled pileup plots show that all tools successfully identified TADs as squares with increased interaction frequencies and dot-corners (Fig. 4.5a). In addition, TAD predictions made by all tools contain vertical and horizontal stripes with increased interaction frequency at its boundary locations. These stripes indicate these tools accurately detect TADs and subTADs from Hi-C contact maps. The number of TAD predicted by different tools from the two contact maps ranges from 500 to 3500, with RobusTAD detecting 2630 and 1710 TADs from the two contact maps (Fig. 4.5b). The mean expectation normalized interaction frequency (O/E) within a TAD further confirms that all tools, including

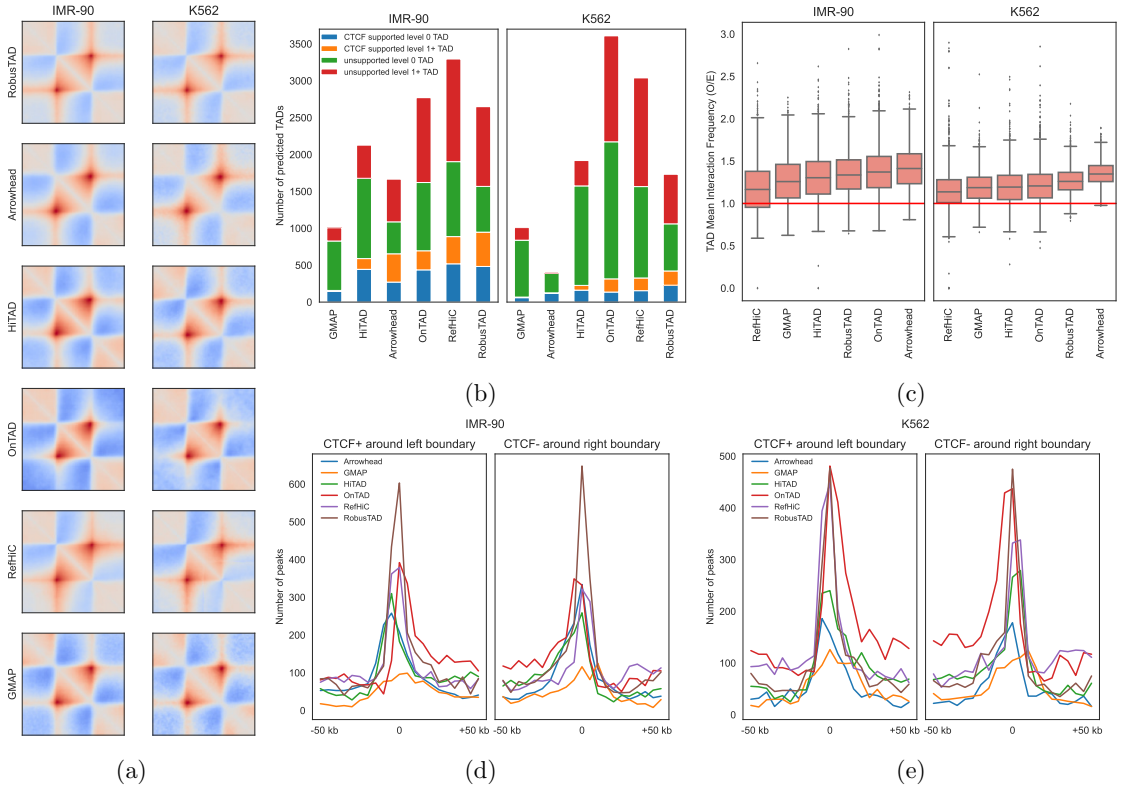


Figure 4.5: **Comparison of RobusTAD, and other five TAD callers on Hi-C data derived from IMR-90 and K562 cell lines. a,** Rescaled pileup plots over predicted domains. **b,** Number of TADs predicted by different tools, and proportion of predicted TAD boundary pairs that are supported by CTCF ChIA-PET data. **c,** TAD mean interaction frequency (observed/expected). Occupancy of ChIP-seq identified forward and reverse CTCF binding site as a function of distance to left (**d**) and right (**e**) boundary annotations.

RobusTAD, successfully identified TADs as a region with increased *cis*-contact pairs (Fig. 4.5c). Next, we compared boundary pairs to CTCF ChIA-PET data (Fig. 4.5b). The ChIA-PET data for IMR-90 contains 4957 contact pairs and the ChIA-PET data for K562 contains 2168 contact pairs. RobusTAD identified the most ChIA-PET supported TADs from both contact maps. Last, we evaluated boundary prediction accuracy by comparing predicted boundary to forward and reverse CTCF binding sites identified by ChIP-Seq experiment. Left and right boundaries predicted by RobusTAD are more enriched by CTCF binding sites than boundaries identified by alternative tools.

4.3.5 RobusTAD reveals multiple types of TADs

Building upon the high accuracy and resolution of RobusTAD, we used it to perform a study of TADs functionalities. We focused on TAD predictions made on the full set of autosomes for a combined Hi-C data set obtained from the GM12878 cell line [4]. We characterized a TAD as a binary vector of dimension $2 \times 116 = 232$, representing the ChIP-seq derived occupancy of 116 transcription factors [119] at its left and right domain boundaries. We identified six TAD groups by applying the UMAP algorithm [180] to project TADs onto a 2D space, followed by K-Means clustering [182] (Fig. 4.6a). The symmetric pattern observed in the UMAP projection (Fig. 4.6a) and the group-averaged occupancy vectors (Fig. 4.6b) indicate that left and right domain boundaries play similar functional roles. Chromatin structural proteins such as ZNF143, CTCF, YY1 and subunits of cohesin complex (SMC3, and RAD21) are the most enriched proteins in all six groups. We also observed TRIM22 highly enriched at domain boundaries in all groups (Fig. 4.6b).

The distribution of transcription factors at TAD boundaries (Fig. 4.6b) and the chromatin accessibility quantified as the average count of ATAC-seq peaks at domain boundaries (Fig. 4.6f) motivate us to interpret group assignments by investigating transcriptional activity and chromatin accessibility.

Groups 3 is characterized by having both boundaries exhibiting evidence of transcriptional activity, with high TF occupancy (Fig. 4.6b) and chromatin accessibility (Fig. 4.6f), often involving pairs of regions annotated as active enhancers or promoters (as annotated by ENCODE's combined Segway ChromHMM segmentations [119]) (Fig. 4.6c,e). This is confirmed this by comparing boundary pairs to Enhancer-Promoter links identified by a POLR2A ChiA-PET experiment [119] (Fig. 4.6d). Additionally, we observe that Enhancer-Promoter pairs can mediate TAD formation even in the absence of CTCF binding sites, with 14.5% of Group-3 TADs corresponding to such links but lacking CTCF occupancy.

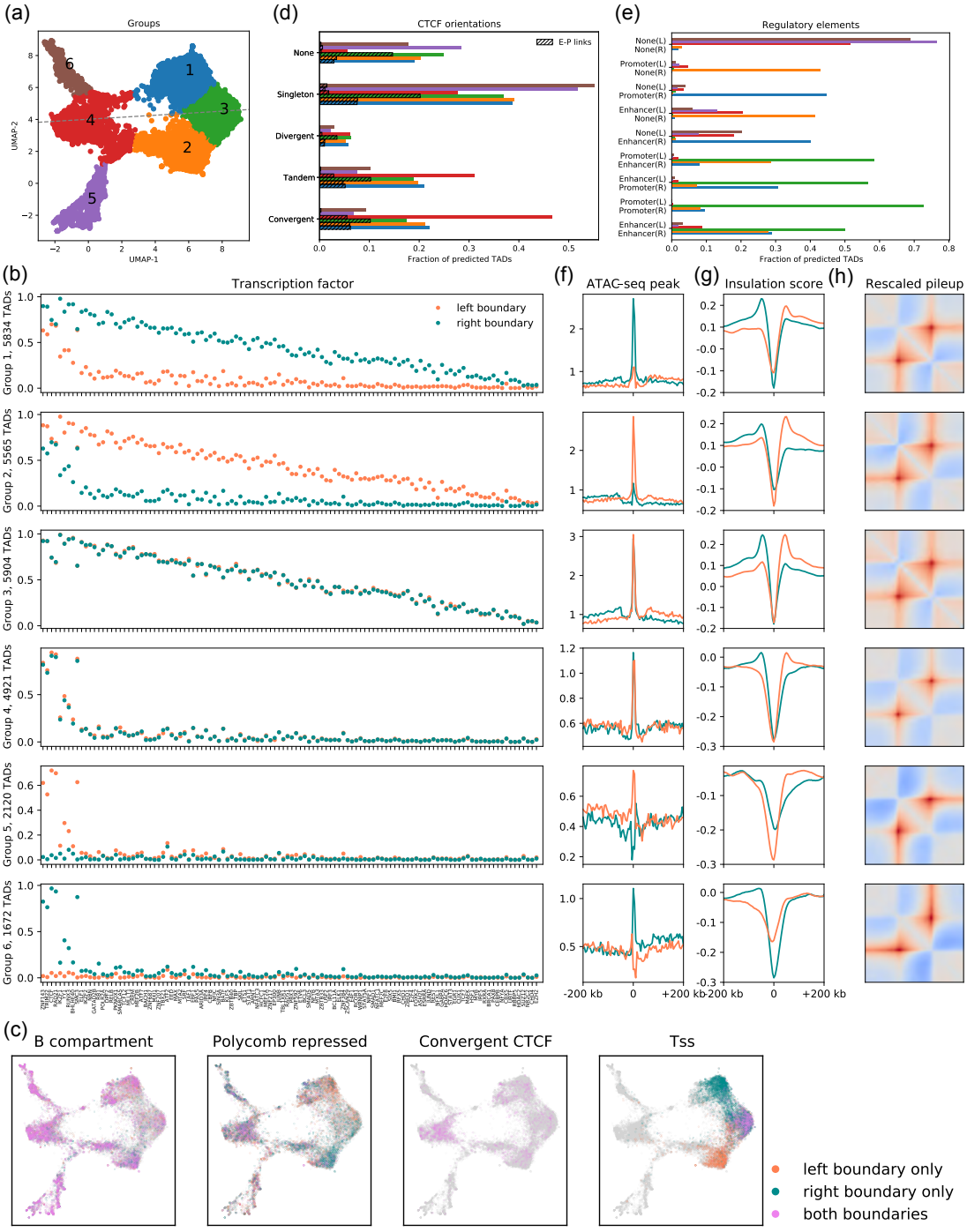


Figure 4.6: **Applying RobusTAD to Hi-C data for GM12878 cells reveals TAD groups.** **a,** a two dimensional UMAP projection of TADs based on the occupancy of transcription factors at domain boundaries. **b,** occupancy of transcription factors in each group of TADs. **c,** two dimensional distributions in the UMAP projected space of TADs associated with different features. **d,** occupancy different pairs of directional CTCF binding sites at domain boundaries. E-P links are domains supported by POLR2A ChiA-PET data. **e,** Proportion of annotated TADs with different regulatory element combinations at domain boundaries. Enrichment of ATAC-seq peaks (**f**), and insulation scores (**g**) at domain boundaries for each TAD group. **h,** Rescaled pileup plots over TAD predictions for each TAD group.

Groups 1 and 2 only display evidence of transcriptional activity at one of their two boundaries, with the inactive boundary showing reduced TF occupancy. Although those three groups are quite different in terms of their activity profiles, their insulation profiles (Fig. 4.6g) and pile-up plots (Fig. 4.6h) are nearly identical.

Groups 4, 5, and 6 are characterized by TADs whose both boundaries are located in repressive chromatin (Fig. 4.6c) with low TF occupancy (Fig. 4.6b) and chromatin accessibility (Fig. 4.6f), and little overlap with active enhancers/promoters (Fig. 4.6e). Group-4 TADs have both boundaries occupied by CTCF and associated structural proteins; these TADs' boundaries also display the highest level of convergent CTCF binding sites (Fig. 4.6d) and have sharper corner dots than domains in other groups (Fig. 4.14), probably because of the reduced level of interactions in surrounding regions. On the contrary, Group-5 and 6 TADs lack CTCF at one or the other of their boundaries. They also exhibit weak insulation scores at the CTCF-free boundary (Fig. 4.6g, 4.12a, 4.13), and weak dot corners (Fig. 4.14).

Among all groups, domain boundaries in active regions are sharper than in repressed regions (Fig. 4.6g, 4.12a, and 4.13). Domains in active regions are more enriched by Hi-C contacts than domains in repressive regions (Fig. 4.12b).

We observe that domain boundaries shared by multiple TADs are more enriched for CTCF bindings sites and activated promoters (Fig. 4.15). We further studied the hierarchy structure of these TADs by classifying them into singleton TADs (isolated TADs that do not overlap with others), TADs (non-singleton TADs that do not reside within other TADs), and sub-TADs (non-singleton TADs found within larger TADs) (Section 4.12.2 and Fig. 4.16). We found TADs frequently associated with boundaries marked by convergent CTCF motifs, and subTADs playing an important role in gene regulation, with a substantial portion being E-P links.

4.4 Discussion

Hi-C experiments and their derivatives have become routine in studying 3D genome organization at the genome-wide scale. Many Hi-C studies have been carried out in the past decade, and hundreds of Hi-C datasets have been published. Though this type of data has enabled the discovery of several key levels of 3D genome organizations (e.g. loops, TADs, and compartments), accurately identifying TAD boundaries and the ways in which they assemble to form a TAD hierarchy remain challenging with existing tools, especially for Hi-C data with typical sequencing coverage. To deepen our understanding of 3D genome organization, high-resolution annotations of TAD hierarchy are required.

RobusTAD annotates high-resolution TAD boundaries and TAD hierarchy from Hi-C contact maps, taking advantage of a reference panel of high-quality Hi-C data sets. Including more reference samples improves annotation accuracy (Fig. 4.17, and 4.18), so one can expect that RobusTAD will continue to get better as its reference panel grows. RobusTAD is based on a novel nonparametric statistic to score both domain boundaries and TADs. It is a distribution-free test to evaluate TAD and TAD boundaries. Thus, it is robust to changes in observation, such as those due to noise and sparsity. In addition, RobusTAD's separate scoring of left and right boundaries eases the dissection of domain boundaries. In contrast, most existing TAD callers do not quantify domain boundaries or amalgamate left and right boundaries as a single insulation locus.

RobusTAD overcomes the statistical challenges caused by high sparsity and signal-to-noise ratio limitation in a Hi-C contact map of typical coverage by identifying and combining many locally matched Hi-C data. At moderate sequencing depths, existing tools often fail at maintaining a low level of false identifications and identify many inaccurate TADs. Although a user can adjust parameters to limit the number of identified TADs in most tools, adjusting parameters is challenging and not necessarily effective at reducing

false identification. Within RobusTAD, we use a simple and statistically sound target-decoy search strategy to select TAD boundaries from a list of candidate boundaries. A user only needs to specify the desired False Discovery Rate (FDR) threshold (α) to ensure that the final predictions to contain at most α expected false-positive boundaries.

Our tools outperformed many TAD callers in accuracy and reproducibility in identifying high-resolution TAD from multiple Hi-C contact maps. For instance, both CTCF ChIA-PET and CTCF occupancy data highlight the superiority of RobusTAD at both boundary detection and TAD assembly. The benefits of RobusTAD were shown to be particularly significant in typical moderate-to-low coverage Hi-C data. As demonstrated in fig. 4.4a,d, RobusTAD can reduce false-positive identifications by identifying slightly fewer TAD boundaries from low-coverage Hi-C contact maps. In contrast, false-positive boundaries increased dramatically in predictions made by most other TAD callers when applied to low-coverage Hi-C contact maps. While RobusTAD was bested by RefHiC in terms of the accuracy of boundary annotation from very low coverage Hi-C data, it outmatched even that tool in terms of predicting CTCF ChIA-PET supported TADs. This advantage of RobusTAD over RefHiC is attributed to the newly developed dynamic programming algorithm for pairing TAD boundaries in RobusTAD. In addition, RobusTAD guarantees full interpretability without a loss of accuracy. When a boundary only appears in the study sample, RobusTAD will usually annotate it without using additional data from its reference panel. LMCC based boundary refinement makes mistake only if the boundary in the study sample is very close (i.e., within 50 kb) to another boundary in the reference panel. Given the superior performance achieved by RobusTAD, we believe this case rarely occurs.

Despite its advantages, RobusTAD has several limitations. First, given RefHiC slightly outperformed RobusTAD in some aspects of TAD boundary annotation, we believe RobusTAD does not fully capitalize on the advantages offered by the reference panel. It

might be due to some weak boundaries that cannot be identified in the first step of RobusTAD. We can overcome this by replacing the first step with the deep learning model in RefHiC, at the cost of losing efficiency and interpretability. Second, the dynamic programming is time consuming, in part due to we need to compare two large sets of interaction frequencies to evaluate the score of a large candidate domain. We are planning to improve the running time by sampling a fraction of elements from the two sets to estimate the score of large candidate TADs.

4.5 Conclusion

RobusTAD allows for precise, high-resolution TAD annotation from Hi-C data of a wide range of sequencing depths, all the way down to only 62.5 million contact pairs. RobusTAD improves the performance of TAD boundary annotation by exploiting locally matched contact maps in a reference panel. By enabling high-resolution and robust analyses of topological domains from standard coverage Hi-C data, RobusTAD paves the way to gaining biological insights that had until now could only be possible from ultra-high coverage (and cost) data.

4.6 Methods

4.6.1 Notations

Consider an intra-chromosomal contact map $M = \{m_{ij}\}$, where m_{ij} represents (normalized) interaction frequency between bin i and j at fixed resolution r. RobusTAD aims to detect TAD boundaries $B = \{B^L, B^R\}$ and TADs $D = \{(B^L_i, B^R_{i'})\}$, where B^L and B^R are lists of left and right boundaries respectively, $(B^L_i, B^R_{i'})$ indicates that the i^{th} left and i'^{th}

right boundaries form a TAD. Define $M_{[a,b]}$ as the submatrix corresponding genomic region $[a, b]$, and $S_{[a,b]}$, $S^L_{[a,b]}$, and $S^R_{[a,b]}$ as the domain score, left and right boundary scores for TAD (a, b).

4.6.2 Boundary and domain scores

To calculate domain and boundary scores for (a, b), we compared interactions for bins within $[a, b]$ and between bins in $[a, b]$ and its left and right flanking regions. Our null hypothesis in the nonparametric test assumes that, for each diagonal of the contact matrix, there are no differences between the distribution of within-TAD and across-TAD-boundary interactions. To quantify TAD, we performed a distance stratified (i.e. diagonal-wise) rank sum test between the two types of interactions. Define $D_k(a, b) = \{(i, i + k) : i \geq a, i + k \leq b\}$. We denote the TAD score evaluated from the k^{th} diagonal as $S^k_{[a,b]}$, and compute it using within-stratum ranks as follows:

$$S^k_{[a,b]} = \frac{\sum_{(i,j)\in D_k(a,b)} \sum_{(i',j')\in D_k(2a-b,a-1)\cup D_k(b+1,2b-a)} \mathbb{1}(m_{i,j} > \gamma m_{i',j'}) - \mathbb{1}(m_{i,j} < \frac{1}{\gamma} m_{i',j'})}{\sum_{(i,j)\in D_k(a,b)} \sum_{(i',j')\in D_k(2a-b,a-1)\cup D_k(b+1,2b-a)} \mathbb{1}(m_{i,j} > \gamma m_{i',j'}) + \mathbb{1}(m_{i,j} < \frac{1}{\gamma} m_{i',j'})} \tag{4.1}$$

where $\gamma \geq 1$ controls the minimum gap allowed between the two types of interactions. Setting $\gamma = 1$ is equivalent to the Wilcoxon rank sum test. The overall TAD score for region $[a, b]$ is a weighted sum of the per-stratum scores:

$$S_{[a,b]} = \frac{2}{3(1 + b - a)(b - a)} \sum_{k=1}^{k=b-a} (b - a + k + 1) S^k_{[a,b]} \tag{4.2}$$

.

$b-a+k+1$ is the number of values used for comparison on the k^{th} diagonal, $\frac{3(1+b-a)(b-a)}{2}$ is the total number of values used for comparison. $S_{[a,b]}$ falls between -1 and 1, where $S_{[a,b]} = -1$ indicates all interactions inside the TAD are smaller than all interactions across TAD boundaries by a factor at least $\frac{1}{\gamma}$, $S_{[a,b]} = 0$ indicates no difference existed between the two types of interactions, $S_{[a,b]} = 1$ indicates all interactions inside the TAD exceed all interactions across TAD boundaries by a factor of at least γ. Note that if a nested TAD (a', b') was already determined to occur within (a, b) (with $a \leq a' < b' \leq b$), we exclude interactions belonging to (a', b') from the calculation.

We define left and right boundary scores $S^L_{[a,b]}$ and $S^R_{[a,b]}$ similarly but only using interactions across the corresponding boundary as the background.

4.6.3 Identifying candidate TAD boundaries

To identify domain boundaries from a normalized intra chromosomal contact map, we first compute a left boundary score $L_a = \max_{w \in \{w_{min},\ldots,w_{max}\}} S^L_{[a,a+w]}$ for each bin a along the whole chromosome. Right boundary scores are computed similarly: $R_a = \max_{w \in \{w_{min},\ldots,w_{max}\}} S^R_{[b-w,b]}$. For our analyses at 5 kb resolution, we use $w_{min} = 50, w_{max} = 250$.

To identify left and right boundaries from boundary scores, we use find_peak function in SciPy [131] to identify local peaks. We assume the minimum distance between two domain boundaries is 25 kb (i.e. five 5-kb bins) and set *distance=5* in find_peak. The set of putative boundaries this identifies usually contains false positives. We use FDR control to select a subset of high confidence boundaries. Briefly, we produced a decoy contact map by shuffling interactions diagonal-wise. The shuffling strategy destroys all domains but maintains the interaction frequency decay pattern. We identify domain boundaries from this decoy contact map and compare scores for boundaries identified in the original and decoy Hi-C contact maps. We select the top boundaries at a FDR of α ($\alpha = 0.05$ for data

containing more than 300M valid read pairs, $\alpha = 0.1$ for data containing less than 300M valid read pairs).

4.6.4 Refining boundary annotation by identifying locally-matched chromosome conformations from the reference panel

Putative TAD boundaries predicted by the single-sample non-parametric test described above are often off by one or more bins, due to the noisy nature of the data. For a given left or right putative TAD boundary predicted at b_i, we define LMCCs as the subset of the Hi-C samples from our reference panel that have a predicted boundary with 25 kb (5 bins) of b_i. We then update the study sample's boundary scores for the 10-bin region centered at b_i as the mean boundary scores of the study sample and all selected reference samples (Section 4.12.1). Last, we update the boundary call as the peak position among the refined boundary scores.

4.6.5 Assembly of nested TADs from predicted boundaries

Given an intra-chromosomal contact map $M = \{m_{ij}\}$ and sets B^L and B^R of previously identified left and right boundaries, we sought to pair left and right boundaries to form hierarchical domains. Similar to OnTAD [183], we do not allow partial overlaps between domains in TAD predictions. While this assumption may inadequately address heterogeneous regions, as illustrated in previous work [183], it remains applicable to the majority of the genome. Moreover, this fully nested TAD hierarchy assumption allows us to use a dynamic programming algorithm to find a globally optimal solution. Our dynamic programming algorithm is inspired by the Nussinov algorithm [177] for RNA secondary structure prediction.

We denote the ordered multi-set of TAD boundaries as $(b_1, b_2, \ldots, b_{n-1}, b_n)$, where $b_i \in B^L \cup B^R$. We define the globally optimal solution as the nested TAD hierarchy that

maximizes the sum of scores of all TADs in the hierarchy, subject to the TADs' left and right boundaries being selected (potentially with repetition) from B^L and B^R. We create the dynamic programming table T of size $n \times n$, where T_{ij} stores the maximum sum of domain scores for all nested domains within region $[b_i, b_j]$. The forward pass of the dynamic programming fills the upper triangular portion of T, using the following recursion:

$$T_{ij} = \max_{i<k<j} T_{ik} + T_{kj} + \delta(i,j) \tag{4.3}$$

$$\delta(i,j) = \begin{cases} S_{[b_i,b_j]} & \text{if } b_i \in B^L \text{ and } b_j \in B^R \text{ and } S_{[b_i,b_j]} \geq \lambda \\ 0 & \text{otherwise} \end{cases} \tag{4.4}$$

λ defines the minimum score for pairing b_i and b_j as a domain (we set λ=0.2). The evaluation of $\delta(i,j)$ depends on k in the recursion function of T_{ij} as $S_{[b_i,b_j]}$ requires excluding nested TADs within $[b_i, b_j]$ which are identified with the dynamic programming algorithm in previous steps. We sequentially fill entries in T from the first to the furthest diagonals. To start, we initialize the first upper diagonal as

$$T_{i,i+1} = \begin{cases} S_{[b_i,b_{i+1}]} & b_i \in B^L, b_{i+1} \in B^R, \text{and } S_{[b_i,b_{i+1}]} \geq \lambda \\ 0 & \text{otherwise} \end{cases} \tag{4.5}$$

Last, we select the optimal set of domains that maximize the sum of TAD scores for genomic region $[b_1, b_n]$ by backtracking from T_{1n}. Some domain boundaries in b may be absent from the TAD hierarchy. They are treated as false positive domain boundaries or involved in partial overlap domains, which does not satisfy our assumption.

4.6.6 Curating Hi-C reference panel

We downloaded 177 published human Hi-C datasets (Table 4.1) from the GEO database and uniformly processed them with distiller [184]. Reads were mapped against hg38 and we discarded reads with a mapping quality < 10. This produced Hi-C contact maps at fixed resolutions and stored processed contact maps in multi-resolution cooler format (.mcool). Lastly, read count matrices were normalized using Cooler's iterative correction algorithm [49, 173]. We applied the single-sample version of robusTAD with default parameters to calculate boundary scores for all of these Hi-C samples at 5 kb resolution and saved boundary scores and boundary calls as a reference database, to be used for the reference-panel based version of RobusTAD.

4.6.7 Enrichment analysis and Measure of Concordance

We followed Zufferey et al. [72] to analyze enrichment of H3K36me3 and H3K27me3 histone marks and CTCF, SMC3, and RAD21 structural proteins within TADs or at their boundaries. For structural protein enrichment, we calculated the fold-change by comparing peak counts in a narrow interval around a boundary to those in distant flanks (Fold change=$\frac{\text{peak}}{\text{background}}$-1). For histone marks, we calculated the average log10-ratio in small intervals within TADs and obtained empirical p-values using ten shuffles.

To compare TAD partitions, we used the Measure of Concordance (MoC) [72], which ranges from 0 (absence of concordance) to 1 (full concordance) and is defined as follows,

$$\text{MoC}(\mathbf{P},\mathbf{Q}) = \begin{cases} 1 & \text{if } N_P = N_Q = 1 \\ \frac{1}{\sqrt{N_P N_Q}-1}\left(\sum_{i=1}^{N_P}\sum_{j=1}^{N_Q}\frac{|\mathbf{F_{i,j}}|^2}{|\mathbf{P_i}||\mathbf{Q_j}|}-1\right) & \text{otherwise} \end{cases} \tag{4.6}$$

where $\mathbf{P} = \{\mathbf{P_i}\}$, and $\mathbf{Q} = \{\mathbf{Q_i}\}$ are sets of TADs including N_P and N_Q TADs, $\mathbf{F_{i,j}}$ is the overlap region between $\mathbf{P_i}$ and $\mathbf{Q_j}$, and $|\cdot|$ represents cardinality. We only included TADs

without any smaller TAD in this analysis.

4.6.8 Alternative approaches

This study compared RobusTAD to 14 other TAD callers. We ran TopDom, Armatus, Arrowhead, EAST, CaTCH, Domaincall (DI), GMAP, ICFinder and HiCSeg as suggested in [72]. As we performed the analysis at 5kb resolution, we have updated parameters related to resolutions accordingly. We ran HiTAD, RefHiC, and deDoc with their default settings. OnTAD: We set maxsz=600 to allow OnTAD to detect TADs as large as 3Mb. Grinch: following Lee and Roy [7], we detected TADs by setting the expected TAD length as 2Mb, 1Mb, and 500Kb in three runs and combined all results. We observed tools such as Grinch reported invalid TAD annotations of length less than three bins and excluded these invalid annotations from further investigation. In addition, the convention of TAD definition often varies from tool to tool (with ± 1 bin shift). We converted them to the convention used in RobusTAD (i.e., boundary refers to the start of the left/right furthermost bin inside the TAD).

4.7 Acknowledgements

The authors thank Dr. Yue Li, and Dr. Jacek Majewski for useful discussions in this project.

4.8 Funding

This work was funded by a Genome Quebec/Canada grant to M.B.. Y.Z. is supported by FRQNT Doctoral (B2X) Research Scholarships.

4.9 Abbreviations

TAD: topologically associating domain

LMCC: locally matched chromosome conformation

CTCF: CCCTC-binding factor

Availability of data and materials

RobusTAD is available at https://github.com/zhyanlin/RobusTAD, under MIT license. All scripts and data required to reproduce figures and analyses are available at https://doi.org/10.5281/zenodo.8306238.

4.10 Competing interests

The authors declare that they have no competing interests.

4.11 Authors' contributions

Y.Z. and M.B. conceived the study. R.D. co-conceived and implemented domain boundary annotation, and wrote portions of the manuscript. Y.Z. implemented RobusTAD, performed data analyses, and wrote the manuscript. M.B. supervised the project and wrote the manuscript. All authors read and approved the final paper.

4.12 Appendix

Supplementary Information

4.12.1 Refined boundary score is a stratified rank-sum test

In RobusTAD, to refine score of a putative domain boundary b_i, we select samples in the reference panel that have domain boundaries within a 50 kb region around b_i. TAD are relatively conserved across cells. Thus, we assume domain boundaries inside this 50 kb region are identical among all samples (i.e., study sample and selected reference samples), and compute refined boundary scores as the mean boundary scores for the study sample and all selected reference samples. Here, we show that a mean boundary score is equivalent to a stratified rank-sum test score where each stratum contains all interaction frequencies of a particular genome distance that comes from a particular sample. This stratified rank-sum test evaluates the strength of a putative boundary using interaction frequencies from all samples that exhibit similar local structures in our study. Each sample is an observation of the same local structure. To simplify, we consider refined boundary scores as the mean of two samples' scores S^1 and S^2 computed with a window of size w. Boundary scores evaluated from the k^{th} diagonal are $S^{1,k}$ and $S^{2,k}$ respectively. Following our definition of the boundary score, we have

$$S^1 = \frac{1}{w \times w} \sum_{k=1}^{w} wS^{1,k}$$
$$S^2 = \frac{1}{w \times w} \sum_{k=1}^{w} wS^{2,k}$$

thus,

$$\begin{aligned}
\text{mean}(S^1, S^2) &= \frac{1}{2}(S^1 + S^2) \\
&= \frac{1}{2}(\frac{1}{w \times w} \sum_{k=1}^{w} wS^{1,k} + \frac{1}{w \times w} \sum_{k=1}^{w} wS^{2,k}) \\
&= \frac{1}{2 \times w \times w} \sum_{i=1}^{2} \sum_{k=1}^{w} wS^{i,k}
\end{aligned}$$

is a stratified rank-sum test, where $2 \times w \times w$ is the total number of interaction frequencies and w is the number of interaction frequencies in each stratum.

4.12.2 A comparison of singleton TADs, TADs, and subTADs

RobusTAD detects TAD hierarchies from Hi-C contact maps, allowing us to classify TAD predictions into distinct categories. These categories include singleton TADs, which are isolated TADs that do not overlap with others; TADs, which are non-singleton TADs that do not reside within larger TADs; and sub-TADs, which are non-singleton TADs found within larger TADs. Sub-TADs are further divided into three groups: sub-TAD A, with left boundaries being left TAD boundaries; sub-TAD B, with right boundaries being right TAD boundaries; and sub-TAD C, encompassing other subTADs. Rescaled pileup plots in Fig. 4.16a show that within-domain interactions in predicted TADs of all groups are larger than their surroundings. Notably, we observe dot-corners displaying increased interactions across all groups, with the strongest dot-corner pattern associated with TADs. Singleton TADs exhibit relatively weak domain boundaries and less involvement in transcription. These boundaries also show lower enrichment of CTCF binding sites and RAD21. In contrast, TAD boundaries (including right boundaries of subTAD A and left boundaries of subTAD B) serve as robust insulating regions compared to other types of domain boundaries. They are more enriched for architectural proteins such as CTCF and RAD21 and more actively involved in transcriptional processes (as indicated by TSS and ATAC-seq signals around these boundaries). Sub-TAD boundaries (including both boundaries of sub-TAD C, right boundaries of sub-TAD A, and left boundaries of sub-TAD B) are relatively weaker and less enriched for architectural proteins. The presence of Tss at sub-TAD boundaries falls between that of TAD boundaries and singleton TAD boundaries. Additionally, both boundaries of sub-TAD C display greater accessibility than domain boundaries of any other type, as measured by ATAC-Seq. Furthermore,

We studied Enhancer-Promoter links by comparing TAD boundary pairs against polII ChIA-PET data (Fig. 4.16b). We found that 5% of TADs are E-P links; 15%-20% of singleton TADs, sub-TADs A and B are E-P links; and more than 28% of sub-TADs C are E-P links. Fig. 4.16c illustrates that TAD boundary pairs are notably enriched in convergent CTCF motifs compared to other boundary pairs, with singleton TADs showing less enrichment in convergent CTCF motifs. In summary, these observations show the importance of TAD hierarchies in facilitating gene expression and regulation, with TADs frequently associated with boundaries marked by convergent CTCF motifs, and subTADs playing an important role in gene regulation, with a substantial portion being E-P links.

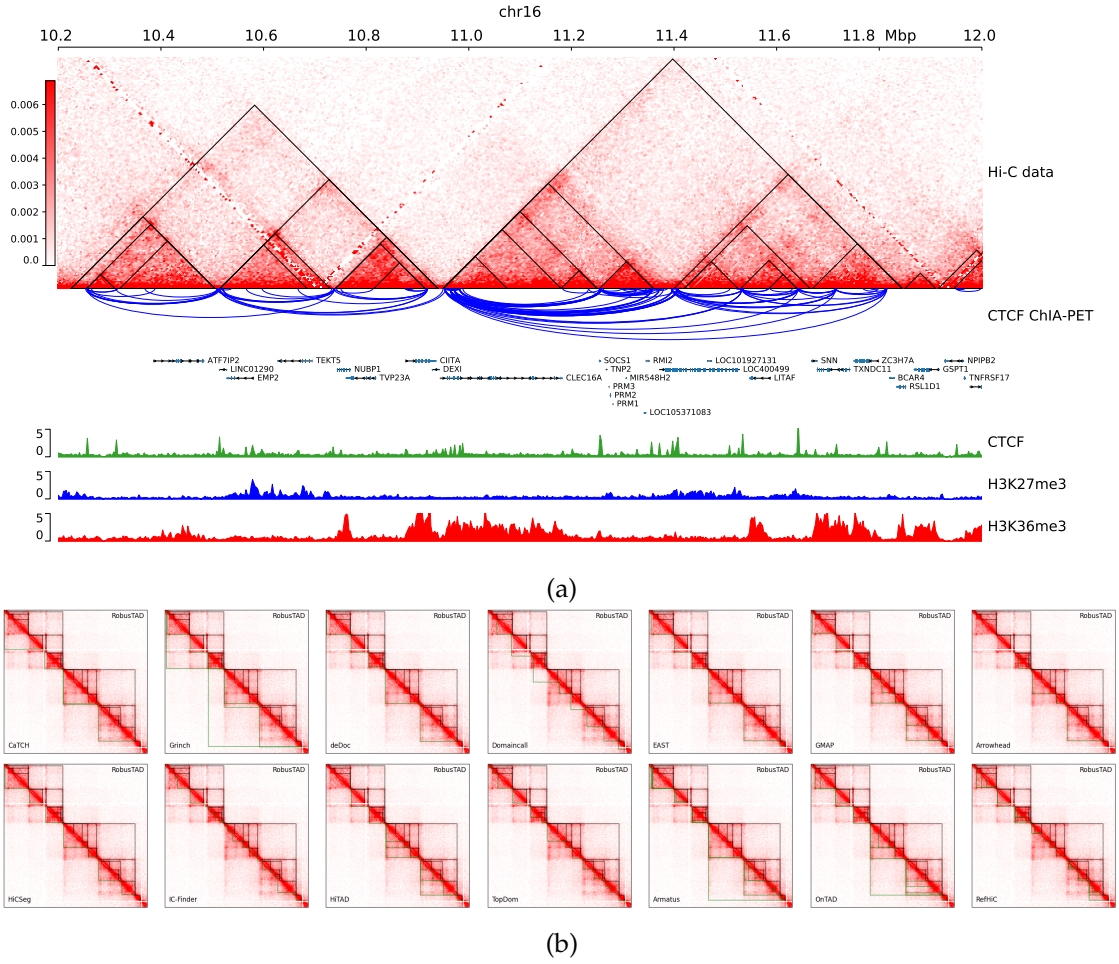


Figure 4.7: **TAD identification for an example genomic region (chr16:10.2 Mb – 12 Mb) of GM12878 cells. a,** TAD identified by RobusTAD on GM12878 cells. Note how TAD predictions are supported by the CTCF ChIA-PET data and consistent with gene annotation and epigenetic features. **b,** Comparison of TADs detected by different tools. RobusTAD annotated two nested sets of TADs. Every gene in this region is included entirely within a TAD. Most predicted TAD boundaries are collocated with ChIP-seq peaks, and loops identified by CTCF ChIA-PET support many TAD predictions; both support the conclusion that RobusTAD produces accurate TAD annotations. We also observed that TADs annotated by RobusTAD are either enriched for either activation (H3K36me3) or repression (H3K27me3) marks, but rarely both. ChIA-PET data suggests that a weak TAD (chr16:10.25Mb-10.7Mb) is missed by RobusTAD because it partially overlaps other TADs. Among all tools, only CaTCH detected this weak TAD.

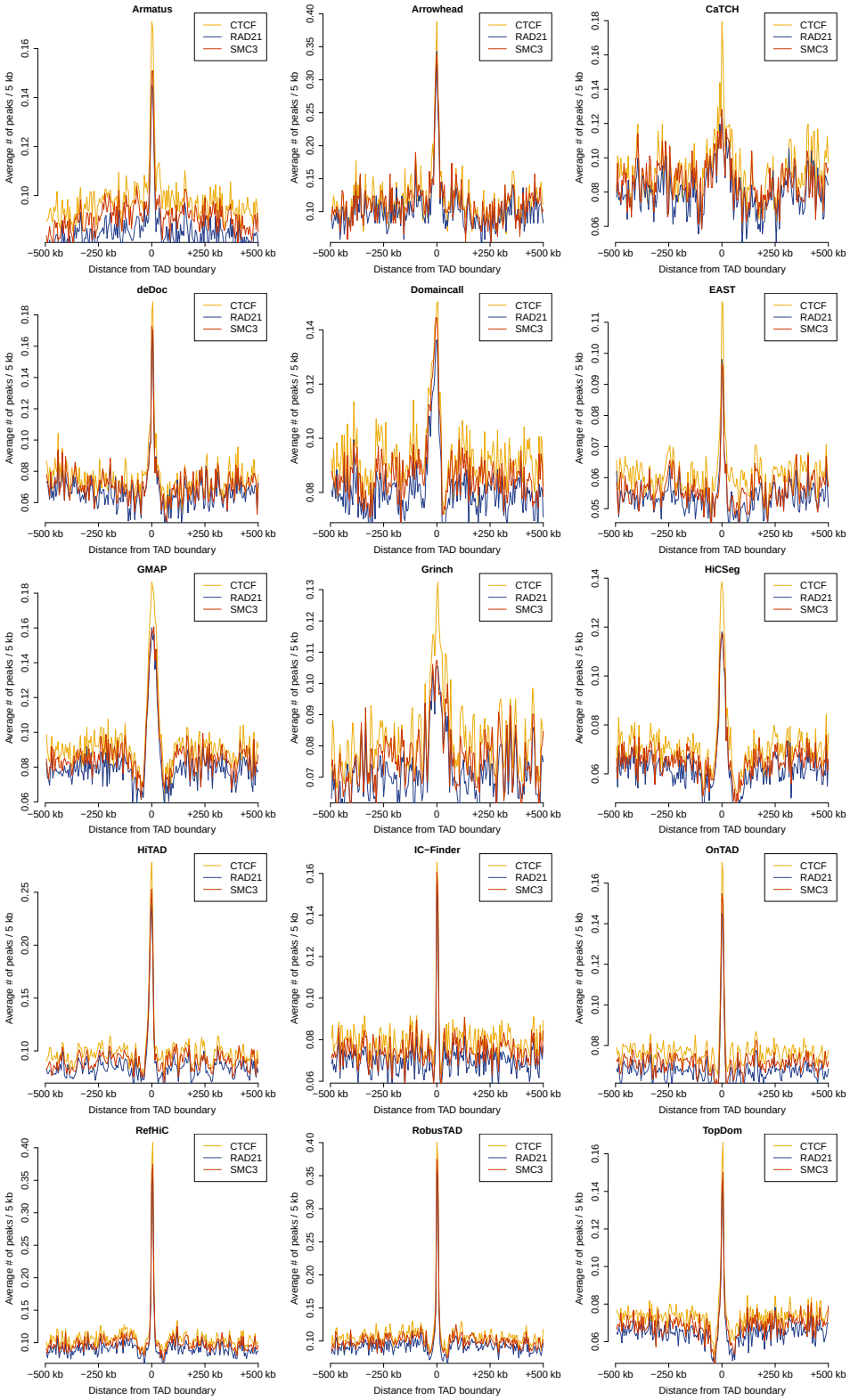


Figure 4.8: **ChIP-seq peak signals for CTCF, RAD21, and SMC3 around TAD boundaries annotated by each tool.**

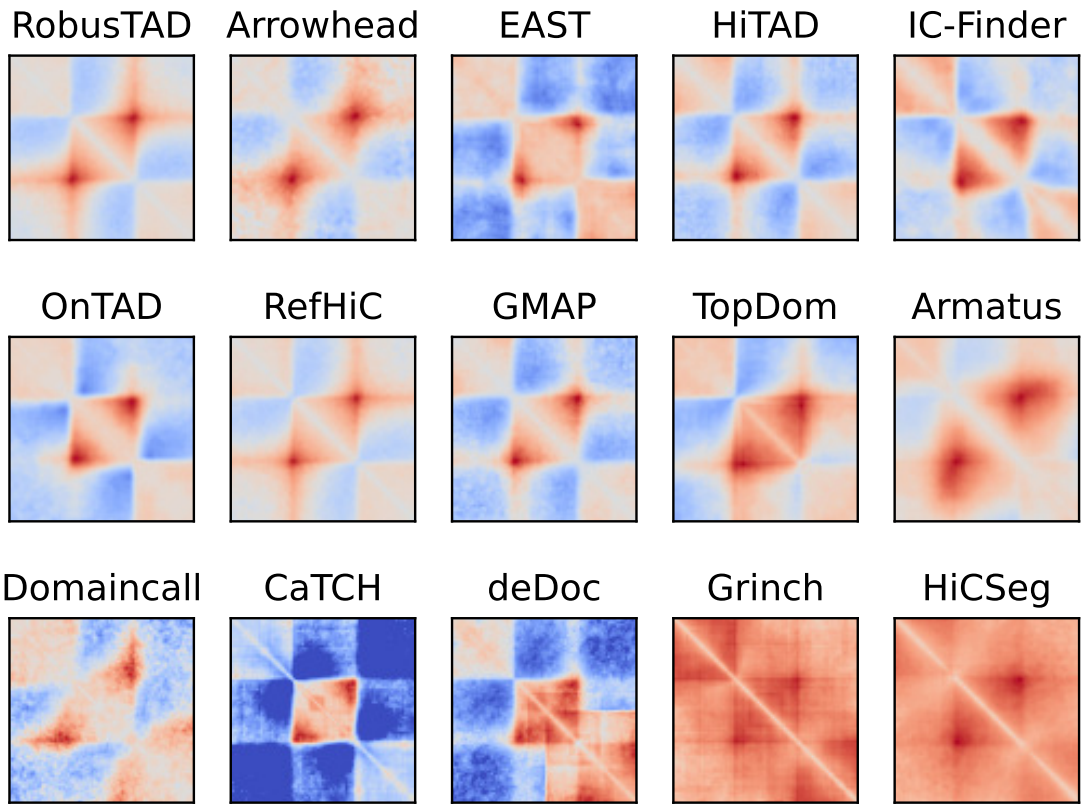


Figure 4.9: **Visual comparison of TADs predicted by RobusTAD and 14 other tools from a GM12878 Hi-C data.** These plots are created by aggregating regions over a Hi-C contact map containing 250M valid read pairs. The regions are TAD predictions used in Fig. 4.3.

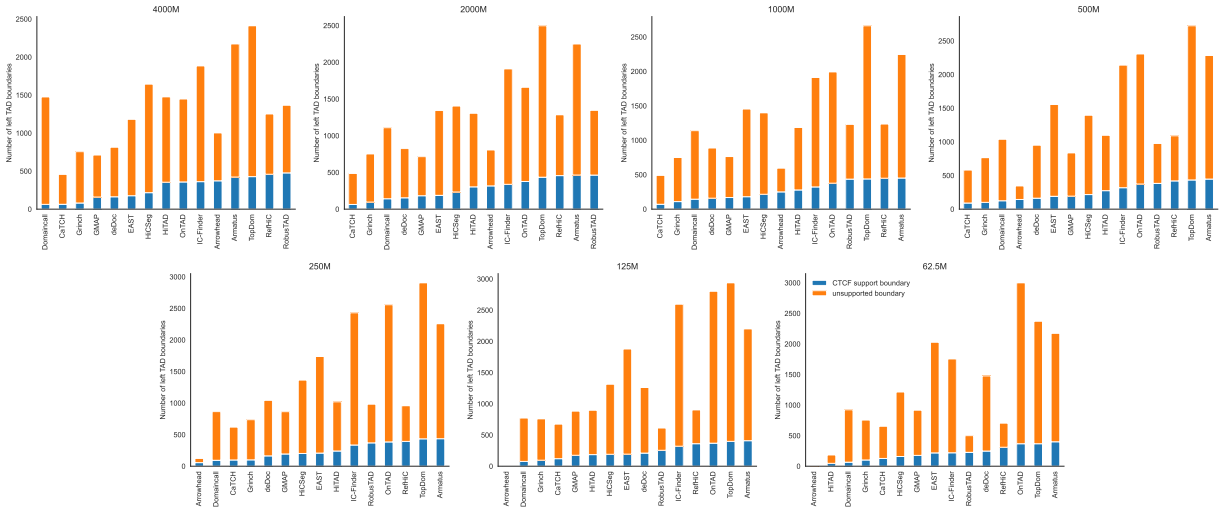


Figure 4.10: **Number of left TAD boundaries predicted by different tools, and proportion of predicted boundaries that are supported by CTCF ChIP-Seq data.**

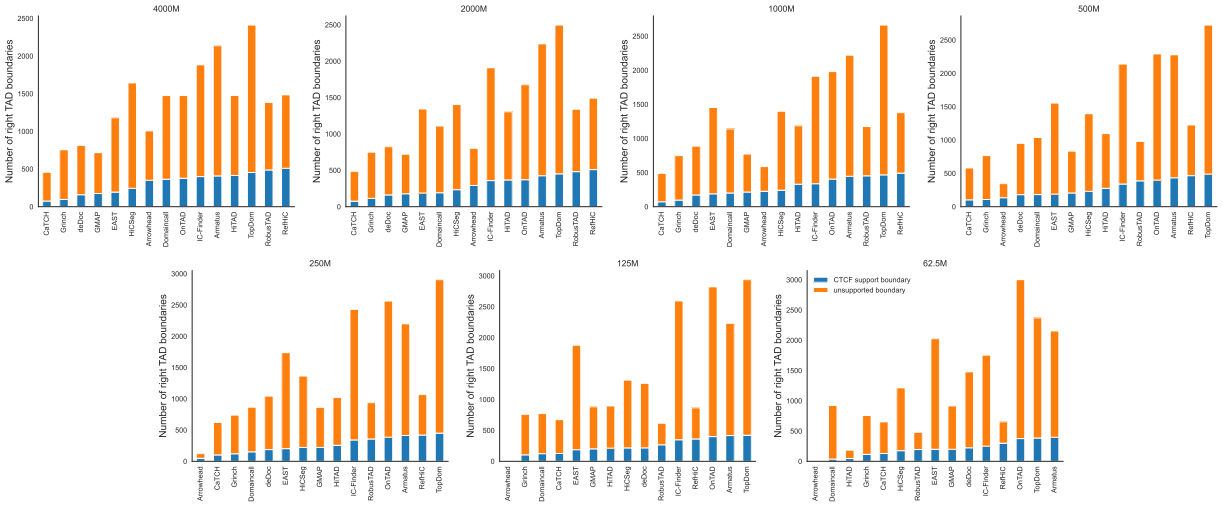


Figure 4.11: **Number of right TAD TAD boundaries predicted by different tools, and proportion of predicted boundaries that are supported by CTCF ChIP-Seq data.**

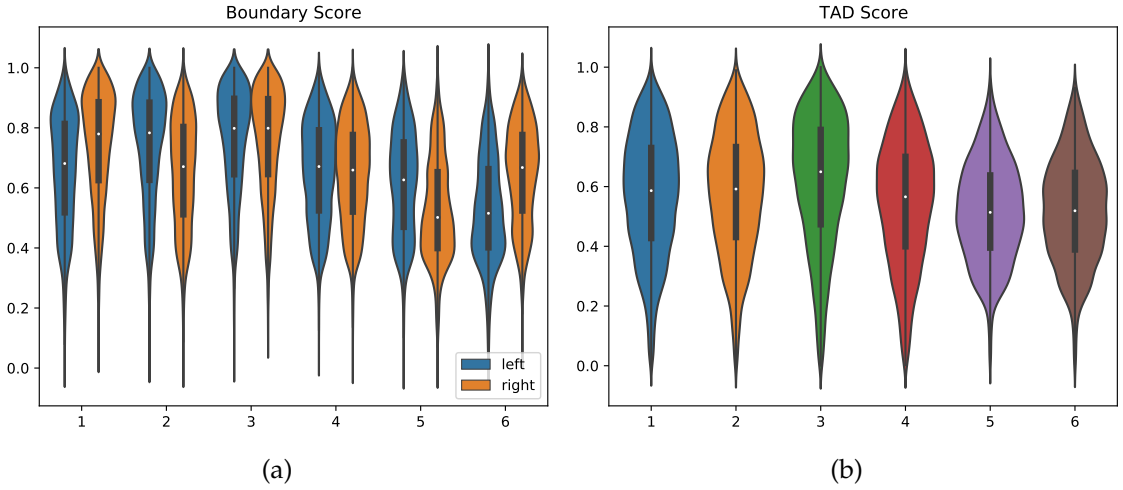
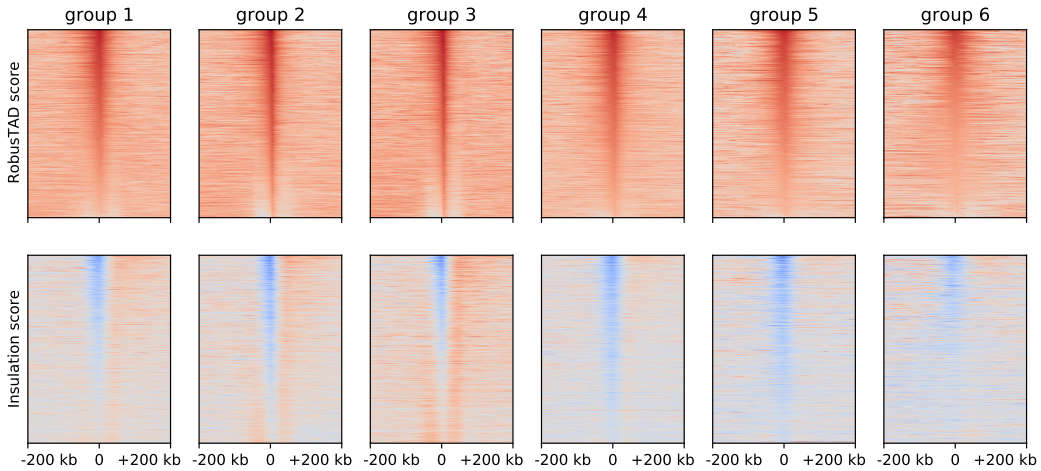
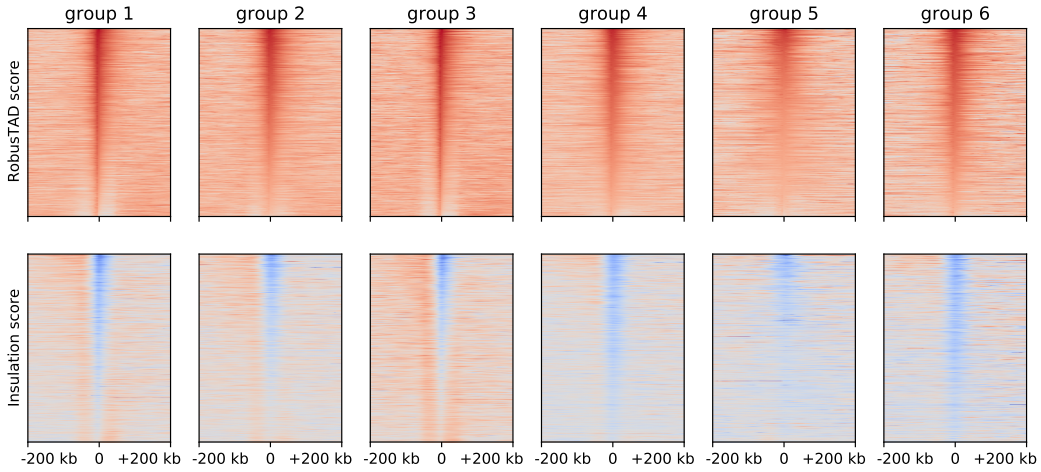


Figure 4.12: **RobusTAD score at domain boundaries (a) and domains (b) of the six groups of TADs predicted from the combined Hi-C data for GM12878 cells.**



(a)



(b)

Figure 4.13: **Insulation score and RobusTAD score around domain boundaries of the six groups of TADs predicted from the combined Hi-C data for GM12878 cells. a,** Left boundary. **b,** Right boundary.

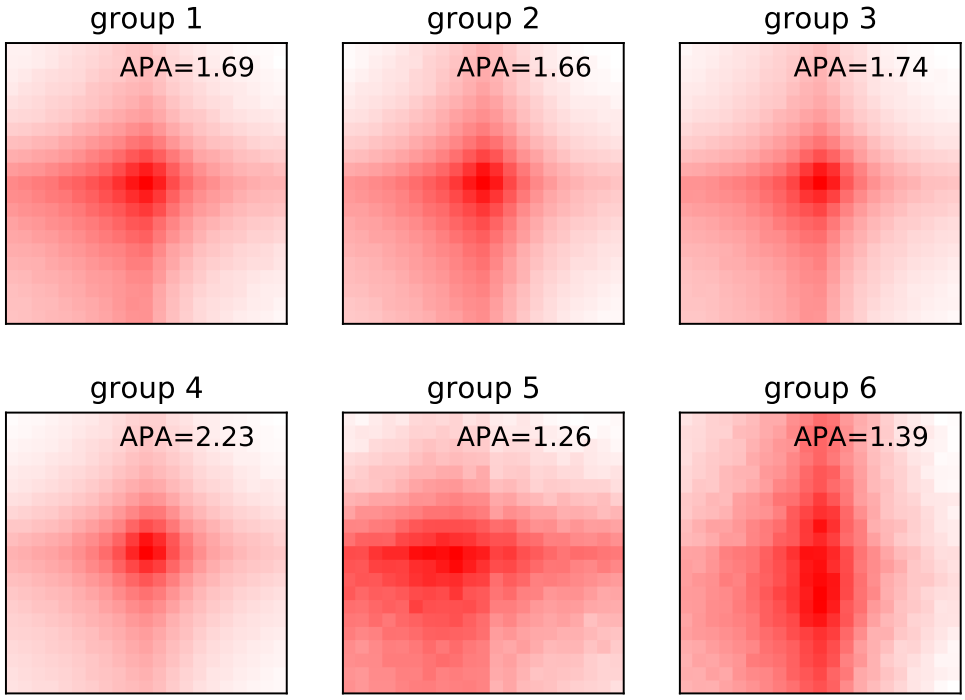
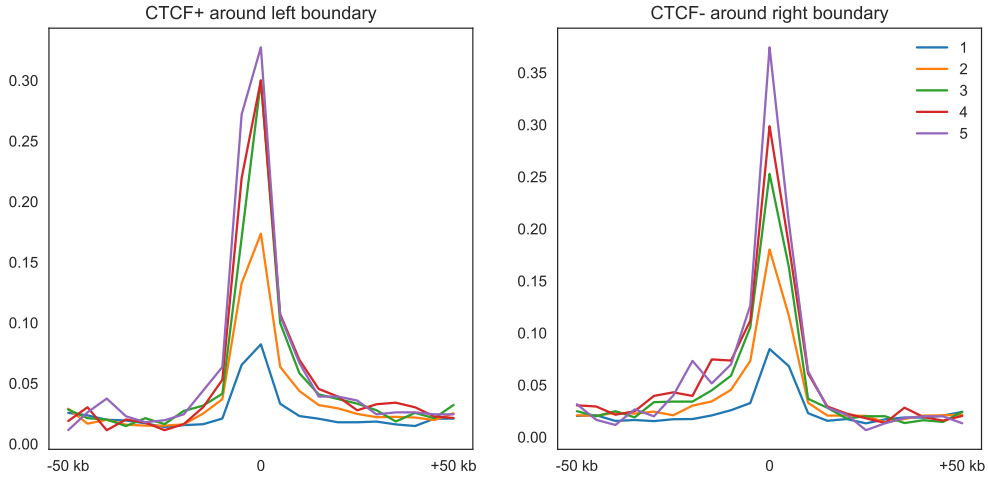
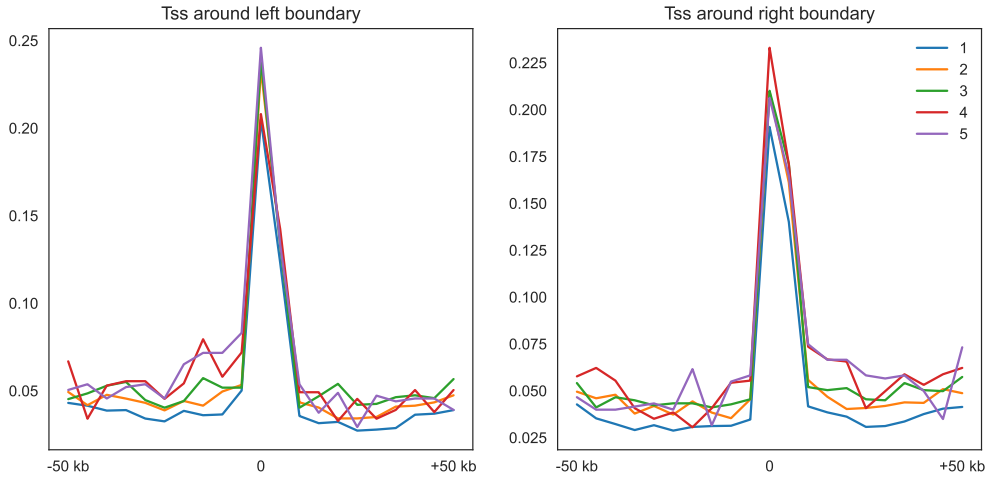


Figure 4.14: **Aggregate peak analysis (APA) at TAD corners for each TAD group identified from the combined GM12878 Hi-C data.**



(a)



(b)

Figure 4.15: **Enrichment of CTCF binding sites and activated promoters around domain boundaries.** We classify a boundary into one of five groups based on the number of times it acts as a domain boundary for different TADs.

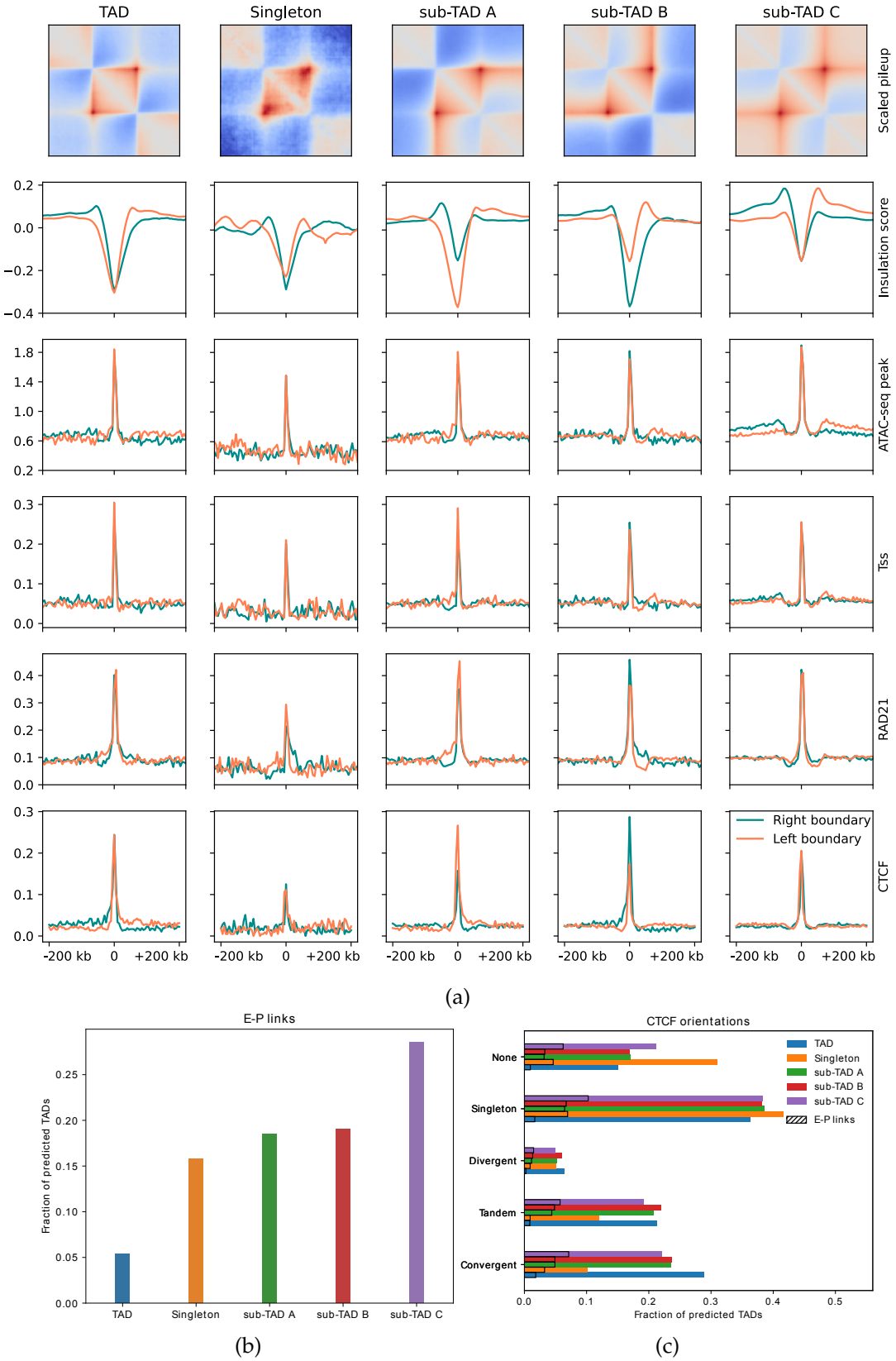


Figure 4.16: **A comparison of singleton TADs, TADs, and subTADs.** **a,** rescaled pileup plots around TADs, and distributions of insulation scores, ATAC-Seq peaks, Tss, RAD21, and CTCF binding sites around domain boundaries. **b,** Proportions of domains being E-P links. **c,** Orientation of CTCF motifs at TAD boundary pairs.

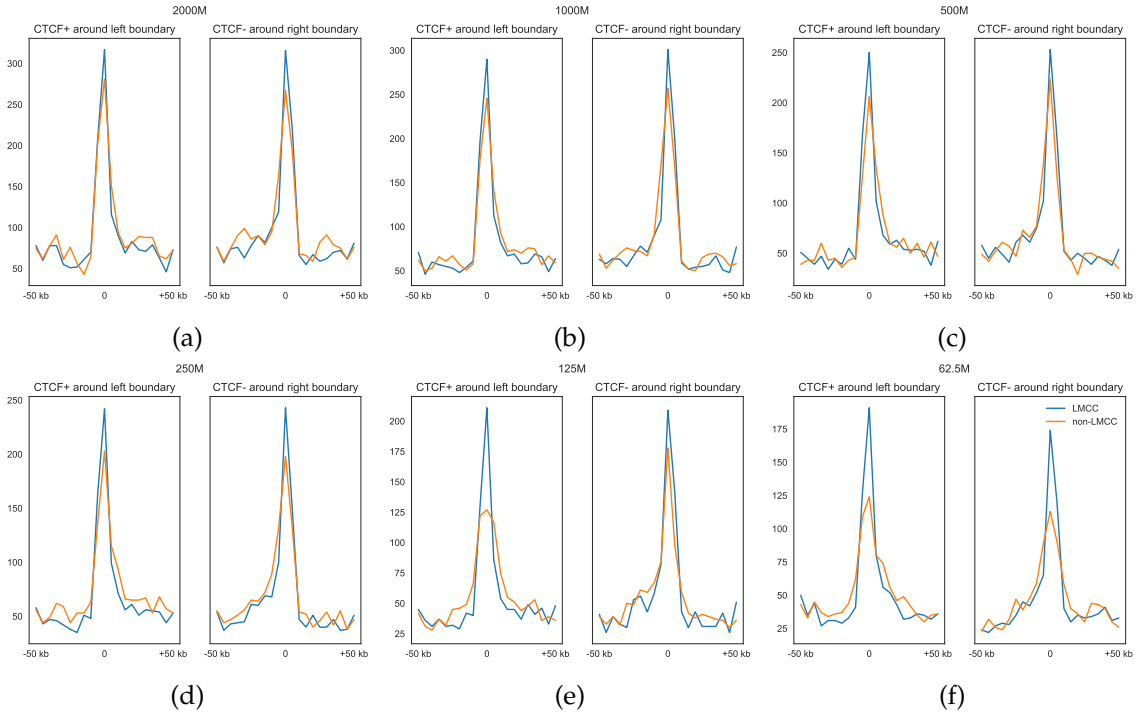


Figure 4.17: **An accuracy comparison of domain boundaries identified by RobusTAD with and without LMCC boundary refinement.** a-f show the occupancy of ChIP-seq identified CTCF binding site as a function of distance to domain boundaries that predicted from Hi-C data containing various number of contact pairs.

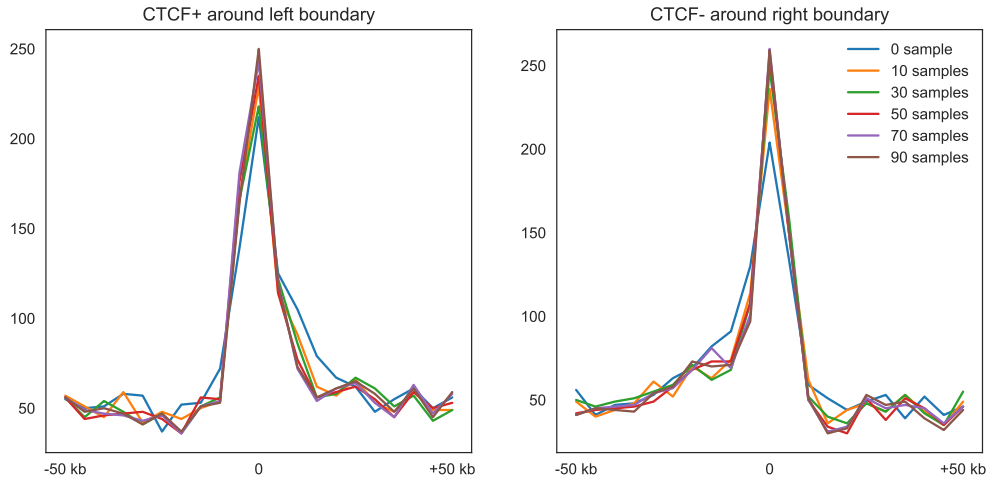


Figure 4.18: **An accuracy comparison of domain boundaries identified by RobusTAD with different number of reference samples.** The two plots show the occupancy of ChIP-seq identified CTCF binding site as a function of distance to domain boundaries that predicted from Hi-C data containing 250M valid read pairs using various samples as a reference panel.

Table 4.1: **Reference Panel**

Accession numbers	Sample	Source
GSM3358191, GSM3358192	22Rv1 (prostate cancer cell line)	[144]
GSM3901271, GSM3901272	293TRex-Flag-BRD4-NUT-HA, treat 1 μg/mL tetracycline for 8 hours	[145]
GSM2631393, GSM2631395	786-M1A cell line (renal cancer cell line)	[185]
GSM2631392, GSM2631394	786-O cell line (renal cancer cell line)	[185]
GSM4198752, GSM4198762	BLaER (lymphoblastic leukemia cell line), CEBPA fused with the estrogen receptor (ER) hormone-binding domain, induced 120hour	[146]
GSM4198753, GSM4198763	BLaER (lymphoblastic leukemia cell line), CEBPA fused with the estrogen receptor (ER) hormone-binding domain, induced 144hour	[146]
GSM4198749, GSM4198759	BLaER (lymphoblastic leukemia cell line), CEBPA fused with the estrogen receptor (ER) hormone-binding domain, induced 48hour	[146]
GSM4198750, GSM4198760	BLaER (lymphoblastic leukemia cell line), CEBPA fused with the estrogen receptor (ER) hormone-binding domain, induced 72hour	[146]
GSM4198751, GSM4198761	BLaER (lymphoblastic leukemia cell line), CEBPA fused with the estrogen receptor (ER) hormone-binding domain, induced 96hour	[146]
GSM4198746, GSM4198756	BLaER (lymphoblastic leukemia cell line), CEBPA fused with the estrogen receptor (ER) hormone-binding domain, induced 9hour	[146]

Table S4.1 continued from previous page

GSM4198768, GSM4198772	BLaER (lymphoblastic leukemia cell line), CTCF-auxin inducible degradation, treat DMSO, 168hour	[146]
GSM3967131, GSM3967132	CUTLL1 (T-ALL cell lines), 1μM DMSO treat every 12h for 72h treat, Arima	[186]
GSM3967126, GSM3967127	CUTLL1 (T-ALL cell lines), 1μM DMSO treat every 12h for 72h, HindIII	[186]
GSM3967129, GSM3967130	CUTLL1 (T-ALL cell lines), 1μM γSI treat every 12 h for 72h	[186]
GSM3967124	Early T-lineage progenitor acute lymphoblastic leukemia (ETP-ALL)	[186]
GSM2825105, GSM2825106	G-401 (kidney cancer cell line)	[187]
GSM3258551	HCC1954 (Breast cancer cell line)	[188]
GSM2809575, GSM2809576, GSM2809577, GSM2809578	HCT-116 (colorectal cancer cell line), RAD21 alleles were tagged with an AID domain and a fluorescent mClover, 6hour axin treat, 180min withdrawal	[147]
GSM3898435, GSM3898437	HCT116 cell, auxin-inducible degron (AID) tag fused to STAG1, auxin treat	[189]
GSM3898434, GSM3898436	HCT116 cell, auxin-inducible degron (AID) tag fused to STAG1, no auxin treat	[189]
GSM3898439, GSM3898441	HCT116 cell, auxin-inducible degron (AID) tag fused to STAG2, auxin treat	[189]
GSM3898438, GSM3898440	HCT116 cell, auxin-inducible degron (AID) tag fused to STAG2, no auxin treat	[189]

Table S4.1 continued from previous page

GSM3489420	HeLa F2 cell, treated for 24 hours with 1000 U/ml of recombinant human IFNg	[190]
GSM2747750	HeLa Kyoto	[191]
GSM4106788	HeLa Kyoto cell, HindIII G1 sync control	[148]
GSM4106796	HeLa Kyoto cell, HindIII G1 sync control, CTCF and ESCO1 siRNA depleted	[148]
GSM4106802	HeLa Kyoto cell, HindIII G1 sync control, CTCF and STAG1 siRNA depleted	[148]
GSM4106795	HeLa Kyoto cell, HindIII G1 sync control, CTCF and STAG2 siRNA depleted	[148]
GSM4106794	HeLa Kyoto cell, HindIII G1 sync control, CTCF siRNA depleted	[148]
GSM4106797	HeLa Kyoto cell, HindIII G1 sync control, ESCO siRNA depleted	[148]
GSM4106792	HeLa Kyoto cell, HindIII G1 sync control, STAG1 siRNA depleted	[148]
GSM4106793	HeLa Kyoto cell, HindIII G1 sync control, STAG2 siRNA depleted	[148]
GSM4106789	HeLa Kyoto cell, MboI G1 sync control	[148]
GSM4106799	HeLa Kyoto cell, MboI G1 sync control, auxin-inducible degron (AID) tag fused to STAG1, auxin treat	[148]
GSM4106798	HeLa Kyoto cell, MboI G1 sync control, auxin-inducible degron (AID) tag fused to STAG1, no auxin treat	[148]

Table S4.1 continued from previous page

GSM4106801	HeLa Kyoto cell, MboI G1 sync control, auxin-inducible degron (AID) tag fused to STAG2, auxin treat	[148]
GSM4106800	HeLa Kyoto cell, MboI G1 sync control, auxin-inducible degron (AID) tag fused to STAG2, no auxin treat	[148]
GSM4106790	HeLa Kyoto cell, MboI G1 sync control, STAG1 siRNA depleted	[148]
GSM4106791	HeLa Kyoto cell, MboI G1 sync control, STAG2 siRNA depleted	[148]
GSM2747751	HeLa Kyoto, CTCF-auxin inducible degradation, 0min	[191]
GSM2747752	HeLa Kyoto, CTCF-auxin inducible degradation, 120min	[191]
GSM2747740	HeLa Kyoto, Pds5SA/B depleted by RNA inference	[191]
GSM2747745, GSM2747748	HeLa Kyoto, Scc1-auxin inducible degradation, 0min	[191]
GSM2747749	HeLa Kyoto, Scc1-auxin inducible degradation, 120min	[191]
GSM2747746	HeLa Kyoto, Scc1-auxin inducible degradation, 15min	[191]
GSM2747747	HeLa Kyoto, Scc1-auxin inducible degradation, 180min	[191]

Table S4.1 continued from previous page

GSM2747753	HeLa Kyoto, Scc1-auxin inducible degradation, WAPL and Pds5SA/B depleted by RNA inference, 0min	[191]
GSM2747754	HeLa Kyoto, Scc1-auxin inducible degradation, WAPL and Pds5SA/B depleted by RNA inference, 15min	[191]
GSM2747755	HeLa Kyoto, Scc1-auxin inducible degradation, WAPL and Pds5SA/B depleted by RNA inference, 180min	[191]
GSM2747738	HeLa Kyoto, synchronized at G1	[191]
GSM2747744	HeLa Kyoto, synchronized at G2	[191]
GSM2747743	HeLa Kyoto, synchronized at S	[191]
GSM2747741	HeLa Kyoto, WAPL and Pds5SA/B depleted by RNA inference	[191]
GSM2747739	HeLa Kyoto, WAPL depleted by RNA inference	[191]
GSM2825569, GSM2825570	HepG2 (hepatocellular carcinoma cell line)	[187]
GSM3304262, GSM3304264	HT1080 (fibrosarcoma cell line)	[192]
GSM2597682, GSM2597683	IMR90 (Lung fibroblast-derived myoblast), control vector	[193]
GSM2597686, GSM2597687	IMR90 (Lung fibroblast-derived myoblast), TET-inducible MYOD, differentiation media	[193]
GSM2597684, GSM2597685	IMR90 (Lung fibroblast-derived myoblast), TET-inducible MYOD, Growth media	[193]
GSM3967128	Jurkat (T-ALL cell lines), 1μM DMSO treat every 12h for 72h	[186]

Table S4.1 continued from previous page

GSM2599093, GSM2599094	MCF10AT1 (hyperplastic breast cell)	[194]
GSM2599095, GSM2599096	MCF10CA1a (fully malignant breast cancer cell)	[194]
GSM3336890, GSM3336891, GSM3336892	MCF-7 (endocrine-sensitive breast cancer cells), endocrine-sensitive ER+ cell	[195]
GSM3336896, GSM3336897,GSM3336898	MCF-7 (endocrine-sensitive breast cancer cells), Fulvestrant-resistant cell	[195]
GSM3756151, GSM3756152	MCF-7 (endocrine-sensitive breast cancer cells), grown without exposure to endocrine therapy, culture 3month	[195]
GSM3756153, GSM3756154	MCF-7 (endocrine-sensitive breast cancer cells), grown without exposure to endocrine therapy, culture 6month	[195]
GSM3756149, GSM3756150	MCF-7 (endocrine-sensitive breast cancer cells), grown without exposure to endocrine therapy, culture start	[195]
GSM3336893, GSM3336894, GSM3336895	MCF-7 (endocrine-sensitive breast cancer cells), Tamoxifen-resistant (TAMR) cell	[195]
GSM3211391	Nalm6 (B cell precursor leukemia cell line)	[150]
GSM3258552	OE33 (Esophegeal adenocarcinoma cell line)	[188]
GSM3967114	Peripheral blood T cells	[186]
GSM4119020, GSM4119025	Primary CD4+ T-cells	[196]
GSM4119022, GSM4119027	Primary CD4+ T-cells, CD3/CD28 stimulated, 1hr	[196]
GSM4119021, GSM4119026	Primary CD4+ T-cells, CD3/CD28 stimulated, 20min	[196]

Table S4.1 continued from previous page

GSM4119024	Primary CD4+ T-cells, CD3/CD28 stimulated, 24hr	[196]
GSM4119023, GSM4119028	Primary CD4+ T-cells, CD3/CD28 stimulated, 4hr	[196]
GSM3392701, GSM3392702	RMG1 (Ovarian clear cell adenocarcinoma cell line), ARID1A Knock Out	[197]
GSM3392703, GSM3392704	RMG1 (Ovarian clear cell adenocarcinoma cell line), NCAPH2 knock Down	[197]
GSM3327706	SNU16 (gastric cancer cell line)	[198]
GSM3258550	SNU-C1 (Colorectal cancer cell line)	[188]
GSM4594449	SW480 (Colorectal cancer cell line), treated with siRNA targeting TCF7L2, 72hour elapsed	[199]
GSM3399745	SW480 (Colorectal cancer cell line)	[188]
GSM3399746	SW480rep1 (Colorectal cancer cell line)	[188]
GSM3258549	SW480rep2 (Colorectal cancer cell line)	[188]
GSM3333325	T2000877 (gastric cancer cell line), CCNE1-rearranged gastric cancer cell line	[198]
GSM3044586, GSM3044588, GSM3044590	T-47D (ductal carcinoma cell line)	[200]
GSM3044586, GSM3044588, GSM3044592	T-47D (ductal carcinoma cell line), treat 110 mM NaCl, 1hour	[200]
GSM3356360	T990275 (gastric cancer cell line), CCNE1-rearranged gastric cancer cell line	[198]
GSM3735784, GSM3735785	WI38 Primary Fibrolabsts, replicative senescence - proliferative	[151]

Table S4.1 continued from previous page

GSM3735786, GSM3735787	WI38 Primary Fibrolabsts, replicative senescence - senescence	[151]
GSM3735782, GSM3735783	WI38 RAF (WI-38hTERT/GFP-RAF1-ER), Oncogene induces senescence day10	[151]
GSM3735776, GSM3735777	WI38 RAF (WI-38hTERT/GFP-RAF1-ER), Oncogene induces senescence day2	[151]
GSM3735778, GSM3735779	WI38 RAF (WI-38hTERT/GFP-RAF1-ER), Oncogene induces senescence day4	[151]
GSM3735788, GSM3735789	WI38 RAF (WI-38hTERT/GFP-RAF1-ER), Oncogene induces senescence day5, treat siDNMT1	[151]
GSM3735790	WI38 RAF (WI-38hTERT/GFP-RAF1-ER), Oncogene induces senescence day5, treat siNT1	[151]
GSM3735780, GSM3735781	WI38 RAF (WI-38hTERT/GFP-RAF1-ER), Oncogene induces senescence day6	[151]
GSM3735774, GSM3735775	WI38 RAF (WI-38hTERT/GFP-RAF1-ER), uninduced	[151]
GSM3417098	MCF10 cell line (ER-/PR- fibrocystic disease)	[201]
GSM3262956, GSM3262957	Embryonic stem cell, Cardiomyocyte differentiation : hESCs (day 0)	[152]
GSM3262962, GSM3262963	Embryonic stem cell, Cardiomyocyte differentiation : cardiac progenitors (day 7)	[152]
GSM3262964, GSM3262965	Embryonic stem cell, Cardiomyocyte differentiation : primitive cardiomyocytes (day 15)	[152]
GSM3262966, GSM3262967	Embryonic stem cell, Cardiomyocyte differentiation : ventricular cardiomyocytes (day 80)	[152]

Table S4.1 continued from previous page

GSM3263085, GSM3263086	Embryonic stem cell	[152]
GSM3263087, GSM3263088	Embryonic stem cell, HERV-H1 Knock-Out, HERV-H elements located TAD boundaries were deleted using CRISPR?Cas9	[152]
GSM3263089, GSM3263090	Embryonic stem cell, HERV-H2 Knock-Out, HERV-H elements located TAD boundaries were deleted using CRISPR?Cas9	[152]
GSM3734958, GSM3734959	Embryonic stem cell, HERV-H2-insertion clone1	[152]
GSM3734960, GSM3734961	Embryonic stem cell, HERV-H2-insertion clone2	[152]
GSM3593256, GSM3593257	teloHAEC (endothelial cell line)	[153]
GSM3593258, GSM3593259	teloHAEC (endothelial cell line), TNFα treated, 4hour	[153]
GSM3560407, GSM3560408	primary white blood cell	-
GSM3560409	primary neutrophil cell	-
GSM3438650, GSM3438651	HUVEC (umblical vein endothelial cells)	[202]
GSM3438652, GSM3438653	HUVEC (umblical vein endothelial cells), treated 10 ng/ml TNF-α, 1hour	[202]
GSM2973922, GSM2973923	ASCs (Adipose-Derived Stem Cells), 0 day of differentiation induction	[203]
GSM2973924, GSM2973925	ASCs (Adipose-Derived Stem Cells), 1 day of differentiation induction	[203]
GSM2973928, GSM2973929	ASCs (Adipose-Derived Stem Cells), 2 days before induction of differentiation	[203]
GSM2973930, GSM2973931	ASCs (Adipose-Derived Stem Cells), 1 day after neuronal induction	[203]

Table S4.1 continued from previous page

GSM2973932, GSM2973933	ASCs (Adipose-Derived Stem Cells), 3 day after neuronal induction	[203]
GSM2410309, GSM2410310	Naïve human embryonic stem cells, growth condition: GSKi + MEKi (2i), Lif, IGF1, FGF	[154]
GSM3506961, GSM3506962, GSM3506963, GSM3506964, GSM3506965, GSM3506966, GSM3506967, GSM3506968, GSM3506969, GSM3506970	GM23248 (primary skin fibroblasts)	[155]
GSM3112369, GSM3112370	HTBE (human tracheobronchial epithelial cells), infect active H5N1 influenza, infection time 12hour	[156]
GSM3112371, GSM3112372	HTBE (human tracheobronchial epithelial cells), infect UV-inactived H5N1 influenza, infection time 12hour	[156]
GSM3112373, GSM3112374	HTBE (human tracheobronchial epithelial cells), infect mock, infection time 12hour	[156]
GSM3112375, GSM3112376	HTBE (human tracheobronchial epithelial cells), infect active H5N1 influenza, infection time 18hour	[156]
GSM3112377, GSM3112378	HTBE (human tracheobronchial epithelial cells), infect UV-inactived H5N1 influenza, infection time 18hour	[156]
GSM3112379, GSM3112380	HTBE (human tracheobronchial epithelial cells), infect mock, infection time 18hour	[156]

Table S4.1 continued from previous page

GSM3112381, GSM3112382	HTBE (human tracheobronchial epithelial cells), infect active H5N1 influenza, infection time 6hour	[156]
GSM3112383, GSM3112384	HTBE (human tracheobronchial epithelial cells), infect UV-inactived H5N1 influenza, infection time 6hour	[156]
GSM3112385, GSM3112386	HTBE (human tracheobronchial epithelial cells), infect mock, infection time 6hour	[156]
GSM3112387, GSM3112388	MDM (monocyte-derived macrophages), infect active H5N1 influenza, infection time 12hour	[156]
GSM3112389, GSM3112390	MDM (monocyte-derived macrophages), infect UV-inactived H5N1 influenza, infection time 12hour	[156]
GSM3112391, GSM3112392	MDM (monocyte-derived macrophages), infect mock, infection time 12hour	[156]
GSM3112395, GSM3112396	MDM (monocyte-derived macrophages), infect UV-inactived H5N1 influenza, infection time 18hour	[156]
GSM3112397, GSM3112398	MDM (monocyte-derived macrophages), infect mock, infection time 18hour	[156]
GSM3112399, GSM3112400, GSM3111878, GSM3111879	MDM (monocyte-derived macrophages), infect active H5N1 influenza, infection time 6hour	[156]
GSM3112401, GSM3112402	MDM (monocyte-derived macrophages), infect UV-inactived H5N1 influenza, infection time 6hour	[156]

Table S4.1 continued from previous page

GSM3112403, GSM3112404, GSM3111876, GSM3111877	MDM (monocyte-derived macrophages), infect mock, infection time 6hour	[156]
GSM3111880, GSM3111881	MDM (monocyte-derived macrophages), infect H5N1-dNS1 influenza, infection time 6hour	[156]
GSM3111882, GSM3111883	MDM (monocyte-derived macrophages), treat IFNb, 6hour	[156]
GSM2816609, GSM2816610	H9 human Embryonic Stem Cell Line, Heat shock condition	[204]
GSM3110157, GSM3110158	MCF10a (epithelial cell line), arrested in G1	[205]
GSM3110159, GSM3110160	MCF10a (epithelial cell line), arrested in G1 and transfected STAG1 siRNA	[205]
GSM3110161, GSM3110162	MCF10a (epithelial cell line), arrested in G1 and transfected STAG2 siRNA	[205]
GSM2595581	HUVEC (Human umbilical vein endothelial cells), donor1	[206]
GSM2595583	HUVEC (Human umbilical vein endothelial cells), donor3	[206]
GSM2595584	IMR90 (fetal lung fibroblast cell), I10	[206]
GSM2595585	IMR90 (fetal lung fibroblast cell), I79	[206]
GSM2595586	MSC (mesenchymal stromal cells)	[206]
GSM2595587	HUVEC (Human umbilical vein endothelial cells), donor1, Oncogenic induced senescence	[206]
GSM2595588	HUVEC (Human umbilical vein endothelial cells), donor2, Oncogenic induced senescence	[206]

Table S4.1 continued from previous page

GSM2595592	MSC (mesenchymal stromal cells), Oncogenic induced senescence	[206]
GSM2845448, GSM2845449	RUES2 (Embryonic stem cells), cardiac differentiation stage : Embryonic stem cells (ESC)	[207]
GSM3452717, GSM3452718	WTC-11 (iPSCs), cardiac differentiation stage : pluripotent stem cells (PSC)	[207]
GSM2627219, GSM2627220	RWPE1 (prostate cell line)	[208]
GSM2828874, GSM2828875	endothelial of hepatic sinusoid primary cell	[187]
GSM2824366, GSM2824367	astrocyte of the cerebellum primary cell	[187]
GSM2247305, GSM2247308	primary epidermal keratinocyte, Differentiation Day 0	[209]
GSM2247306, GSM2247309	primary epidermal keratinocyte, Differentiation Day 3	[209]
GSM2247307, GSM2247310	primary epidermal keratinocyte, Differentiation Day 6	[209]
GSM2494290, GSM2494294, GSM2494298	HAP1 (near-haploid cell line)	[157]
GSM2494291, GSM2494295, GSM2494299	HAP1 (near-haploid cell line), WAPL knock Out	[157]
GSM2494292, GSM2494296, GSM2494300	HAP1 (near-haploid cell line), SSC Knock Out	[157]
GSM2494293, GSM2494297, GSM2494301	HAP1 (near-haploid cell line), WAPL and SSC Knock OUT	[157]
GSM2225739, GSM2225740	purified human naïve B cells	[158]
GSM1267198, GSM1267199	H1 Mesendoderm Cell	[160]

Table S4.1 continued from previous page

GSM1267200, GSM1267201	H1 Mesenchymal Stem Cell	[160]
GSM2437834, GSM2437835, GSM2437836, GSM2437837, GSM2437838, GSM2437839, GSM2437840, GSM2437841	A549 00h 100 nM dexamethasone	[187]
GSM2437749, GSM2437750, GSM2437751, GSM2437752, GSM2437753, GSM2437754, GSM2437755	A549 01h 100 nM dexamethasone	[187]
GSM2437783, GSM2437784, GSM2437785, GSM2437786, GSM2437787, GSM2437788, GSM2437789, GSM2437790	A549 04h 100 nM dexamethasone	[187]
GSM2437857, GSM2437858, GSM2437859, GSM2437860, GSM2437861, GSM2437862, GSM2437863, GSM2437864	A549 08h 100 nM dexamethasone	[187]
GSM2437806, GSM2437807, GSM2437808, GSM2437809, GSM2437810, GSM2437811, GSM2437812, GSM2437813	A549 12h 100 nM dexamethasone	[187]
GSM1551629, GSM1551630, GSM1551631	HUVEC, in-situ MboI	-

Table S4.1 continued from previous page

GSM1551599, GSM1551600, GSM1551601, GSM1551602, GSM1551603, GSM1551604, GSM1551605	IMR90, in-situ MboI	-
GSM1551618, GSM1551619, GSM1551620, GSM1551621, GSM1551622, GSM1551623	K562, in-situ MboI	-
GSM1551624, GSM1551625, GSM1551626, GSM1551627, GSM1551628	KBM7, in-situ MboI	[161]
GSM1551614, GSM1551615, GSM1551616	NHEK, in-situ MboI	[161]
GSM2297252, GSM2297253, GSM2297254, GSM2297255	H1-derived Mesenchymal Stem Cell	[162]

Chapter 5

Discussion and conclusion

5.1 Summary of contributions

In this thesis, we introduced a reference panel enabled framework to computational 3D genomics, revolutionizing the analysis of chromatin structures based on Hi-C contact maps. This is a paradigm shift in 3D genome studies. Although the usage of a reference panel is a common strategy in biological data analysis, its application to Hi-C data analysis remains unexplored. Different cells have different 3D genomes, but often share local sub-structures. The existence of such shared local sub-structures is the foundation of introducing a panel of reference Hi-C samples to facilitate the analysis of a given study sample. However, detecting which reference sample is locally similar to a portion of the study sample is made difficult by the low coverage and systematic bias of both study and reference samples. Within this framework, we have developed three tools that address various challenges in computational 3D genomics. Specifically, we have improved topological structure annotations and contact map enhancement by incorporating a panel of reference Hi-C contact maps into our tools.

In Chapter 2, we introduced RefHiC, the first reference panel enabled application in Hi-C data analysis. It detects topological structures from a Hi-C contact map of interest.

RefHiC combines the encoding-decoding framework and the scaled dot product attention to incorporate a study sample and a reference panel in an end-to-end fashion. This model only extracts information from reference samples that are highly related to the study sample as it learned to assign more weight to reference samples that are more similar to the study sample. In addition, the SHAP values [210] show this model actually focuses more on horizontal and vertical central strips as well as central regions in each patches during prediction. By utilizing a panel of reference Hi-C samples, RefHiC can accurately annotate TADs and chromatin loops from very low-coverage Hi-C contact maps. In our study, we showed that RefHiC significantly outperformed existing tools in annotating TADs and loops under different conditions. As the first reference panel enabled application in Hi-C data analysis, we performed extensive studies to learn the behavior of RefHiC. Our results showed that RefHiC is applicable to any type of cell from species for which a sufficiently large reference panel is available (currently human and mouse). Both rare and common structures can benefit from the introduction of a reference panel. The major contribution of this work includes the introduction of a novel reference panel enabled framework to computational 3D genomics, together with the development of a pretraining and training strategy for deep learning applications in Hi-C data analysis.

In Chapter 3, we proposed RefHiC-SR, a reference panel enabled model for contact map enhancement. This model uses the U-net [112] architecture as a backbone. This backbone allows RefHiC-SR to balance local and global information to predict per-pixel values of a contact map. To integrate a panel of reference samples into the computation graph, we extensively modified the U-net architecture. Briefly, RefHiC-SR projects a study sample and reference samples to embeddings at different scales with its encoding blocks and uses the dot product attention and embedding concatenation to combine information of the study sample and reference samples. Existing contact map enhancement tools are exclusively study sample based and use image super-resolution algorithms to

enhance Hi-C contact maps. They have limited power in enhancing very low coverage Hi-C data and often introduce artifacts in prediction. In contrast, the introduction of a reference panel allows RefHiC-SR to accurately enhance sparse Hi-C contact maps. The major contribution of this work is through developing RefHiC-SR to address a fundamental task (i.e., contact map enhancement), we allow a wide range of downstream tools to directly benefit from the reference panel enabled framework.

In Chapter 4, we demonstrated that the proposed reference panel enabled computational 3D genomics framework is a generic computational framework instead of a deep learning model. We developed RobusTAD using this framework to statistically identify TAD hierarchies from Hi-C contact maps. In detail, we proposed a stratified rank-sum test for scoring and detecting domain boundaries, domains, and nested TADs. Benchmarking evaluation indicates that RobusTAD outperformed existing tools in accurately detecting domain boundaries and entire domains from low to high-coverage Hi-C contact maps.

In summary, this thesis aims to leverage a panel of reference Hi-C samples to address the data insufficiency issue in Hi-C data analysis and consequently improve TAD annotation, loop annotation, and contact map enhancement. Throughout this thesis, we have shown the effectiveness of our reference panel enabled framework. Furthermore, the superior performance achieved by both deep learning and statistical inference applications indicates the proposed framework is a generic strategy for Hi-C data analysis.

5.2 Impact

We anticipate that our framework and applications will have a broad and significant impact in the field of 3D/4D genomics, enabling the analysis of chromatin architecture at unprecedented resolutions and accuracy. As illustrated in previous chapters, applying tools introduced in this thesis allow researchers to identify topological structures more precisely, and investigate 3D genome organization in finer detail even from Hi-C contact

maps with very low sequencing coverage. Our tools enable researchers to investigate 3D genome organization in greater detail, whether by analyzing existing Hi-C data sets or conducting low-coverage Hi-C experiments, thereby substantially reducing sequencing costs and potentially broadening the application of Hi-C technology to other domains. Given that our tools provide accurate high-resolution annotations for topological structures, we anticipate that researchers will better understand the formation and functional roles of TADs and chromatin loops in 3D genome studies. Within this thesis, we have developed applications for annotating TADs and loops from Hi-C contact maps, as well as enhancing Hi-C contact maps. The reference panel based framework can also be employed for other tasks, including developing tools for stripe detection and single-cell Hi-C data analysis.

5.3 Future Work

Despite the significant efforts in developing the framework and applications, there is room for improvement in tools described in this thesis. Applying RefHiC to detect chromatin loops from the entire genome requires 2-3 hours. In contrast, the fastest alternative tool only requires several minutes. The intensive computation involved in loop prediction is in part due to that we evaluate each pixel in a contact map separately. As we consider a $(2 \times w + 1) \times (2 \times w + 1)$ submatrix centered at (i, j) when analyzing pixel (i, j), a large portion of regions are overlapped when we evaluate neighboring pixels. We expect to reduce running time by predicting loops of multiple pixels (i.e., 2×2, 3×3, etc.) around the center of the submatrix instead of a single pixel in the center. Although we have conducted different experiments to demonstrate that our reference panel enabled framework can annotate structures being absent from the reference panel (i.e., rare loop, TAD, etc.), it only confirms our approach can annotate structures that significantly differ

from structures in the reference panel. If there is a loop in the study sample that differs from another loop in the reference panel by only a few bins, our framework might misidentify its location. We did not observe such an issue in practice, but this issue is worth investigating if we have ultra-high coverage data to support this analysis. It can be performed by first comparing loops identified from ChIA-PET experiments and more recent ultra-high coverage Hi-C [211] and micro-C [5] experiments for different cell types at the whole-genome level to collect a list of such differing loops. We can subsequently assess whether our tool can accurately detect them or not. In this thesis, we focused on annotating topological structures purely from Hi-C contact maps and significantly improved model performance. Another direction worth exploring is to integrate Hi-C and other biological data such as sequence motifs and gene expression to comprehend the formation and functional roles of chromatin structures [212, 213, 19]. This type of study can be explored in several ways. First, since epigenetics and sequence grammar such as motifs also dictates TAD boundaries and downstream gene regulation [106], we can improve models proposed in this thesis by taking one dimensional sequence information as additional input to further improve both loop and TAD annotations. This can be achieved by applying multi-modal learning and data integration techniques. Indeed, researchers already combined DNA sequences and Hi-C contact maps to annotate CTCF-medicate loops and achieved more accurate annotations [214]. Second, the correlation between the spatial features of the genome and its functionality still lacks comprehensive understanding. By leveraging our tool to enhance the precision of structural annotations through its application to various Hi-C datasets, we gain the ability to investigate the sequence determinants underlying these structural features. This involves training a deep learning model to predict structures, as identified by our tools, by relying on DNA sequences

or other types experimental data (such as ChIP-Seq, ATAC-Seq, etc.). Subsequently, employing explainable machine learning techniques facilitates the dissection and interpretation of these predictions. We can also represent DNA sequences as graphs induced from RefHiC-annoated loops to design new models to capture long-range interactions in tasks such as gene expression prediction [215].

Finally, despite our methodologies developed in this thesis focus on bulk Hi-C data, this reference panel enabled framework is not limited to performing bulk Hi-C data analysis. We can expand this method to analyze other biological data. In single-cell Hi-C data analysis, applications such as Higashi [60] already improved various single-cell Hi-C analysis tasks by exploiting shared contacts between cells at each contact loci independently. We expect our framework can further improve single-cell Hi-C data analysis as it differs from existing tools by taking both spatial neighboring information within a cell and shared contacts between cells into consideration. A single single-cell Hi-C experiment already produces hundreds to thousands of Hi-C contact maps. The reference panel will be extremely large and our attention approach can be time-consuming. We need to design a new way to define the reference panel as well as integrate the study sample and the reference panel. We anticipate that such an application possesses the potential to better enhance single-cell Hi-C contact maps and improve the accuracy of annotating TAD-like structures and loops from single-cell Hi-C data.

References

1. Bonev, B. & Cavalli, G. Organization and function of the 3D genome. *Nature Reviews Genetics* **17,** 661–678 (2016).
2. Bintu, B. *et al.* Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells. *Science* **362,** eaau1783 (2018).
3. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *science* **326,** 289–293 (2009).
4. Rao, S. S. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159,** 1665–1680 (2014).
5. Krietenstein, N. *et al.* Ultrastructural details of mammalian chromosome architecture. *Molecular cell* **78,** 554–565 (2020).
6. Beagan, J. A. & Phillips-Cremins, J. E. On the existence and functionality of topologically associating domains. *Nature genetics* **52,** 8–16 (2020).
7. Lee, D.-I. & Roy, S. GRiNCH: simultaneous smoothing and detection of topological units of genome organization from sparse chromatin contact count matrices with matrix factorization. *Genome biology* **22,** 1–31 (2021).
8. Hong, H. *et al.* DeepHiC: A Generative Adversarial Network for Enhancing Hi-C Data Resolution. *BioRxiv,* 718148 (2019).
9. Liu, Q., Lv, H. & Jiang, R. hicGAN infers super resolution Hi-C data with generative adversarial networks. *Bioinformatics* **35,** i99–i107 (2019).

10. Hansen, A. S., Cattoglio, C., Darzacq, X. & Tjian, R. Recent evidence that TADs and chromatin loops are dynamic structures. *Nucleus* **9,** 20–32 (2018).

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

12. Fudenberg, G. *et al.* Formation of chromosomal domains by loop extrusion. *Cell reports* **15,** 2038–2049 (2016).

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

14. Zhang, Y., Zhang, X., Dai, H.-Q., Hu, H. & Alt, F. W. The role of chromatin loop extrusion in antibody diversification. *Nature Reviews Immunology* **22,** 550–566 (2022).

15. Flemming, W. *Zellsubstanz, kern und zelltheilung* (Vogel, 1882).

16. Matthey-Doret, C. *et al.* Computer vision for pattern detection in chromosome contact maps. *Nature communications* **11,** 1–11 (2020).

17. Salameh, T. J. *et al.* A supervised learning framework for chromatin loop detection in genome-wide contact maps. *Nature communications* **11,** 1–12 (2020).

18. Roayaei Ardakany, A., Gezer, H. T., Lonardi, S. & Ay, F. Mustache: multi-scale detection of chromatin loops from Hi-C and Micro-C maps using scale-space representation. *Genome biology* **21,** 1–17 (2020).

19. Zhang, Y. *et al.* Computational methods for analysing multiscale 3D genome organization. *Nature Reviews Genetics,* 1–19 (2023).

20. Bianco, S. *et al.* Polymer physics predicts the effects of structural variants on chromatin architecture. *Nature genetics* **50,** 662–667 (2018).

21. Conte, M., Esposito, A., Vercellone, F., Abraham, A. & Bianco, S. Unveiling the Machinery behind Chromosome Folding by Polymer Physics Modeling. *International Journal of Molecular Sciences* **24,** 3660 (2023).

22. Banigan, E. J. & Mirny, L. A. Loop extrusion: theory meets single-molecule experiments. *Current Opinion in Cell Biology* **64,** 124–138 (2020).

23. Zheng, H. & Xie, W. The role of 3D genome organization in development and cell differentiation. *Nature reviews Molecular cell biology* **20,** 535–550 (2019).

24. Marchal, C., Sima, J. & Gilbert, D. M. Control of DNA replication timing in the 3D genome. *Nature Reviews Molecular Cell Biology* **20,** 721–737 (2019).

25. Dubois, F., Sidiropoulos, N., Weischenfeldt, J. & Beroukhim, R. Structural variations in cancer and the 3D genome. *Nature Reviews Cancer* **22,** 533–546 (2022).

26. Yang, L. *et al.* 3D genome alterations associated with dysregulated HOXA13 expression in high-risk T-lineage acute lymphoblastic leukemia. *Nature Communications* **12,** 3708 (2021).

27. Deng, S., Feng, Y. & Pauklin, S. 3D chromatin architecture and transcription regulation in cancer. *Journal of hematology & oncology* **15,** 1–23 (2022).

28. Van Bemmel, J. G. *et al.* The bipartite TAD organization of the X-inactivation center ensures opposing developmental regulation of Tsix and Xist. *Nature genetics* **51,** 1024–1034 (2019).

29. Jerkovic, I. & Cavalli, G. Understanding 3D genome organization by multidisciplinary methods. *Nature Reviews Molecular Cell Biology* **22,** 511–528 (2021).

30. Eskeland, R. *et al.* Ring1B compacts chromatin structure and represses gene expression independent of histone ubiquitination. *Molecular cell* **38,** 452–464 (2010).

31. Patterson, B. *et al.* Female naïve human pluripotent stem cells carry X chromosomes with Xa-like and Xi-like folding conformations. *Science Advances* **9,** eadf2245 (2023).

32. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *science* **295,** 1306–1311 (2002).

33. Simonis, M. *et al.* Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture–on-chip (4C). *Nature genetics* **38,** 1348–1354 (2006).

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

35. Hsieh, T.-H. S. *et al.* Mapping nucleosome resolution chromosome folding in yeast by micro-C. *Cell* **162,** 108–119 (2015).

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

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

38. Belaghzal, H., Dekker, J. & Gibcus, J. H. Hi-C 2.0: An optimized Hi-C procedure for high-resolution genome-wide mapping of chromosome conformation. *Methods* **123,** 56–65 (2017).

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

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

41. Ramani, V. *et al.* Massively multiplex single-cell Hi-C. *Nature methods* **14,** 263–266 (2017).

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

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

44. Zhou, T., Zhang, R. & Ma, J. The 3D genome structure of single cells. *Annual review of biomedical data science* **4,** 21–41 (2021).

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

46. Quinodoz, S. A. *et al.* Higher-order inter-chromosomal hubs shape 3D genome organization in the nucleus. *Cell* **174,** 744–757 (2018).

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

48. Durand, N. C. *et al.* Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell systems* **3,** 95–98 (2016).

49. Abdennur, N. & Mirny, L. A. Cooler: scalable storage for Hi-C data and other genomically labeled arrays. *Bioinformatics* **36,** 311–316 (2020).

50. Servant, N. *et al.* HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome biology* **16,** 259 (2015).

51. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997* (2013).

52. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature methods* **9,** 357–359 (2012).

53. Hu, M. *et al.* HiCNorm: removing biases in Hi-C data via Poisson regression. *Bioinformatics* **28,** 3131–3133 (2012).

54. Kerpedjiev, P. *et al.* HiGlass: web-based visual exploration and analysis of genome interaction maps. *Genome biology* **19,** 1–12 (2018).

55. Robinson, J. T. *et al.* Juicebox. js provides a cloud-based visualization system for Hi-C data. *Cell systems* **6,** 256–258 (2018).

56. Wolff, J. *et al.* Galaxy HiCExplorer: a web server for reproducible Hi-C data analysis, quality control and visualization. *Nucleic acids research* **46,** W11–W16 (2018).

57. Li, D., Harrison, J. K., Purushotham, D. & Wang, T. Exploring genomic data coupled with 3D chromatin structures using the WashU Epigenome Browser. *Nature Methods* **19,** 909–910 (2022).

58. Zhu, X. *et al.* Nucleome browser: an integrative and multimodal data navigation platform for 4D nucleome. *Nature methods* **19,** 911–913 (2022).

59. Zhang, S. *et al.* DeepLoop robustly maps chromatin interactions from sparse allele-resolved or single-cell Hi-C data at kilobase resolution. *Nature Genetics,* 1–13 (2022).

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

61. Ay, F., Bailey, T. L. & Noble, W. S. Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. *Genome research* **24,** 999–1011 (2014).

62. Carty, M. *et al.* An integrated model for detecting significant chromatin interactions from high-resolution Hi-C data. *Nature communications* **8,** 15454 (2017).

63. Dali, R. & Blanchette, M. A critical assessment of topologically associating domain prediction tools. *Nucleic acids research* **45,** 2994–3005 (2017).

64. Szabo, Q., Bantignies, F. & Cavalli, G. Principles of genome folding into topologically associating domains. *Science advances* **5,** eaaw1668 (2019).

65. Shin, H. *et al.* TopDom: an efficient and deterministic method for identifying topological domains in genomes. *Nucleic acids research* **44,** e70–e70 (2016).

66. Filippova, D., Patro, R., Duggal, G. & Kingsford, C. Identification of alternative topological domains in chromatin. *Algorithms for Molecular Biology* **9,** 1–11 (2014).

67. Li, A. *et al.* Decoding topologically associating domains with ultra-low resolution Hi-C data by graph structural entropy. *Nature communications* **9,** 1–12 (2018).

68. Zhan, Y. *et al.* Reciprocal insulation analysis of Hi-C data shows that TADs represent a functionally but not structurally privileged scale in the hierarchical folding of chromosomes. *Genome research* **27,** 479–490 (2017).

69. Crane, E. *et al.* Condensin-driven remodelling of X chromosome topology during dosage compensation. *Nature* **523,** 240–244 (2015).

70. Chen, J., Hero III, A. O. & Rajapakse, I. Spectral identification of topological domains. *Bioinformatics* **32,** 2151–2158 (2016).

71. An, L. *et al.* OnTAD: hierarchical domain structure reveals the divergence of activity among TADs and boundaries. *Genome biology* **20,** 1–16 (2019).

72. Zufferey, M., Tavernari, D., Oricchio, E. & Ciriello, G. Comparison of computational methods for the identification of topologically associating domains. *Genome biology* **19,** 1–18 (2018).

73. Soler-Vila, P., Cusco, P., Farabella, I., Di Stefano, M. & Marti-Renom, M. A. Hierarchical chromatin organization detected by TADpole. *Nucleic acids research* **48,** e39–e39 (2020).

74. Liu, K., Li, H.-D., Li, Y., Wang, J. & Wang, J. A comparison of topologically associating domain callers based on Hi-C data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **20,** 15–29 (2022).

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

76. Liu, T. & Wang, Z. HiCNN: a very deep convolutional neural network to better enhance the resolution of Hi-C data. *Bioinformatics* (2019).

77. Liu, T. & Wang, Z. HiCNN2: enhancing the resolution of Hi-C data using an ensemble of convolutional neural networks. *Genes* **10,** 862 (2019).

78. Cameron, C. J., Dostie, J. & Blanchette, M. Estimating DNA-DNA interaction frequency from Hi-C data at restriction-fragment resolution. *bioRxiv,* 377523 (2018).

79. Zhou, J. *et al.* Robust single-cell Hi-C clustering by convolution-and random-walk–based imputation. *Proceedings of the National Academy of Sciences* **116,** 14011–14018 (2019).

80. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS genetics* **5,** e1000529 (2009).

81. Delaneau, O., Marchini, J. & Zagury, J.-F. A linear complexity phasing method for thousands of genomes. *Nature methods* **9,** 179–181 (2012).

82. Källberg, M. *et al.* Template-based protein structure modeling using the RaptorX web server. *Nature protocols* **7,** 1511–1522 (2012).

83. Browning, B. L., Zhou, Y. & Browning, S. R. A one-penny imputed genome from next-generation reference panels. *The American Journal of Human Genetics* **103,** 338–348 (2018).

84. Pareek, C. S., Smoczynski, R. & Tretyn, A. Sequencing technologies and genome sequencing. *Journal of applied genetics* **52,** 413–435 (2011).

85. Liao, W.-W. *et al.* A draft human pangenome reference. *Nature* **617,** 312–324 (2023).

86. Kuhlman, B. & Bradley, P. Advances in protein structure prediction and design. *Nature Reviews Molecular Cell Biology* **20,** 681–697 (2019).

87. David, A., Islam, S., Tankhilevich, E. & Sternberg, M. J. The AlphaFold database of protein structures: a biologist's guide. *Journal of molecular biology* **434,** 167336 (2022).

88. Martí-Renom, M. A. *et al.* Comparative protein structure modeling of genes and genomes. *Annual review of biophysics and biomolecular structure* **29,** 291–325 (2000).

89. Kaczanowski, S. & Zielenkiewicz, P. Why similar protein sequences encode similar three-dimensional structures? *Theoretical Chemistry Accounts* **125,** 643–650 (2010).

90. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *Journal of molecular biology* **215,** 403–410 (1990).

91. Chenna, R. *et al.* Multiple sequence alignment with the Clustal series of programs. *Nucleic acids research* **31,** 3497–3500 (2003).

92. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596,** 583–589 (2021).

93. Baek, M. *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373,** 871–876 (2021).

94. Yang, Y., Zhang, Y., Ren, B., Dixon, J. R. & Ma, J. Comparing 3D genome organization in multiple species using Phylo-HMRF. *Cell systems* **8,** 494–505 (2019).

95. Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning* http://www.deeplearningbook.org (MIT Press, 2016).

96. Bahdanau, D., Cho, K. & Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).

97. Xu, K. *et al. Show, attend and tell: Neural image caption generation with visual attention* in *International conference on machine learning* (2015), 2048–2057.

98. Niu, Z., Zhong, G. & Yu, H. A review on the attention mechanism of deep learning. *Neurocomputing* **452,** 48–62 (2021).

99. Vaswani, A. *et al.* Attention is all you need. *Advances in neural information processing systems* **30** (2017).

100. Baltrušaitis, T., Ahuja, C. & Morency, L.-P. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence* **41,** 423–443 (2018).

101. Zhang, A., Lipton, Z. C., Li, M. & Smola, A. J. Dive into deep learning. *arXiv preprint arXiv:2106.11342* (2021).

102. Sapoval, N. *et al.* Current progress and open challenges for applying deep learning across the biosciences. *Nature Communications* **13,** 1728 (2022).

103. Johansson-Åkhe, I. & Wallner, B. Improving peptide-protein docking with AlphaFold-Multimer using forced sampling. *Frontiers in Bioinformatics* **2,** 85 (2022).

104. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379,** 1123–1130 (2023).

105. Elofsson, A. Progress at protein structure prediction, as seen in CASP15. *Current Opinion in Structural Biology* **80,** 102594 (2023).

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

107. Cui, H. *et al.* scGPT: Towards Building a Foundation Model for Single-Cell Multi-omics Using Generative AI. *bioRxiv.* eprint: https://www.biorxiv.org/content/early/2023/07/02/2023.04.30.538439.full.pdf. https://www.biorxiv.org/content/early/2023/07/02/2023.04.30.538439 (2023).

108. Chen, J. *et al.* Transformer for one stop interpretable cell type annotation. *Nature Communications* **14,** 223 (2023).

109. Yang, F. *et al.* scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. *Nature Machine Intelligence* **4,** 852–866 (2022).

110. Tan, J. *et al.* Cell-type-specific prediction of 3D chromatin organization enables high-throughput in silico genetic screening. *Nature biotechnology,* 1–11 (2023).

111. Murtaza, G., Wagner, J., Zook, J. M. & Singh, R. GrapHiC: An integrative graph based approach for imputing missing Hi-C reads. *bioRxiv,* 2022–10 (2022).

112. Ronneberger, O., Fischer, P. & Brox, T. *U-net: Convolutional networks for biomedical image segmentation* in *International Conference on Medical image computing and computer-assisted intervention* (2015), 234–241.

113. Sanborn, A. L. *et al.* Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proceedings of the National Academy of Sciences* **112,** E6456–E6465 (2015).

114. Forcato, M. *et al.* Comparison of computational methods for Hi-C data analysis. *Nature methods* **14,** 679–685 (2017).

115. Dali, R., Bourque, G. & Blanchette, M. RobusTAD: A Tool for Robust Annotation of Topologically Associating Domain Boundaries. *bioRxiv,* 293175 (2018).

116. Peng, J. & Xu, J. RaptorX: exploiting structure information for protein alignment by statistical inference. *Proteins: Structure, Function, and Bioinformatics* **79,** 161–171 (2011).

117. Luong, M.-T., Pham, H. & Manning, C. D. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025* (2015).

118. Tang, Z. *et al.* CTCF-mediated human 3D genome architecture reveals chromatin topology for transcription. *Cell* **163,** 1611–1627 (2015).

119. Consortium, E. P. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57 (2012).

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

121. Rao, S. S. *et al.* Cohesin loss eliminates all loop domains. *Cell* **171,** 305–320 (2017).

122. Bonev, B. *et al.* Multiscale 3D genome rewiring during mouse neural development. *Cell* **171,** 557–572 (2017).

123. Wang, X.-T., Cui, W. & Peng, C. HiTAD: detecting the structural and functional hierarchies of topologically associating domains from chromatin interactions. *Nucleic acids research* (2017).

124. Roayaei Ardakany, A. & Lonardi, S. *Efficient and accurate detection of topologically associating domains from contact maps* in *17th International Workshop on Algorithms in Bioinformatics (WABI 2017)* (2017).

125. Yu, W., He, B. & Tan, K. Identifying topologically associating domains and subdomains by Gaussian mixture model and proportion test. *Nature communications* **8,** 1–9 (2017).

126. Lévy-Leduc, C., Delattre, M., Mary-Huard, T. & Robin, S. Two-dimensional segmentation for analyzing Hi-C data. *Bioinformatics* **30,** i386–i392 (2014).

127. Haddad, N., Vaillant, C. & Jost, D. IC-Finder: inferring robustly the hierarchical organization of chromatin folding. *Nucleic acids research* **45,** e81–e81 (2017).

128. Gao, T., Yao, X. & Chen, D. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821* (2021).

129. Dsouza, K. B. *et al.* Learning representations of chromatin contacts using a recurrent neural network identifies genomic drivers of conformation. *Nature Communications* **13,** 1–19 (2022).

130. Rodriguez, A. & Laio, A. Clustering by fast search and find of density peaks. *science* **344,** 1492–1496 (2014).

131. Virtanen, P. *et al.* SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* **17,** 261–272 (2020).

132. Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).

133. Lin, T.-Y., Goyal, P., Girshick, R., He, K. & Dollár, P. *Focal loss for dense object detection* in *Proceedings of the IEEE international conference on computer vision* (2017), 2980–2988.

134. You, Y., Gitman, I. & Ginsburg, B. Large batch training of convolutional networks. *arXiv preprint arXiv:1708.03888* (2017).

135. Kruse, K., Hug, C. B. & Vaquerizas, J. M. FAN-C: a feature-rich framework for the analysis and visualisation of chromosome conformation capture data. *Genome biology* **21,** 1–19 (2020).

136. Venev, S. *et al. open2c/cooltools: v0.4.1* version v0.4.1. Aug. 2021. https://doi.org/10.5281/zenodo.5214125.

137. Karolchik, D. *et al.* The UCSC Table Browser data retrieval tool. *Nucleic acids research* **32,** D493–D496 (2004).

138. Paszke, A. *et al.* in *Advances in Neural Information Processing Systems 32* (eds Wallach, H. *et al.*) 8024–8035 (Curran Associates, Inc., 2019). http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

139. McKinney, W. *et al. Data structures for statistical computing in python* in *Proceedings of the 9th Python in Science Conference* **445** (2010), 51–56.

140. Harris, C. R. *et al.* Array programming with NumPy. *Nature* **585,** 357–362. https://doi.org/10.1038/s41586-020-2649-2 (Sept. 2020).

141. Van der Walt, S. *et al.* scikit-image: image processing in Python. *PeerJ* **2,** e453. ISSN: 2167-8359. https://doi.org/10.7717/peerj.453 (June 2014).

142. Yang, T. *et al.* HiCRep: assessing the reproducibility of Hi-C data using a stratum-adjusted correlation coefficient. *Genome research* **27,** 1939–1949 (2017).

143. Lin, D., Sanders, J. & Noble, W. S. HiCRep. py: fast comparison of Hi-C contact matrices in python. *Bioinformatics* **37,** 2996–2997 (2021).

144. Guo, Y. *et al.* CRISPR-mediated deletion of prostate cancer risk-associated CTCF loop anchors identifies repressive chromatin loops. *Genome Biol* **19,** 160 (Oct. 2018).

145. Rosencrance, C. D. *et al.* Chromatin Hyperacetylation Impacts Chromosome Folding by Forming a Nuclear Subcompartment. *Mol Cell* **78,** 112–126 (Apr. 2020).

146. Stik, G. *et al.* CTCF is dispensable for immune cell transdifferentiation but facilitates an acute inflammatory response. *Nat Genet* **52,** 655–661 (July 2020).

147. Rao, S. S. P. *et al.* Cohesin Loss Eliminates All Loop Domains. *Cell* **171,** 305–320 (Oct. 2017).

148. Wutz, G. *et al.* from WAPL. *Elife* **9** (Feb. 2020).

149. Jacobson, E. C. *et al.* Migration through a small pore disrupts inactive chromatin organization in neutrophil-like cells. *BMC Biol* **16,** 142 (Nov. 2018).

150. Tian, L. *et al.* Long-read sequencing unveils IGH-DUX4 translocation into the silenced IGH allele in B-cell acute lymphoblastic leukemia. *Nat Commun* **10,** 2789 (June 2019).

151. Sati, S. *et al.* 4D Genome Rewiring during Oncogene-Induced and Replicative Senescence. *Mol Cell* **78,** 522–538 (May 2020).

152. Zhang, Y. *et al.* Transcriptionally active HERV-H retrotransposons demarcate topologically associating domains in human pluripotent stem cells. *Nat Genet* **51,** 1380–1388 (Sept. 2019).

153. Lalonde, S. *et al.* Integrative analysis of vascular endothelial cell genomic features identifies AIDA as a coronary artery disease candidate gene. *Genome Biol* **20,** 133 (July 2019).

154. Battle, S. L. *et al.* Enhancer Chromatin and 3D Genome Architecture Changes from Naive to Primed Human Embryonic Stem Cell States. *Stem Cell Reports* **12,** 1129–1144 (May 2019).

155. Nir, G. *et al.* Walking along chromosomes with super-resolution imaging, contact maps, and integrative modeling. *PLoS Genet* **14,** e1007872 (Dec. 2018).

156. Heinz, S. *et al.* Transcription Elongation Can Affect Genome 3D Structure. *Cell* **174,** 1522–1536 (Sept. 2018).

157. Haarhuis, J. H. I. *et al.* The Cohesin Release Factor WAPL Restricts Chromatin Loop Extension. *Cell* **169,** 693–707 (May 2017).

158. Bunting, K. L. *et al.* Multi-tiered Reorganization of the Genome during B Cell Affinity Maturation Anchored by a Germinal Center-Specific Locus Control Region. *Immunity* **45,** 497–512 (Sept. 2016).

159. Leung, D. *et al.* Integrative analysis of haplotype-resolved epigenomes across human tissues. *Nature* **518,** 350–354 (Feb. 2015).

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

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

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

163. Monahan, K., Horta, A. & Lomvardas, S. LHX2-and LDB1-mediated trans interactions regulate olfactory receptor choice. *Nature* **565,** 448–453 (2019).

164. Vian, L. *et al.* The energetics and physiological impact of cohesin extrusion. *Cell* **173,** 1165–1178 (2018).

165. Kubo, N. *et al.* Promoter-proximal CTCF binding promotes distal enhancer-dependent gene activation. *Nature structural & molecular biology* **28,** 152–161 (2021).

166. Yan, J. *et al.* Histone H3 lysine 4 monomethylation modulates long-range chromatin interactions at enhancers. *Cell research* **28,** 204–220 (2018).

167. Huang, H. *et al.* CTCF mediates dosage-and sequence-context-dependent transcriptional insulation by forming local chromatin domains. *Nature genetics* **53,** 1064–1074 (2021).

168. Lindeman, R. E. *et al.* The conserved sex regulator DMRT1 recruits SOX9 in sexual cell fate reprogramming. *Nucleic acids research* **49,** 6144–6164 (2021).

169. Yan, X. *et al. Towards content-independent multi-reference super-resolution: Adaptive pattern matching and feature aggregation* in *European conference on computer vision* (2020), 52–68.

170. Xia, B. *et al.* Coarse-to-Fine Embedded PatchMatch and Multi-Scale Dynamic Aggregation for Reference-based Super-Resolution. *arXiv preprint arXiv:2201.04358* (2022).

171. Borgeaud, S. *et al.* Improving language models by retrieving from trillions of tokens. *arXiv preprint arXiv:2112.04426* (2021).

172. Zhang, Y. & Blanchette, M. Reference panel guided topological structure annotation of Hi-C data. *Nature Communications* **13,** 7426 (2022).

173. Imakaev, M. *et al.* Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nature methods* **9,** 999 (2012).

174. Chen, F., Li, G., Zhang, M. Q. & Chen, Y. HiCDB: a sensitive and robust method for detecting contact domain boundaries. *Nucleic acids research* **46,** 11239–11250 (2018).

175. Li, Y., Willer, C., Sanna, S. & Abecasis, G. Genotype imputation. *Annual review of genomics and human genetics* **10,** 387–406 (2009).

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

177. Nussinov, R. & Jacobson, A. B. Fast algorithm for predicting the secondary structure of single-stranded RNA. *Proceedings of the National Academy of Sciences* **77,** 6309–6313 (1980).

178. Weinreb, C. & Raphael, B. J. Identification of hierarchical chromatin domains. *Bioinformatics* **32,** 1601–1609 (2016).

179. Zhang, Y. W., Wang, M. B. & Li, S. C. SuperTAD: robust detection of hierarchical topologically associated domains with optimized structural information. *Genome biology* **22,** 1–20 (2021).

180. McInnes, L., Healy, J. & Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).

181. Flyamer, I. M., Illingworth, R. S. & Bickmore, W. A. Coolpup. py: versatile pile-up analysis of Hi-C data. *Bioinformatics* **36,** 2980–2985 (2020).

182. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12,** 2825–2830 (2011).

183. An, L. *et al.* OnTAD: hierarchical domain structure reveals the divergence of activity among TADs and boundaries. *Genome biology* **20,** 1–16 (2019).

184. Open2C *et al.* Pairtools: from sequencing data to chromosome contacts. *bioRxiv,* 2023–02 (2023).

185. Rodrigues, P. *et al.* B-Dependent Lymphoid Enhancer Co-option Promotes Renal Carcinoma Metastasis. *Cancer Discov* **8,** 850–865 (July 2018).

186. Kloetgen, A. *et al.* Three-dimensional chromatin landscapes in T cell acute lymphoblastic leukemia. *Nat Genet* **52,** 388–400 (Apr. 2020).

187. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (Sept. 2012).

188. Akdemir, K. C. *et al.* Disruption of chromatin folding domains by somatic genomic rearrangements in human cancer. *Nature genetics* **52,** 294–305 (2020).

189. Casa, V. *et al.* Redundant and specific roles of cohesin STAG subunits in chromatin looping and transcriptional control. *Genome Res* **30,** 515–527 (Apr. 2020).

190. Raviram, R. *et al.* Analysis of 3D genomic interactions identifies candidate host genes that transposable elements potentially regulate. *Genome Biol* **19,** 216 (Dec. 2018).

191. Wutz, G. *et al.* Topologically associating domains and chromatin loops depend on cohesin and are regulated by CTCF, WAPL, and PDS5 proteins. *EMBO J* **36,** 3573–3599 (Dec. 2017).

192. Kantidze, O. L. *et al.* The anti-cancer drugs curaxins target spatial genome organization. *Nat Commun* **10,** 1441 (Mar. 2019).

193. Dall'Agnese, A. *et al.* Transcription Factor-Directed Re-wiring of Chromatin Architecture for Somatic Cell Nuclear Reprogramming toward trans-Differentiation. *Mol Cell* **76,** 453–472 (Nov. 2019).

194. Fritz, A. J. *et al.* Intranuclear and higher-order chromatin organization of the major histone gene cluster in breast cancer. *J Cell Physiol* **233,** 1278–1290 (Feb. 2018).

195. Achinger-Kawecka, J. *et al.* Epigenetic reprogramming at estrogen-receptor binding sites alters 3D chromatin landscape in endocrine-resistant breast cancer. *Nat Commun* **11,** 320 (Jan. 2020).

196. Yang, J. *et al.* Analysis of chromatin organization and gene expression in T cells identifies functional genes for rheumatoid arthritis. *Nat Commun* **11,** 4402 (Sept. 2020).

197. Wu, S. *et al.* ARID1A spatially partitions interphase chromosomes. *Sci Adv* **5,** eaaw5294 (May 2019).

198. Ooi, W. F. *et al.* enhancer hijacking in primary gastric adenocarcinoma. *Gut* **69,** 1039–1052 (June 2020).

199. Brown, M. A. *et al.* TCF7L2 silencing results in altered gene expression patterns accompanied by local genomic reorganization. *Neoplasia* **23,** 257–269 (Feb. 2021).

200. Amat, R. *et al.* Rapid reversible changes in compartments and local chromatin organization revealed by hyperosmotic shock. *Genome Res* **29,** 18–28 (Jan. 2019).

201. Le Dily, F. *et al.* Hormone-control regions mediate steroid receptor-dependent genome organization. *Genome Res* **29,** 29–39 (Jan. 2019).

202. Higashijima, Y. *et al.* Coordinated demethylation of H3K9 and H3K27 is required for rapid inflammatory responses of endothelial cells. *EMBO J* **39,** e103949 (Apr. 2020).

203. Paulsen, J. *et al.* Long-range interactions between topologically associating domains shape the four-dimensional genome during differentiation. *Nat Genet* **51,** 835–843 (May 2019).

204. Lyu, X., Rowley, M. J. & Corces, V. G. Architectural Proteins and Pluripotency Factors Cooperate to Orchestrate the Transcriptional Response of hESCs to Temperature Stress. *Mol Cell* **71,** 940–955 (Sept. 2018).

205. Kojic, A. *et al.* Distinct roles of cohesin-SA1 and cohesin-SA2 in 3D chromosome organization. *Nat Struct Mol Biol* **25,** 496–504 (June 2018).

206. Zirkel, A. *et al.* HMGB2 Loss upon Senescence Entry Disrupts Genomic Organization and Induces CTCF Clustering across Cell Types. *Mol Cell* **70,** 730–744 (May 2018).

207. Bertero, A. *et al.* Dynamics of genome reorganization during human cardiogenesis reveal an RBM20-dependent splicing factory. *Nat Commun* **10,** 1538 (Apr. 2019).

208. Luo, Z., Rhie, S. K., Lay, F. D. & Farnham, P. J. A Prostate Cancer Risk Element Functions as a Repressive Loop that Regulates HOXA13. *Cell Rep* **21,** 1411–1417 (Nov. 2017).

209. Rubin, A. J. *et al.* Lineage-specific dynamic and pre-established enhancer-promoter contacts cooperate in terminal differentiation. *Nat Genet* **49,** 1522–1528 (Oct. 2017).

210. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. *Advances in neural information processing systems* **30** (2017).

211. Harris, H. L. *et al.* Chromatin alternates between A and B compartments at kilobase scale for subgenic organization. *Nature Communications* **14,** 3303 (2023).

212. Feng, F., Yao, Y., Wang, X. Q. D., Zhang, X. & Liu, J. Connecting high-resolution 3D chromatin organization with epigenomics. *Nature communications* **13,** 2054 (2022).

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

214. Wang, F. *et al.* GILoop: Robust chromatin loop calling across multiple sequencing depths on Hi-C data. *Iscience* **25** (2022).

215. Karbalayghareh, A., Sahin, M. & Leslie, C. S. Chromatin interaction–aware gene regulatory modeling with graph attention networks. *Genome Research* **32,** 930–944 (2022).