A Heuristic Strategy for Multi-mapping Reads to Enhance Hi-C Data

Chanaka Bulathsinghalage
Department of Computer Science
North Dakota State University
Fargo, USA
chanaka.cooray@ndsu.edu

Lu Liu*
Department of Computer Science
North Dakota State University
Fargo, USA
lu.liu.2@ndsu.edu

***Abstract*—Current Hi-C analysis approaches focus on uniquely mapped reads and little research has been carried out to include multi-mapping reads, which leads to a lack of biological signals from DNA repetitive regions. We propose a heuristic strategy to assign multi-mapping reads to loci according to the distance to their closest restriction enzyme cutting sites. We demonstrate that the heuristic strategy can rescue multi-mapping reads thus enhance the quality of Hi-C data. Compared with mHi-C, it not only improves replicate reproducibility in the same cell type, but also maintains the difference between replicates of different cell types. Moreover, the strategy identifies much more common statistically significant chromatin interactions between Hi-C experiments of different restriction enzymes and has a huge advantage on computing resources. Therefore, the heuristic strategy can be used to enhance Hi-C data by utilizing multi-mapping reads.**

***Index Terms*—heuristic strategy, Hi-C, multi-mapping reads**

I. Introduction

Three-dimensional genome organization plays important roles in many biological processes, which include long-range gene regulation [1], DNA replication and repair [2], [3]. The alteration of three-dimensional genome architecture leads to human diseases, such as cancer [4], [5]. As the development of chromosome conformation capture-based technologies, high-throughput chromosome conformation capture (Hi-C) [6] emerges as a popular method to detect genome-wide chromatin interactions. In Hi-C experiments, crosslinked DNA is fragmented with restriction enzymes. Then DNA fragments are ligated, selected, sheared and finally sequenced as paired-end reads. After these paired-end reads are processed by Hi-C analysis pipelines, chromatin contact maps are generated for downstream analysis and exploration. Recent studies have discovered some multi-scale spatial genomic structures, which include A/B compartment [6], topologically associating domains (TADs) [7], chromatin loops [8] and frequently interacting regions (FIREs) [9].

Owing to the sequencing cost, few studies generate high-resolution data sets. To enable high-resolution structure discovery on low-resolution data sets, some computational methods are proposed to enhance Hi-C data with machine learning algorithms. HiCPlus [10] and HiCNN [11] both use deep convolutional neural networks. HicGAN [12] and DeepHiC [13] infers high-resolution Hi-C data with generative adversarial networks. However, all of these methods depend on one high-resolution data set as their training sets and ignore heterogeneity among cell types.

Though machine learning algorithms are popular, they are not the only method to enhance Hi-C data. In fact, for each Hi-C data set, a large number of reads are discarded at the very beginning. Because most Hi-C pipelines only consider uniquely mapped reads (unique reads) and ignore multi-mapping reads, which are mapped to multiple genomic loci. To the best of our knowledge, there is only one study, mHiC [14], accounting for multi-mapping reads. mHi-C assigns multi-mapping reads according to the interacting patterns learned from unique reads, therefore the multi-mapping read assignment depends on unique reads. Here we propose a heuristic strategy which doesn't depend on unique reads to utilize multi-mapping reads. The heuristic strategy not only enhances Hi-C data, but also enables exploration of new interacting patterns.

Our contributions may be stated as follows:

- We propose a heuristic strategy to utilize multi-mapping reads for Hi-C data processing.
- We demonstrate that using our proposed strategy on Hi-C data sets can enhance Hi-C data in quantity and reproducibility and recover more common statistically significant chromatin interactions between experiments of different restriction enzymes.

The rest of paper is organized as follows. The second section delineates the heuristic strategy to use multi-mapping reads. The third section introduces two human cell lines and two Arabidopsis data sets as our test data. The fourth section evaluates the heuristic strategy by comparing it with mHi-C and a method that only considers unique reads. The last one concludes that the heuristic strategy complements multi-mapping reads in Hi-C analysis.

II. Method

We propose a heuristic strategy to utilize multi-mapping reads in Hi-C experiments to strengthen chromatin interaction data. As shown in Figure 1A, for Hi-C read ends, there are three possible outcomes, unaligned, unique and multi-mapping reads. Compared with unaligned reads, multi-mapping reads

*Corresponding author.

are reads with high quality alignment scores, but their alignment loci cannot be uniquely determined. To avoid the abuse of utilizing multi-mapping reads, we only rescue multi-mapping reads with less than a specific number of alignments. For example, mHi-C by default utilizes multi-mapping reads with less than 100 alignments. In order to assign a multi-mapping read to a unique locus among its alignments, we hypothesize that the locus closer to restriction enzyme cutting sites has a higher probability to be the origin as shown in Figure 1B. The hypothesis is based on the Hi-C processing of unique reads. In Hi-C processing pipelines, the closest restriction enzyme cutting sites are picked to filter unique reads. Second according to our empirical experience, an object's breakage because of outside forces is most likely to happen at the object's periphery with defects. In Hi-C experiments at the shearing step, shearing may happen preferentially close to the restriction enzyme cutting sites, which can be viewed as defects as these sites are cut by restriction enzymes before. Therefore, we select the loci for multi-mapping reads according to the distance to the closest restriction enzyme cutting sites. What is more important, as our multi-mapping read assignment is carried out at the sequence alignment step, there is no impact on following Hi-C data processing and the same filtering criteria (such as distance to restriction enzyme cutting sites) can be applied to unique and multi-mapping reads to remove invalid chromatin interactions.

III. Data

To demonstrate that the heuristic strategy can rescue multi-mapping reads in Hi-C experiments, thus increasing detected chromatin interactions and expanding the breadth of genome coverage, we test the strategy on Hi-C experiments of two cell lines from a study [7] on revealing topological domains in mammalian genomes and Hi-C experiments of Arabidopsis thaliana seedling tissues from two studies [15], [16] with different restriction enzymes. The first cell line is human embryonic stem cell (hESC) and the second cell line is derived from human fetal lung (IMR90). For each cell line, Hi-C experiments were conducted independently with two biological replicates (r1 and r2) using HindIII as the restriction enzyme to cut crosslinked DNA into fragments. Thereafter, DNA fragments in close proximity were ligated in a diluted environment and the resulting ligation products were sonicated, filtered and finally sequenced by paired-end sequencing. Therefore, two paired read files were generated for each replicate, e.g. hESC_r1_1 and hESC_r1_2. For Arabidopsis thaliana seedling tissues, the first study [15] carried out the Hi-C experiments using HindIII with two biological replicates (r1 and r2), which are named HindIII_r1 and HindIII_r2. The second study [16] carried out the Hi-C experiments using DpnII with three biological replicates (r1, r2 and r3), which are named DpnII_r1, DpnII_r2 and DpnII_r3.

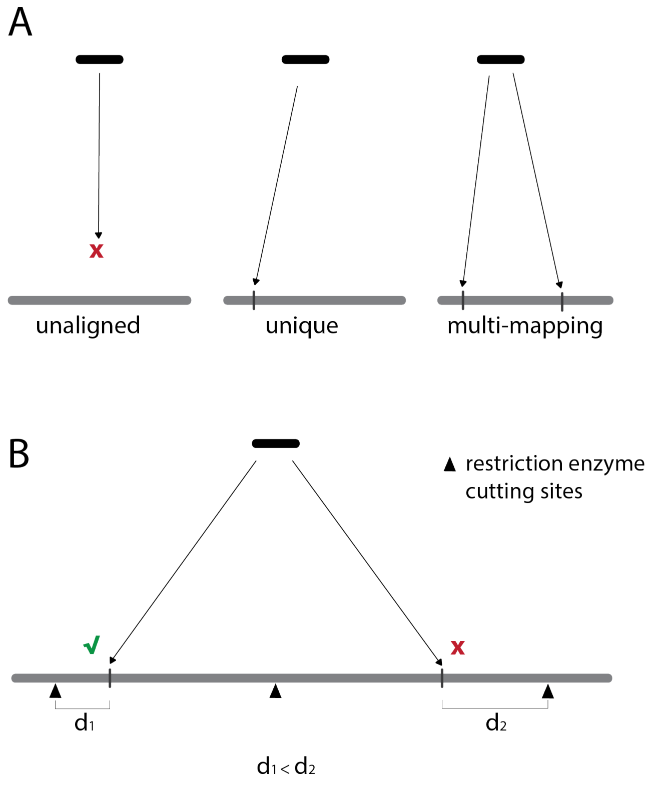

Fig. 1. Hi-C read alignment outcomes and the heuristic strategy for multi-mapping reads. A: three types of reads, unaligned, unique and multi-mapping reads, B: a multi-mapping read is assigned to a locus closest to restriction enzyme cutting sites.

IV. Results

A. Sequence alignment statistics necessitate utilizing multi-mapping reads

We adopt Hi-C processing pipelines consisting of a sequence of processing functions/commands, for example, Hiclib [17], to process paired reads of hESC and IMR90's replicates. Because it is convenient to incorporate the heuristic strategy into these pipelines and understanding the inner complex logic of a holistic tool is not this study's research focus. As Hi-C processing pipelines ignore multi-mapping reads at the sequence alignment step, we need to carry out our own sequence alignment to keep multi-mapping reads. A sequence alignment tool, for example, Bowtie 1 [18], is applied to align two ends of Hi-C reads independently with its default settings and the statistics of sequence alignment for each replicate are listed in Table I. For each replicate, multi-mapping reads are more than unaligned reads at both ends. This means there are more multi-mapping reads than unaligned reads to be rescued. This phenomenon can be explained by the fact that these reads are short reads which are more likely to be aligned to multiple loci than nowhere. In addition, prevalent short-read sequencing in Hi-C experiments necessitates the need of utilizing multi-mapping reads to enhance chromatin interaction data.

TABLE I
hESC AND IMR90 PAIRED-END SEQUENCE ALIGNMENT STATISTICS.TWO ENDS OF HI-C PAIRED-END READS ARE MAPPED INDEPENDENTLY BECAUSE DISTANCE CONSTRAINT OF PAIRED-END READS DOESN'T APPLY TO HI-C READS.

replicate	hESC_r1		hESC_r2		IMR90_r1		IMR90_r2	
#reads	237,662,270		496,522,946		397,194,480		259,123,992	
unique reads(%)	69.77	68.74	72.31	70.96	71.65	69.04	70.44	70.26
unaligned reads(%)	11.99	13.16	9.79	11.45	10.82	13.87	11.74	11.70
multi-mapping reads(%)	**18.24**	**18.10**	**17.9**	**17.59**	**17.53**	**17.09**	**17.82**	**18.04**

B. The heuristic strategy increases detected chromatin interactions

To demonstrate that the heuristic strategy can strengthen chromatin interaction data, we test the strategy on each replicate with hiclib and mHi-C respectively. hiclib only considers unique reads and incorporating our strategy takes both unique and multi-mapping reads into count. mHi-C leverages multi-mapping reads in a sequence of commands and it is convenient to replace its multi-mapping read assignment method with our strategy. The numbers of detected chromatin interactions for each replicate are shown in Table II. Compared with unique reads, the heuristic strategy increases millions of chromatin interactions because it also accounts for multi-mapping reads. Compared with mHi-C, the heuristic strategy gains chromatin interactions marginally because they both leverage unique and multi-mapping reads.

C. The heuristic strategy enhances the reproducibility of chromatin interaction data

Replicate reproducibility is an important measurement used to assess the quality of chromatin interaction data. We calculate the reproducibility scores among hESC and IMR90's replicates by chromosome (from chromosome 1 to chromosome 22) with HiCRep [19]. As shown in Figure 2, for each configuration [mHi-C (unique), mHi-C and mHi-C+], there are two types of replicate reproducibility scores. The first type (at the top) represents the average of replicate reproducibility scores in the same cell line (hESC_r1 VS hESC_r2 and IMR90_r1 VS IMR90_r2). The second type (at the bottom) represents the difference between the average of replicate reproducibility scores in the same cell line and the average of replicate reproducibility scores between different cell lines (hESC_r1 VS IMR90_r1, hESC_r1 VS IMR90_r2, hESC_r2 VS IMR90_r1 and hESC_r2 VS IMR90_r2). For the first type of replicate reproducibility scores, mHi-C and mHi-C+ are better than mHi-C(unique). This means compared with the configuration only utilizing unique reads, configurations utilizing both unique and multi-mapping reads improve the reproducibility between replicates in the same cell line. In addition, mHiC's multi-mapping read assignment method (mHi-C) is slightly better than our strategy (mHi-C+) on improving the reproducibility between replicates in the same cell line. But for the second type of replicate reproducibility scores, our strategy performs better than mHi-C. Among the 22 chromosomes, our strategy has noticeably larger differences on 7 chromosomes, while mHi-C's multi-mapping read assignment method has 2 noticeably larger differences on 2 chromosomes. What is more important, our strategy achieves similar performance with the method only utilizing unique reads. Taking these two types of replicate reproducibility scores into consideration, we conclude that our strategy not only improves the replicate reproducibility in the same cell line, but also maintains the difference between different cell lines.

D. The heuristic strategy improves statistically significant chromatin interactions

Enhanced chromatin interaction data enable downstream analysis and exploration for new discoveries. Therefore, we apply Fit-Hi-C [20] to normalized chromatin interactions to identify statistically significant chromatin interactions with respect to a false discovery rate of 0.05. In Table III, both configurations utilizing unique and multi-mapping reads report more statistically significant chromatin interactions than the configuration utilizing only unique reads. In addition, mHi-C's multi-mapping read assignment method seems identifying more statistically significant chromatin interactions than our strategy. It can be explained if we further examine detected chromatin interactions and keep only unique chromatin interactions. As shown in Table IV, incorporating our strategy gains much more unique chromatin interactions because mHi-C assigns multi-mapping reads according to the interacting patterns in the unique reads. Therefore, interacting patterns in the unique reads would be enriched to be statistically significant. The heuristic strategy doesn't assign multi-mapping reads according to unique reads and consequently it can explore more interacting patterns. However, these dispersed interacting patterns may become less statistically significant.

To further investigate two approaches utilizing multi-mapping reads on identifying statistically significant chromatin interactions, we apply them on Hi-C experiments of Arabidopsis thaliana seedling tissues from two studies [15], [16] with different restriction enzymes, HindIII and DpnII. Fit-Hi-C is used to identify statistically significant chromatin interactions with respect to a false discovery rate of 0.05 for each replicate respectively. Pairwise comparison is carried out between replicates of different restriction enzymes and the common statistically significant chromatin interactions are counted as shown in Table V. Our strategy identifies much more common statistically significant chromatin interactions than mHi-C (>32%) because when assigning multi-mapping reads, our strategy does not depend on unique reads and therefore improving the identification of common statistically significant chromatin interactions.

TABLE II

hESC and IMR90 chromatin interactions with hiclib and mHi-C under different configurations. hiclib+ represents incorporating hiclib with the heuristic strategy. mHi-C(unique) represents limiting mHi-C to unique reads. mHi-C+ represents replacing mHi-C's multi-mapping read assignment method with the heuristic strategy.

method	hiclib	hiclib+	mHi-C(unique)	mHi-C	mHi-C+
hESC_r1	16,156,824	**21,528,337**	17,043,308	20,325,529	**20,819,070**
hESC_r2	117,150,577	**139,527,552**	105,617,771	124,622,391	**124,955,453**
IMR90_r1	81,524,268	**97,985,497**	83,161,703	97,444,579	**98,380,530**
IMR90_r2	89,322,274	**104,647,014**	83,381,123	96,099,798	**98,325,832**

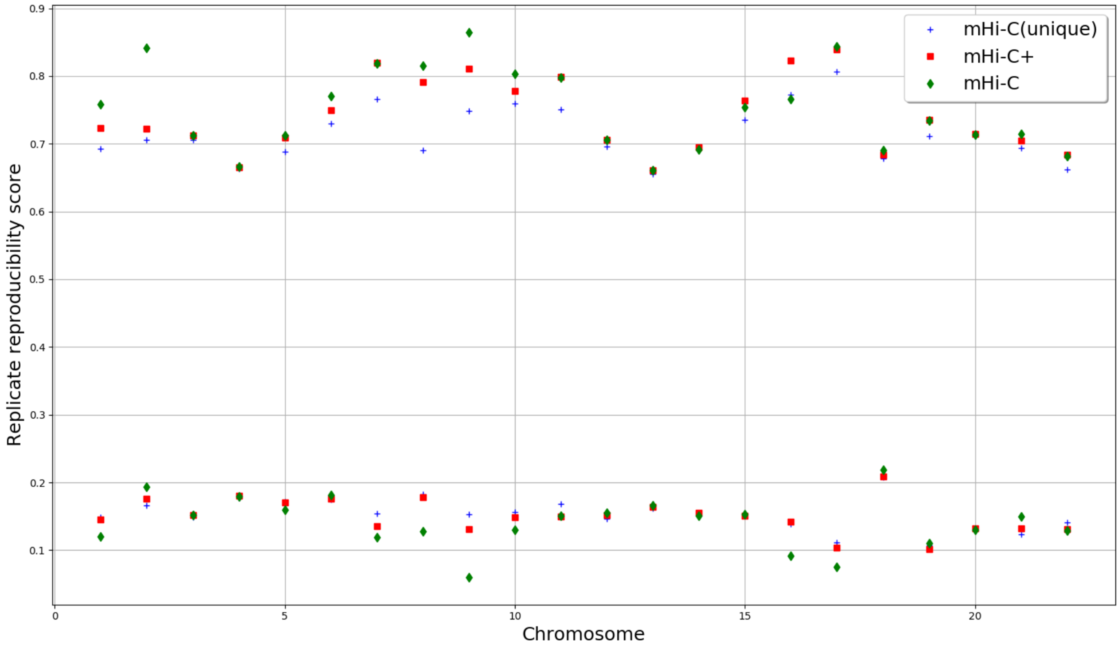

Fig. 2. Replicate reproducibility scores for human chromosome 1-22. HiCRep is used to calculate reproducibility scores among hESC and IMR90's replicates. For each configuration [mHi-C(unique), mHi-C and mHi-C+], there are two types of replicate reproducibility scores. The first type (at the top) represents the average of replicate reproducibility scores in the same cell line. The second type (at the bottom) represents the difference between the average of replicate reproducibility scores in the same cell line and the average of replicate reproducibility scores between different cell lines.

TABLE III

Statistically significant chromatin interactions identified by Fit-Hi-C. mHi-C(unique) represents limiting mHi-C to unique reads. mHi-C+ represents replacing mHi-C's multi-mapping read assignment method with the heuristic strategy.

method	mHi-C(unique)	mHi-C	mHi-C+
hESC_r1	4,206	8,412	7,226
hESC_r2	34,630	54,642	53,236
IMR90_r1	49,500	78,574	69,476
IMR90_r2	55,124	85,160	74,396

TABLE IV

hESC and IMR90's unique chromatin interactions with mHi-C under different configurations. mHi-C(unique) represents limiting mHi-C to unique reads. mHi-C+ represents replacing mHi-C's multi-mapping read assignment method with the heuristic strategy.

method	mHi-C(unique)	mHi-C	mHi-C+
hESC_r1	11,589,365	12,696,565	**14,656,936**
hESC_r2	48,065,862	51,792,951	**61,564,215**
IMR90_r1	54,974,139	58,975,514	**66,763,164**
IMR90_r2	63,548,605	67,705,423	**76,033,914**

E. The heuristic strategy has a huge advantage on computing resources

Computing resources are essential to bioinformatics research, especially for researchers and students with a limited budget. We compare the running time and memory usage on the same computing resource. As some commands (such as sequence alignment) in the pipeline are shared under different configurations, we only summarize the computing resources pertaining to the multi-mapping read assignment in Figure 3. mHi-C's multi-mapping read assignment method takes at least

TABLE V
COMMON STATISTICALLY SIGNIFICANT CHROMATIN INTERACTIONS ON ARABIDOPSIS THALIANA HI-C EXPERIMENT. HindIII and DpnII were used on Arabidopsis thaliana seedling tissues. Pairwise comparision between replicates of different restriction enzymes is carried out.

mHiC VS mHiC+	DpnII_r1	DpnII_r2	DpnII_r3
HindIII_r1	1561, **2064**	2079, **2838**	2067, **2877**
HindIII_r2	2020, **3250**	2817, **4083**	2757, **4084**

five-fold running time and ten-fold RAM than our strategy. When two configurations are applied to high resolution Hi-C data sets, the difference on computing resources becomes more glaring. Therefore, the heuristic strategy has a huge advantage on computing resources than mHi-C's multi-mapping read assignment method.

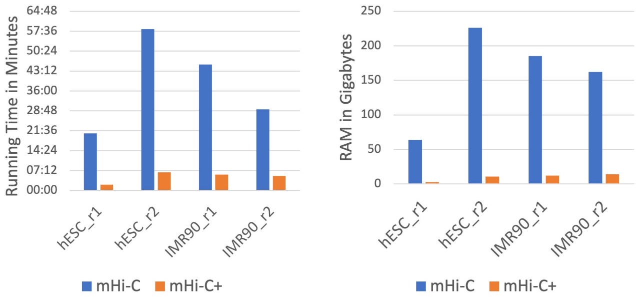

Fig. 3. Comparison of computing resources (runnting time in minutes and RAM in gigabytes) with mHi-C under different configurations.

V. CONCLUSION

In this paper, we introduce a heuristic strategy to include multi-mapping reads into Hi-C analysis by assigning these reads according to the distance to their closest restriction enzyme cutting sites. Through the evaluation of Hi-C human data, we display that there are more multi-mapping reads than unaligned reads to be rescued. Compared with methods only considering unique reads, the strategy improves the quantity and reproducibility of Hi-C data, which enables new discoveries of statistically significant chromatin interactions. Compared with mHi-C, the strategy maintains the difference between replicates of different cell lines, reports more common statistically significant chromatin interactions (>32%) between experiments with different restriction enzymes and shows a huge advantage on computing resources (at least 5-fold in running time and 10-fold in RAM). Therefore, our strategy is an important complement to incorporating Hi-C multi-mapping reads.

Due to most Hi-C reads used in this paper are short reads (36 base pairs), we didn't rescue unaligned reads. For longer sequence reads, more efforts can be extended to study whether Hi-C data can be further enhanced by rescuing both unaligned reads with recursive mapping and multi-mapping reads with our proposed strategy. We also plan to combine our proposed strategy and machine learning algorithms to achieve high-resolution and high coverage Hi-C data.

ACKNOWLEDGMENT

The work is supported by the National Science Foundation under NSF EPSCoR Track-1 Cooperative Agreement OIA #1946202 and used resources of the Center for Computationally Assisted Science and Technology (CCAST) at North Dakota State University.

REFERENCES

[1] J. Dekker and T. Misteli, "Long-range chromatin interactions," *Cold Spring Harbor perspectives in biology*, vol. 7, no. 10, p. a019356, 2015.

[2] J. Ma and Z. Duan, "Replication timing becomes intertwined with 3d genome organization," *Cell*, vol. 176, no. 4, pp. 681–684, 2019.

[3] E. Fabre and C. Zimmer, "From dynamic chromatin architecture to dna damage repair and back," *Nucleus*, vol. 9, no. 1, pp. 161–170, 2018.

[4] G. Fundenberg, G. Getz, M. Meyerson, and L. Mirny, "High-order chromatin architecture determines the landscape of chromosomal alterations in cancer," *Nature precedings, hdl*, vol. 10101, 2011.

[5] C. Anania and D. G. Lupiáñez, "Order and disorder: abnormal 3d chromatin organization in human disease," *Briefings in Functional Genomics*, vol. 19, no. 2, pp. 128–138, 2020.

[6] E. Lieberman-Aiden, N. L. Van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner *et al.*, "Comprehensive mapping of long-range interactions reveals folding principles of the human genome," *science*, vol. 326, no. 5950, pp. 289–293, 2009.

[7] J. R. Dixon, S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. S. Liu, and B. Ren, "Topological domains in mammalian genomes identified by analysis of chromatin interactions," *Nature*, vol. 485, no. 7398, p. 376, 2012.

[8] S. S. Rao, M. H. Huntley, N. C. Durand, E. K. Stamenova, I. D. Bochkov, J. T. Robinson, A. L. Sanborn, I. Machol, A. D. Omer, E. S. Lander *et al.*, "A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping," *Cell*, vol. 159, no. 7, pp. 1665–1680, 2014.

[9] A. D. Schmitt, M. Hu, I. Jung, Z. Xu, Y. Qiu, C. L. Tan, Y. Li, S. Lin, Y. Lin, C. L. Barr *et al.*, "A compendium of chromatin contact maps reveals spatially active regions in the human genome," *Cell reports*, vol. 17, no. 8, pp. 2042–2059, 2016.

[10] Y. Zhang, L. An, J. Xu, B. Zhang, W. J. Zheng, M. Hu, J. Tang, and F. Yue, "Enhancing hi-c data resolution with deep convolutional neural network hicplus," *Nature communications*, vol. 9, no. 1, pp. 1–9, 2018.

[11] T. Liu and Z. Wang, "Hicnn: a very deep convolutional neural network to better enhance the resolution of hi-c data," *Bioinformatics*, vol. 35, no. 21, pp. 4222–4228, 2019.

[12] Q. Liu, H. Lv, and R. Jiang, "hicgan infers super resolution hi-c data with generative adversarial networks," *Bioinformatics*, vol. 35, no. 14, pp. i99–i107, 2019.

[13] H. Hong, S. Jiang, H. Li, G. Du, Y. Sun, H. Tao, C. Quan, C. Zhao, R. Li, W. Li *et al.*, "Deephic: A generative adversarial network for enhancing hi-c data resolution," *PLOS Computational Biology*, vol. 16, no. 2, p. e1007287, 2020.

[14] Y. Zheng, F. Ay, and S. Keles, "Generative modeling of multi-mapping reads with mhi-c advances analysis of hi-c studies," *eLife*, vol. 8, p. e38070, 2019.

[15] H. Zhang, R. Zheng, Y. Wang, Y. Zhang, P. Hong, Y. Fang, G. Li, and Y. Fang, "The effects of arabidopsis genome duplication on the chromatin organization and transcriptional regulation," *Nucleic acids research*, vol. 47, no. 15, pp. 7857–7869, 2019.

[16] M. J. Rowley, M. H. Rothi, G. Böhmdorfer, J. Kuciński, and A. T. Wierzbicki, "Long-range control of gene expression via rna-directed dna methylation," *PLoS genetics*, vol. 13, no. 5, p. e1006749, 2017.

[17] M. Imakaev, G. Fudenberg, R. P. McCord, N. Naumova, A. Goloborodko, B. R. Lajoie, J. Dekker, and L. A. Mirny, "Iterative correction of hi-c data reveals hallmarks of chromosome organization," *Nature methods*, vol. 9, no. 10, p. 999, 2012.

[18] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg, "Ultrafast and memory-efficient alignment of short dna sequences to the human genome," *Genome biology*, vol. 10, no. 3, p. R25, 2009.

[19] T. Yang, F. Zhang, G. G. Yardımcı, F. Song, R. C. Hardison, W. S. Noble, F. Yue, and Q. Li, "Hicrep: assessing the reproducibility of hi-c data using a stratum-adjusted correlation coefficient," *Genome research*, vol. 27, no. 11, pp. 1939–1949, 2017.

[20] F. Ay, T. L. Bailey, and W. S. Noble, “Statistical confidence estimation for hi-c data reveals regulatory chromatin contacts,” *Genome research*, vol. 24, no. 6, pp. 999–1011, 2014.