UNIVERSITY OF CALIFORNIA
RIVERSIDE

Efficient Algorithms for the Analysis of Hi-C Contact Maps

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Computer Science

by

Abbas Roayaei Ardakany

December 2019

Dissertation Committee:

Dr. Stefano Lonardi, Chairperson
Dr. Tao Jiang
Dr. Tamar Shinar
Dr. Eamonn Keogh

The Dissertation of Abbas Roayaei Ardakany is approved:

Committee Chairperson

University of California, Riverside

Acknowledgments

I am grateful to my advisor, Prof. Stefano Lonardi without whose help, I would not have been here. I would like to thank my fellow labmates Hamid, Hind, Rachid, Weihua, Abid, Qihua, Dipankar and Saleh for making five years of PhD working fun. I thank my friends Saeed, Ebi, Mostafa, Mojtaba, Mamad, Mamad-safar, Ali-mehr, Saeedeh, Rasoul, Alireza, Farshad-koochike, Farshid, Ebi-Amiri, Seyed, Hamed, Hossein, Shabnam, Joobin, Mehdi, Amir-fegh, Mahdi Ghamkhari, Amin Ghiasi, Reza, Asieh, Iman, Mamad-Ranjbar and especially Mina, without whom I would not enjoy the PhD.

To my family for all the support.

ABSTRACT OF THE DISSERTATION

Efficient Algorithms for the Analysis of Hi-C Contact Maps

by

Abbas Roayaei Ardakany

Doctor of Philosophy, Graduate Program in Computer Science
University of California, Riverside, December 2019
Dr. Stefano Lonardi, Chairperson

This dissertation deals with the analysis of high-throughput chromatin conformation capture (Hi-C) data. Hi-C experiments provide genome-wide maps of chromatin interactions and has enabled Life Scientists to investigate the role of the three-dimensional structure of genomes in gene regulation and other essential cellular functions. Several studies have confirmed the existence of fundamental 3D structural features of different scales that are stable across cell types and conserved across species, e.g., topological associating domains (TADs) and chromatin loops.

The research presented here is articulated around three main topics on the analysis of contact maps, namely (1) the detection of TADs, (2) how to compare two maps, and (3) how to detect chromatin loops. The detection of TADs has become a critical step in the analysis of Hi-C data, e.g., to identify enhancer-promoter associations. First, we present EAST, a novel TAD identification algorithm based on fast 2D convolution of Haar-like features, that is as accurate as the state-of-the-art method based on the directionality index, but 75-80× faster.

Another fundamental problem in the analysis of Hi-C data is to compare two contact maps derived from Hi-C experiments to identify the functional differences. Detecting similarities and differences between contact maps is critical in evaluating the reproducibility of replicate experiments and identifying differential genomic regions with biological significance. Due to the complexity of chromatin conformations and the presence of technology-driven and sequence-specific biases, the comparative analysis of Hi-C data is analytically and computationally challenging. Second, we present a novel approach called Selfish for the comparative analysis of Hi-C data that takes advantage of the structural self-similarity in contact maps. We define a self-similarity measure to design algorithms for (i) measuring reproducibility for Hi-C replicate experiments and (ii) finding differential chromatin interactions between two contact maps. Extensive experimental results on simulated and real data show that Selfish is more accurate and robust than state-of-the-art methods.

Regulatory elements at large genomic distances can engage in gene regulation by making direct physical contacts to their target genes or loci bringing distant loci in close spatial proximity of each other forming chromatin loops. These long-range interactions form complex regulatory networks that need to be carefully studied. Analyzing chromatin interactions between regulatory elements and genes at high resolution using high-throughput chromosome conformation capture method Hi-C, can provide fundamental insights into the spatial organization of chromosomes and its effect on gene regulation. Third, we present a new method Mustache to detect significant chromatin interactions genome-wide. Mustache robustly finds chromatin pairs of loci that interacts significantly compared with the expected interaction. We show that detected interactions are biologically supported by running a wide

range of experiments. The experiments indicate that these interactions are associated with contacts between promoters and enhancers, promoters to promoters, mediated by different proteins and are stable between cell types.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This dissertation describes research on the analysis of High-throughput Chromosome Conformation Capture (Hi-C) data. Specifically, it introduces efficient and accurate algorithms for the detection of local genome structures at different scales and the comparative analysis of Hi-C data.

Recent studies have revealed that genomic DNA in eukaryotes is not arbitrarily packed into the nucleus. The chromatin has a well organized and regulated structure in accordance to the stage of the cell cycle and environmental conditions [43, 48]. The chromatin structure in the nucleus plays a critical role in many essential cellular processes, including regulation of gene expression and DNA replication [17, 16, 50, 54]

Therefore, it is of great importance to systematically study the 3D conformation of the chromatin and how its folding is associated to specific cell function and the transcriptional control of genes. Historically, the spatial organization of the genome had been studied by conducting fluorescent in situ hybridization (FISH), which are low-throughput, time con-

suming experiments. As a consequence these study were limited to a few genomic loci, in a few dozens cells at once, and limited in spatial resolution. Newer and more advanced super-resolution microscopy approaches such as STORM and PALM have enabled Life Scientists to look into more detailed structures of chromatin at an much higher resolution [4].

Despite these remarkable technical advancements, microscopy-based approaches are limited in different aspects. First, they can be carried out only for a limited number of loci, thus they do not allow a genome-wide analysis of chromatin structure. Second, observed folding patterns cannot be directly translated to genomic loci, which substantially limits its integration with other genomic data [31]. Proximity ligation techniques, which are variations of the chromosome conformation capture (3C) experiment [14], were introduced to experimentally quantify chromatin interactions between different genomic loci in the genome [53]. 3C works by randomly cross-linking genomic loci in close physical proximity in the nucleus.

3C was followed by many derivative techniques (4C, 5C, ChIA-PET, Hi-C), which all start with a similar set of steps, even though they differ by the way the ligation product is measured and quantified. Chromosome conformation capture-on-chip (4C) identifies the interaction of unknown DNA regions with a locus of interest [58]. Chromosome conformation capture carbon copy (5C) detects all interactions in a given region. It generates a library of any ligation products which are then analyzed by next-generation sequencing for a given region of interest (typically no greater than 1Mb) [19].

Chromatin Interaction Analysis by Paired-End Tag Sequencing (ChIA-PET) combines chromatin immunoprecipitation (ChIP) with 3C to determine long-range genome-wide chromatin interactions [25].

Hi-C is considered the least biased method for discovering genome-wide chromatin interactions [36]. In Hi-C, after the initial restriction enzyme digestion as it is done in 3C, (1) the ends of fragments are filled by biotinylated nucleotides, (2) the blunt-end fragments are ligated, (3) the ligated DNA is sheared and (4) the marked fragments (chimeric products) are enriched by a streptavidin pull-down step. The resulting DNA (chimeric) fragments are then paired-end sequenced to find the corresponding interacting loci of each fragment (Figure 1.1).

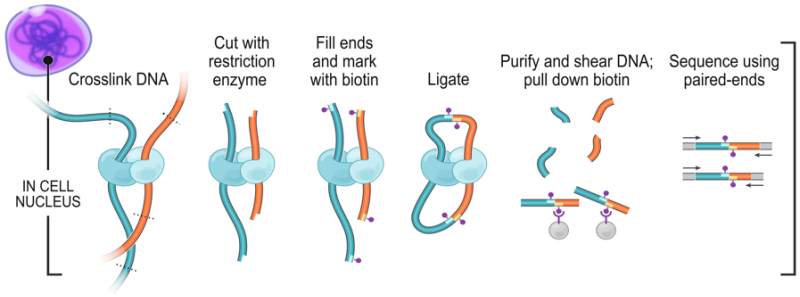


Figure 1.1: Overview of Hi-C experiments. Cells are first cross-linked using formaldehyde. This results in segments of DNA which are in close spatial proximity linked together. Then the chromatin is digested by a restriction enzyme and the ends are biotinylated. Then the blunt-end fragments are ligated, the ligated DNA is sheared and the marked fragments (chimeric products) are enriched by a streptavidin pull-down step. The resulting DNA (chimeric) fragments are then undergo a paired-end sequencing to find the corresponding interacting loci of each fragment (source [36]).

Each paired-end read represents a potential interaction between its two corresponding loci. The frequency of interactions between two fragments of DNA can be translated to the proximity of those fragments in 3D space. In the last step, interaction frequencies are

binned into equal-sized regions forming contact maps. In a Hi-C *contact map* (or *interaction matrix*) A, each entry $A(i, j)$ represents the number of times segments i and j are observed together in a DNA proximity ligation experiment. Larger values of $A(i, j)$ indicate closer loci i and j in 3D space inside the nucleus. The average size of each bin is directly proportional to the sequencing depth of the experiment. Higher coverage of sequencing provides more accuracy in detecting longer chromatin interactions, less noise, and higher resolution. The first Hi-C study provided chromatin interactions at a 1Mb resolution [36], which was improved to 40kb [17] and later to 1kb resolution [50].

The genome structure is thought to be organized hierarchically. Each chromosome occupies its own "territory": intra-chromosomal interactions occur significantly more frequent than inter-chromosomal interactions. Inside each chromosome territory, the organization of genomic regions is not random and is associated to the transcriptional activity of the cell. Transcriptionally active regions interact with other active regions more frequently when compared to inactive regions which tend to interact to other inactive regions [67]. Active regions, termed as *compartment A*, contain higher levels of gene density, chromatin accessibility and active histone modifications while inactive regions, termed as *compartment B* tend to be gene depleted and contain heterochromatin.

With the decreasing cost of Hi-C experiments and the higher availability of Hi-C data for different cell types in diverse conditions, there is a growing need for reliable and robust measures to systematically compare contact maps to discover similarities and differences.

The analysis Hi-C data presents computational and analytical challenges which are due to technology-driven and sequence-specific biases. Technology-driven biases include non-uniform sequencing depth, cross-linking conditions, circularization issues, and restriction enzyme biases [46, 11, 59]. Sequence-specific biases include GC content of trimmed ligation junctions, sequence uniqueness, and nucleotide composition [64].

For instance, it is well-known that contact maps produced as a result of replicate experiments can contain significant difference solely due to these biases, which could be falsely interpreted as biological differences if these biases are not accounted for [66]. Several normalization methods have been developed to compensate for these biases and improve the reproducibility of Hi-C experiments, e.g., [36, 64, 34, 32]. While several computational methods have been proposed to extract statistically significant contacts from normalized contact maps [50, 1, 6, 52], their performance is still not entirely satisfactory due to the inherent complexity, inter-dependency and unaccounted biases in chromatin interaction data.

In this dissertation, we developed a set robust and efficient methods to analyze Hi-C data. By conducting an extensive set of experiments we showed that our methods outperform the state-of-the-art methods for the analysis of Hi-C data both on simulated and real Hi-C datasets. In Chapter 2, we present a novel TAD calling algorithm called EAST (for "efficient and accurate summed-area-table-based TAD calling") that takes advantage of fast 2D convolution. Experimental results show that EAST is as accurate in detecting TADs as the DI method [17], which is considered the state-of-the art. EAST is however, 75-80$\times$ faster than DI.

In Chapter 3, we present a new comparative method SELFISH (for "discovery of differential Chromatin Interactions via a Self-Similarity measure") for the analysis of Hi-C data based on the notion of self-similarity [55]. We show that our self-similarity measure is robust to biases and does not need complex and computationally intensive normalization steps, such as MA [20] or MD [59].

In Chapter 4, we introduce MUSTACHE (for "multi-scale detection of statistically significant interactions") for genome-wide detection of significant chromatin interactions (loops). Mustache uses a scale-space representation of the contact map to model chromatin loops produced by interacting of chromatin segments with different sizes. We show that MUSTACHE detects the majority of chromatin loops detected by HICCUPS [50], as well as new biologically significant loops which HICCUPS fails to capture.

Finally, Chapter 5 concludes the dissertation by summarizing the main findings of the research and discusses possible directions for extending this work.

Chapter 2

Calling Topologically Associating Domains

2.1 Introduction

Topological associating domains (TADs) are large, megabase-sized contiguous local chromatin interaction domains that have a high average interaction within and a low average interaction with their surrounding regions. Because of the role that TADs play in cellular functions they have been widely explored since their discovery [69]. TADs are stable across different cell types and highly conserved across species [17]. TADs tend to interact with each other in a tree-like structure and form a hierarchy of domains-within-domains (metaTAD), which can scale up to the size of chromosomes. metaTADs show correlation with genetic and epigenomic features. TAD boundaries are enriched for the insulator binding protein CTCF, housekeeping genes, transfer RNAs and histone modifications [17, 23]. More importantly,

enhancers tend to interact with gene promoters within the same TAD [33]. Disruption of TAD boundaries can affect the expression of nearby genes and lead to developmental disorders or cancer [42].

Several methods have been developed to identify TADs genome-wide [24]. Dixon *et al.* were the first group to identify and define TADs [17]. In their seminal work, they proposed an identification method based on the *directionality index* (DI) which measures the frequency of interaction of a genomic locus with a fixed-sized neighborhood. Drastic changes of the DI score are expected at TAD boundaries where the region tends to have a high rate of both upstream and downstream interactions.

Filippova *et al.* [23] introduced a single parameter, two-step dynamic programming method for detection of TADs. Assuming that there exist a few characteristic resolutions across which TADs are similar, they identify a set of non-overlapping domains that are persistent across the resolutions.

Crane *et al.* [12] proposed a method based on the *insulation index* (IS). For each chromosome segment, IS score is the average number of interactions that cross the segment in a pre-specified size neighborhood. Given that interactions tend to be isolated within TADs, IS local minima are expected to occur at TAD boundaries. The IS score can be computed efficiently by sliding a window across the diagonal of the contact matrix and computing the average number of interactions that fall inside the window.

Chen *et al.* [10] translated the TAD identification problem into a graph segmentation/clustering problem. In this method, domains at different scales are identified by running the spectral graph cuts algorithm recursively until the connectivity of the graph reaches some predefined threshold.

In a Hi-C contact map, segments that are close in genomic 1D distance tend to form dense areas which can be seen as isolated high frequency blocks along the matrix diagonal, namely, TADs. TADs have high intra-frequency within and low inter-frequency with their neighboring blocks. The aim is to identify TADs efficiently and accurately.

We propose an algorithm called EAST that utilizes rectangular Haar-like features [62] and dynamic programming to identify TADs. Genomic regions are scored based on an objective function that measures their likelihood of containing a TAD with respect to the characteristics mentioned above. We use Haar-like features to describe such a scoring function.

2.2 Summed area table and Haar-like features

A *Haar-like feature* is a set of adjacent rectangular regions each of which has a certain weight. Weights of rectangular regions indicate certain characteristics of a particular area of the image. By convolving Haar-like features, i.e., by computing the weighted sum of pixel values for a particular location, we obtain a value that represents how well a region (window) satisfies the characteristics we are looking for. To compute the weighted sum efficiently we use the summed area table.

A *summed area table* (SAT), also known as *integral image* in computer vision, is a data structure used for efficiently calculating the sum of values in a rectangular region. By precomputing the summed area table one can obtain the sum of values in any arbitrary rectangular region using only a constant number of operations. SAT was first introduced to computer graphics in 1984 by Frank Crow [13] and later to computer vision in 2001 by Lewis [62] in a popular face detection framework called Viola-Jones. The value of a point (x, y) in a summed area table A_{SAT} is the sum of all pixels above and to the left of that point in the original grid A, including the (x, y) point itself.

$$A_{\text{SAT}}(x, y) = \sum_{x' \leq x,\ y' \leq y} A(x', y')$$

Since the value of each point in the SAT can be computed based on the values of neighboring points, the formula can be rewritten as

$$A_{\text{SAT}}(x, y) = A(x, y) + A_{\text{SAT}}(x - 1, y) + A_{\text{SAT}}(x, y - 1) - A_{\text{SAT}}(x - 1, y - 1)$$

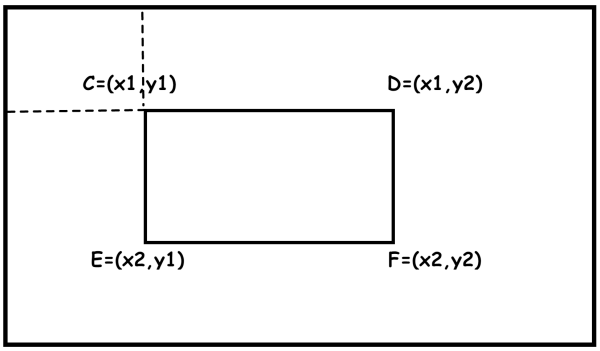
Given the summed area table, computing the sum of values in an arbitrary size rectangular region can be done in $O(1)$ time (see Figure 2.1).

2.3 TAD objective function

To score TADs we need to define a function f that quantifies the quality of an arbitrary region along the matrix diagonal with respect to the following properties

1. The average frequency inside the region must be "high"

2. The average frequency with the neighborhood must be "low"



$$\sum_{\substack{x_1 \le x \le x_2 \\ y_1 \le y \le y_2}} A(x,y) = A_{\text{SAT}}(C) + A_{\text{SAT}}(F) - A_{\text{SAT}}(D) - A_{\text{SAT}}(E)$$

Figure 2.1: If the summed area table A_{SAT} is available, computing the sum of values in any rectangular region takes $O(1)$ time.

3. The average frequency between start and end segments of the region must be higher than the average frequency inside the region

The last property derives from the fact that TADs are the result of a compact locality or loop formation in the chromatin. To explain the design of the objective function f we refer to Figure 2.2, where different colors indicates different weighting. The area in green color is the region we expect to have a high frequency of interaction (intra-frequency), as opposed to the area in red where lower frequency is expected (inter-frequency). The corner region which is colored in blue in Figure 2.2 has a higher weight in order to account for the last property in the list above. Using the SAT data structure, function f can be computed as follows.

$$f([i,j]) = \frac{CDEF^{\diamond} - \alpha \cdot (ABGH^{\diamond} - CDEF^{\diamond}) + \beta \cdot IDJK^{\diamond}}{\mathcal{N}}$$

where $CDEF^{\diamond}$, $ABGH^{\diamond}$ and $IDJK^{\diamond}$ represent the sum of pixel values inside the rectangular regions $CDEF$ (defined by interval $[i,j]$), $ABGH$ and $IDJK$ respectively, which can

be computed in $O(1)$ time from the SAT of the interaction matrix A. Parameters α and β are dataset-independent, and they can be determined experimentally. Parameter $\mathcal{N}$ is a normalization factor discussed in subsection 2.5.1.

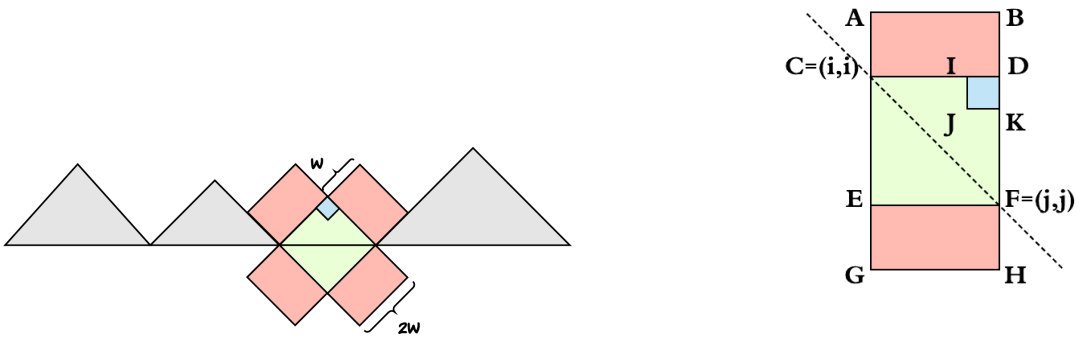


Figure 2.2: Objective function f. (LEFT) Representation of a TAD of size $2w$. High interaction frequency expected inside the TAD's domain (green) while low interaction frequency is expected between the TAD and surrounding domains (red) (RIGHT) Coordinates of Haar-like representation of a TAD.

2.4 Finding the optimal set of domains

Given a $n \times n$ interaction matrix A, the problem of TAD identification is an optimization problem aimed at identifying the set of contiguous non-overlapping domains for which the

$$\sum_{d_i \in D} f(d_i)$$

is maximized, where $D = \{d_i | d_i = [s_i, e_i]\}$ is a set of non-overlapping intervals, i.e., $e_j < s_i$ or $e_i < s_j$ for $i \neq j$.

We use dynamic programming to solve this optimization problem. The optimal solution $OPT(i)$ for the sub-problem $[1, i]$ can be expressed by following recurrence relation

$$OPT(i) = \max_{0 \leq k \leq i-1} \{OPT(k) + f([k+1, i])\}$$

By gradually increasing the size of the sub-problem and keeping track of the set of extracted domains, the optimal set of TADs for the entire interaction matrix can be computed. As we grow the size of the sub-problem, for each bin i, we need to find the optimal location to break the sub-problem $[1, i]$ into a sub-problem $[1, k]$ and a domain $d = [k+1, i]$. The overall time-complexity is $O(n^2)$, where n is the number of bins/segments.

If we do not allow TADs to be larger than L, the optimal break point for a sub-problem $[1, i]$ can always be found in the interval $[max\{i - L, 0\}, i - 1]$. Therefore, the overall time complexity decreases to $O(nL)$.

Theorem 1 () *Let $D^* = \{[a_1, a_2], [a_2, a_3], \ldots, [a_{s-1}, a_s]\}$ be an optimal set of domains for the interaction matrix A for which*

$$\sum_{d_i \in D^*} f(d_i)$$

is maximized. Then,

$$OPT^*(n) = OPT(n)$$

where

$$OPT^*(i) = \max_{max\{i-L,0\} \leq k \leq i-1} \{OPT^*(k) + f([k+1, i])\}$$

Proof. *We prove the theorem by induction. For the base case* $OPT^*(a_1) = OPT(a_1) = 0$. *Now, suppose* $OPT^*(a_{i-1}) = OPT(a_{i-1})$ *then we have*

$$\begin{aligned} OPT^*(a_i) &= OPT^*(a_{i-1}) + f([a_{i-1}+1, a_i]) \\ &= OPT(a_{i-1}) + f([a_{i-1}+1, a_i]) \\ &= OPT(a_i) \text{ for } k = a_{i-1} \end{aligned}$$

where k *satisfies the inequality* $\max\{a_i - L, 0\} \leq k \leq a_i - 1$. ■

2.5 Experimental results

We performed the analysis on Hi-C data for two mouse cell types (cortex and embryonic stem cell), and one human cell type (embryonic stem cell) at bin resolution of 40kb. The Hi-C data was obtained from [17].

2.5.1 Parameter settings

In addition to α, β and L, EAST relies on two additional parameters. The first is the minimum quality threshold τ that is used to filter out low-quality TADs. If we assume that TAD quality scores are distributed according to a Gaussian distribution, we define the threshold $\tau = \mu - \sigma$ where μ and σ are the mean and standard deviation of the distribution of scores. Observe that parameter τ can be computed from the analysis of the dataset.

The second parameter is the normalization parameter $\mathcal{N}$ for the function f. Since the quality measure f is proportional to the sum of interactions inside the domains, f grows as the TAD size increases. Figure 2.3 illustrates how the sum of interactions inside a domain grows as the size of the TAD increases for the three datasets used in the experimental

results below and for a synthetic interaction matrix. In the synthetic data, the number of interactions was set to be inversely proportional to the genomic distance. For purpose of comparison, the sum of interactions is normalized by the sum of the largest domain.

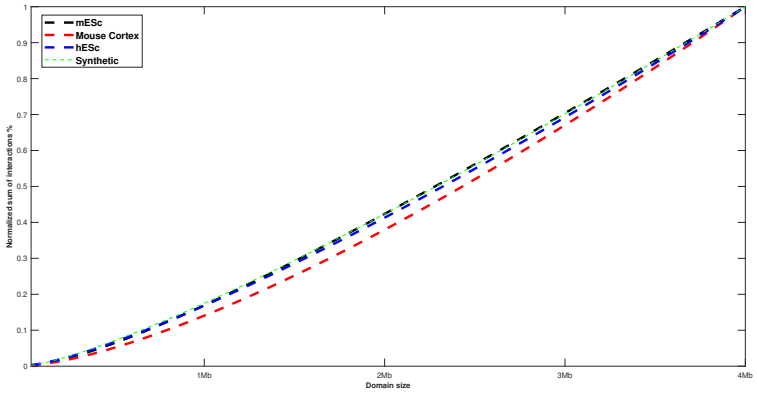


Figure 2.3: Growth of the quality measure f as the size of the TAD increases on the three datasets used in the experimental results and for a synthetic interaction matrix (see text).

Observe that the curve for the mouse embryonic data roughly matches the curve for the synthetic data. This suggests that the average interaction frequency of two loci in the mESC dataset is inversely proportional to their genomic distance. The growth function of the synthetic data can be estimated by $(n/L)^{1.2}$ where L is the largest domain size we are evaluating.

Also observe the hESC and mouse cortex curves are slightly different from the curve for the synthetic data, and they can be estimated by $(n/L)^{1.36}$ and $(n/L)^{1.4}$ respectively. We experimentally determined that as the curves diverge from the curve for the synthetic data, the normalization factor needs to adjusted accordingly. We set $\mathcal{N} = n^{0.4}$, $\mathcal{N} = n^{0.43}$ and $\mathcal{N} = n^{0.38}$ for hESC, mESC and mouse cortex, respectively. Parameters α and β were optimized experimentally to values $\alpha = 0.2$ and $\beta = 0.2$, and they are dataset-independent.

2.5.2 Comparison with existing methods

Based on the availability and popularity of TAD calling methods, we decided to compare EAST with the directionality index method [17], insulation score method [12] and multiscale method in [23]. We hereafter refer to these methods as DI, INS and MR respectively.

EAST, DI, INS and MR were ran on an Intel Core-i7 2.7GHz CPU with 16GB of memory. For the DI method we ran the experiments with posterior marginal probability threshold 0.99 and up/downstream span size of 2Mb (default parameters according to [17]). For the INS method, we set the insulation delta span to 200kb and the insulation square size to 500kb. For the MS method, we set the highest resolution parameter to 0.5.

In our experiment we investigated the enrichment of epigenetic characteristics of chromatin near the TAD boundaries. Although the mechanism behind the formation of TADs and their role in gene regulation are not fully understood, multiple studies have shown that some proteins and histone marks are enriched at the TAD boundary regions, implying that these boundaries play a role in gene transcription. As it was done in other studies [23, 57, 10], we can therefore use these genomic markers to evaluate the quality of the computed TADs.

To produce enrichment plots, we used each method to determine the boundary locations of TADs. Then, the frequency of each marker was calculated in 10kb bins in a window of 1Mb centered at the TAD boundaries. Each plot show the distribution of specific markers for each tool in the region centered at the TAD boundaries.

For mouse cortex and stem cells we evaluated the enrichment of transcription factor CTCF, promoter related marks RNA Polymerase II and H3K4me3, and enhancer-related histone modification H3K27ac. This marker data was collected from [56]. For human stem cells we assessed the enrichment of CTCF near TAD boundaries. The CTCF data was obtained from [35].

Figure 2.4 shows that CTCF binding sites are almost twice as enriched near the TAD boundaries than the surrounding regions, suggesting that TAD boundaries are associated with insulator genomic regions and their mediator protein CTCF. Figure 2.5 and Figure 2.6 show that promoter marks RNA Polymerase II and H3K4me3 peak within the TAD boundaries for both mouse cortex and embryonic stem cells. Observe in Figure 2.7 that histone modification mark H3K27ac is highly enriched around TAD boundaries in mouse embryonic stem cells but not in mouse cortex cells. Also observe in Figure 2.8 that enhancer marks are highly depleted around TAD boundaries in mouse cortex cells but not in mouse embryonic stem cells.

Overall, observe in Figures 2.4-2.8 that the blue curve for EAST is almost always higher than the other three tools, suggesting that our tool generates TADs with very accurate boundaries. The closest competitor is DI (green curve), but EAST is significantly faster than DI.

We compared the running time of EAST with that of DI, MS and INS on Hi-C data for human embryonic stem cells, mouse embryonic stem cells and mouse cortex [17].

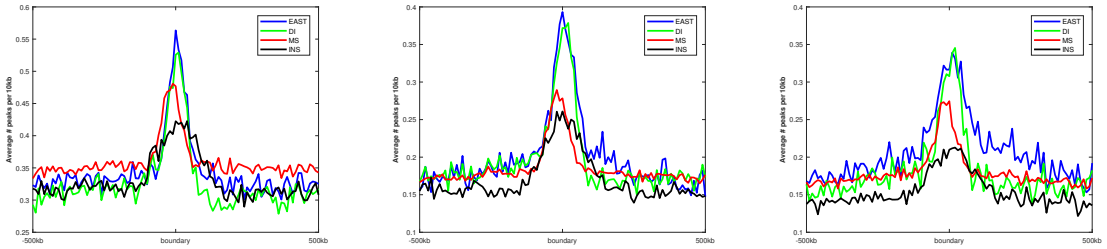


Figure 2.4: CTCF enrichment in human embryonic stem cells, (left) mouse embryonic stem cells (center) and mouse cortex cells (right).

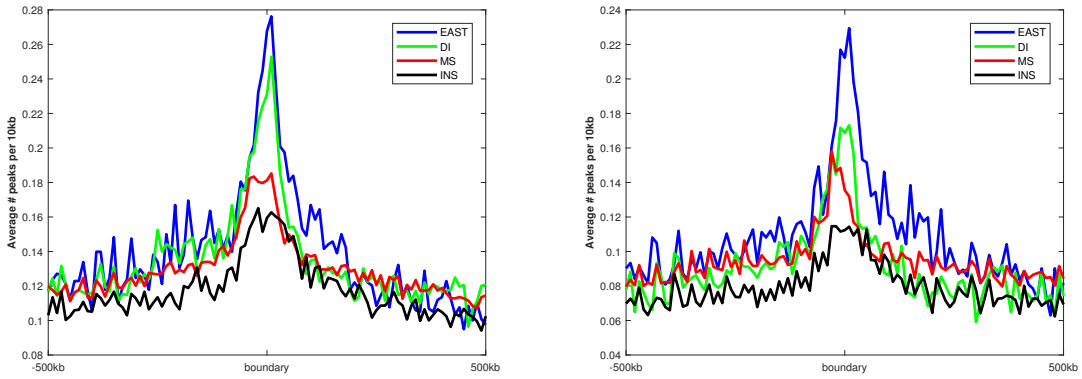


Figure 2.5: H3K4me3 enrichment in mouse embryonic stem cells (left) and mouse cortex cells (right).

Table 3.1 shows that EAST and INS are comparable in speed, MS is 10-14× slower, DI is 75-80× slower.

Figure 2.9 illustrates the size distribution of TADs for all four methods for the human embryonic stem cells. The numbers of TADs extracted by EAST, DI, MS and INS are 2229, 2429, 12427 and 4708 respectively. Observe that EAST and DI roughly produce the same size distribution.

In summary, these experimental results show that while EAST can identify the TAD boundaries as accurately as the best method (DI), but it is much more time efficient.

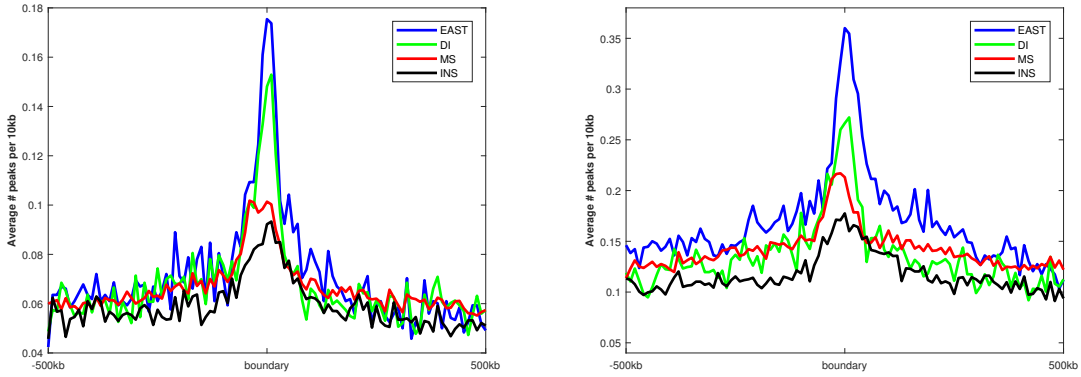


Figure 2.6: polII enrichment in mouse embryonic stem cells (left) and mouse cortex cells (right).

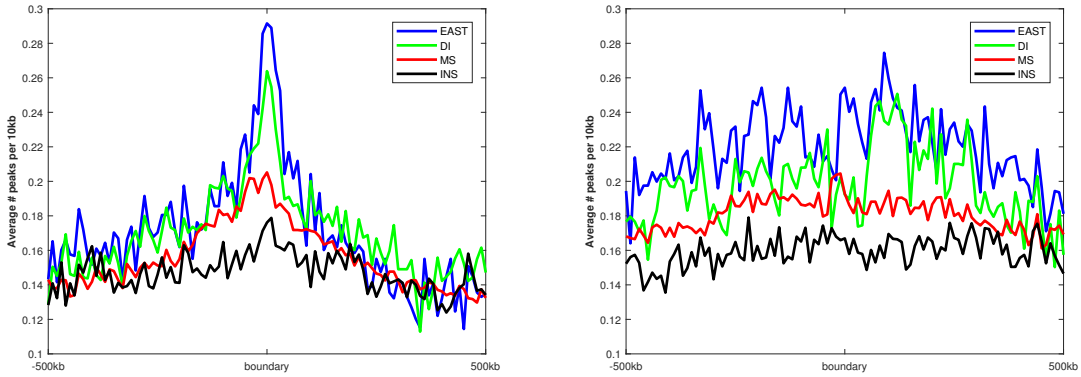


Figure 2.7: H3K27ac enrichment in mouse embryonic stem cells (left) and mouse cortex cells (right).

2.6 Conclusion

In this chapter, we introduced an efficient algorithm called EAST, to accurately identify topological associating domains in chromatin from interaction matrices obtained from high-throughput chromosome conformation capture (Hi-C). EAST can be downloaded from `https://github.com/ucrbioinfo/EAST`.

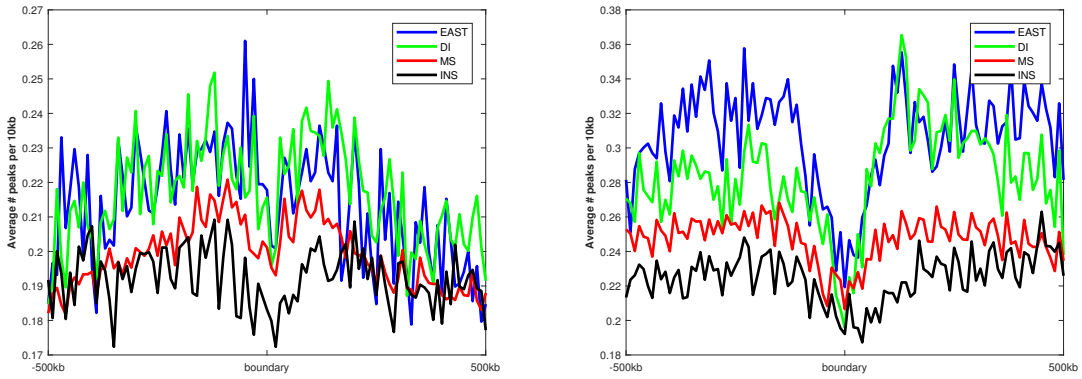


Figure 2.8: Enhancer enrichment in mouse embryonic stem cells (left) and mouse cortex cells (right).

	hESC	mESC	Cortex
EAST	58s	50s	48s
INS	52s	44s	42s
DI	4,721s	3,845s	3,628s
MS	762s	545s	520s

Table 2.1: Running time of EAST, INS, MS and DI on the three datasets used in this work.

We performed a comparative evaluation of EAST on Hi-C data for human stem cells, mouse stem cells and mouse cortex cells. We showed that our algorithm extracts TADs as accurately as the state-of-the art. TADs identified by EAST show substantial enrichment of various epigenetic modification factors at their boundaries, confirming similar findings in previous studies. By comparing the running time of EAST with the other published methods, we showed that our method is very time efficient. For a given Hi-C dataset, the only parameter in EAST that might need to be tuned by the user is the normalization factor for which we have given some guidance in Subsection 2.5.1.

The framework we presented here for TAD identification is based on fast 2D-convolution of Haar-like features. We believe that this framework could be adapted to

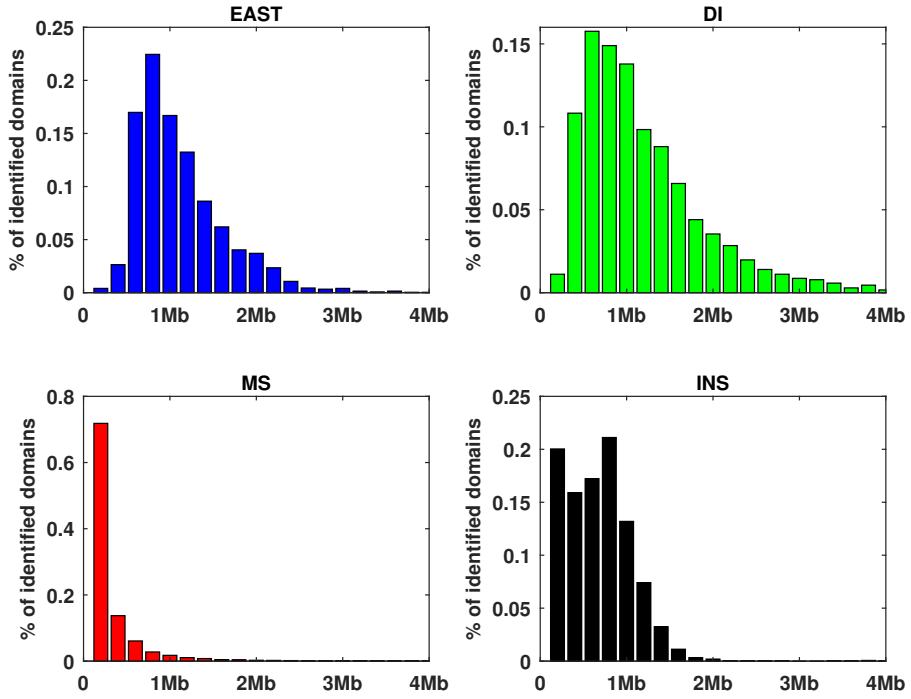


Figure 2.9: Comparison of the distribution of TAD size.

other chromatin feature detection problems such as chromatin loops [50]. We also plan to extend our work to efficiently identify chromatin features at arbitrary scales.

Chapter 3

Discovery of Differential Chromatin Interactions via a Self-Similarity Measure

3.1 Introduction

Recent studies have revealed that genomic DNA in eukaryotes is not arbitrarily packed into the nucleus. The chromatin has a well organized and regulated structure in accordance to the stage of the cell cycle and environmental conditions [43, 48]. The chromatin structure in the nucleus plays a critical role in many essential cellular processes, including regulation of gene expression and DNA replication [17, 16, 50, 54].

Technological and scientific advancements in genome-wide DNA proximity ligation (Hi-C) have enabled life scientists to study how chromatin folding regulates cellular

functions [36, 8, 9, 27]. The analysis of Hi-C led to the discovery of new structural features of chromosomes such as topologically associating domains (see chapter 2) [17, 69] and chromatin loops [50, 7].

With the decreasing cost of Hi-C experiments and the higher availability of Hi-C data for different cell types in diverse conditions, there is a growing need for reliable and robust measures to systematically compare contact maps to discover similarities and differences. However, the comparative analysis of Hi-C data presents computational and analytical challenges due to presence of technology-driven and sequence-specific biases. Technology-driven biases include sequencing depth, cross-linking conditions, circularization length, and restriction enzyme sites length [46, 11, 59]. Sequence-specific biases include GC content of trimmed ligation junctions, sequence uniqueness, and nucleotide composition [64]. For instance, it is well-known that contact maps from replicate experiments can contain significant differences solely due to these biases, which could be falsely interpreted as biological differences if these biases were not accounted for [66]. Several normalization methods have been developed to compensate for these biases and improve the reproducibility of Hi-C experiments, e.g., [36, 64, 34, 32]. While several computational methods have been proposed to extract statistically significant differences in normalized contact maps [50, 1, 6, 52], their performance is still not entirely satisfactory due to the inherent complexity, inter-dependency and unaccounted biases in chromatin interaction data.

There are two major domains of application for the comparative analysis of Hi-C contact maps. The first application domain is focused on quantifying the reproducibility of Hi-C biological/technical replicate experiments [66]. For instance, [65] defined a repro-

ducibility measure based on the *stratum-adjusted correlation coefficient statistic* defined on the unique spatial features of Hi-C data. Their method HiCRep (i) reduces the effect of noise and biases by applying a 2D averaging filter on the data, (ii) addresses the distance-dependence of Hi-C data by stratifying the data with respect to the genomic distance, (iii) calculates a Pearson correlation coefficient for each stratum, and (iv) aggregates the computed stratum-specific correlation coefficients using a weighted average.

In the same application domain aimed at quantifying reproducibility or concordance of contact maps, [61] presented a method called GenomeDISCO to measure the differences between smoothed contact maps. GenomeDISCO represents contact map as a graph, where each node represents a genomic locus and each edge represents an interaction between two loci. Edges are weighted by the normalized frequency of the corresponding pairs of loci. GenomeDisco executes iteratively the two following steps: (i) traverses the graph using random walks, which has the effect of denoising (smoothing) the data, (ii) computes the normalized difference between smoothed contact maps using the L_1 distance between two contact maps.

The second application domain is aimed at finding statistically significant differences between contact maps for cells in different states (tissues, developmental states, healthy/diseased, time-points, etc). It is well-known that chromatin interactions that are mediated by specific protein can have distinct frequencies in different cell types or in different cell conditions [47, 26]. Differences in chromatin interactions can be associated with cell type-specific gene expression or mis-regulation of oncogenes or anti-oncogenes [39, 53, 30].

[63] proposed the first method to discover differences in Hi-C contact maps. The authors used a simple fold-change of the normalized local interactions to discover that estrogen stimulation significantly impacts chromatin interactions in MCF7 cells. Building on this idea, [16] proposed a method that (i) quantile-normalizes contact maps to compensate for the bias induced by different sequencing depth, and (ii) determines the significance of normalized differences between two contact maps (augmented by feature vectors representing epigenetic signals) using a Random Forest model. Their method can (i) determine whether the epigenetic signal is predictive of changes in interaction frequency and (ii) discover which epigenetic signals are most predictive of changes in higher-order chromatin structure.

[59] developed a non-parametric method to account for between-datasets biases. They used locally weighted polynomial regression to fit a simple model trained on the difference between the two datasets. Based on the assumption that the majority of the interactions should be relatively unchanged among similar Hi-C datasets and by centering the average difference to zero, loci which are far from the average are considered potentially significant differential interaction.

Unlike other methods which assume independence among pairwise interactions (which holds true only for low resolution Hi-C) [18] presented a method that takes into account the dependency of adjacent loci in higher resolutions. Based on the fact that interacting neighboring loci are known to be inter-dependent, structural differences can be detected by observing the differences in a neighborhood of the corresponding loci pair. In contrast, random noises tend to affect singular pairwise interactions only. By considering a three-dimensional space in which the x and y are the coordinates of the genomic loci and z

is their pairwise interaction frequency, the authors define a chromatin interaction between two conditions to be *differential* when the intensity of the majority of k-nearest neighbors of (x, y) exhibit a significant change.

In this work, we address three major weaknesses of these existing methods for the comparative analysis of contact maps, namely, a) ignoring the inter-dependency of chromatin interactions, b) requiring a pre-processing (normalization) step based on a flawed assumption that biases between two contact maps can be accurately modeled and c) being extremely computationally demanding for the analysis of high-resolution Hi-C data. We present new comparative methods for the analysis of Hi-C data based on the notion of self-similarity [55]. We show that our self-similarity measure is robust to biases and does not need complex and computationally intensive normalization steps, such as Minus vs. Average (MA) [20] or Minus vs. Distance (MD) [59]. In the first part of the paper, we show that our self-similarity measure can be used as a tool to quantify the reproducibility of Hi-C biological/technical replicate experiments. In the second part, we show that our measure can also be employed for finding statistically significant differences between Hi-C contact maps. Although existing methods for comparing contact maps vary widely, they all share the following assumption. Given two contact maps that are expected to be similar (e.g., technical replicates of the same biological experiment) it is possible to devise (or train) a common underlying model can faithfully represent both. We believe that this approach is fundamentally flawed, because the inherent biases present in the Hi-C data are very hard to model and completely eliminate. Here we propose to use the intrinsic self-similarity structure in contact maps to avoid dealing with the modeling problem.

In the application domain of object detection in complex visual data, the notion of self-similarity was first introduced by [55]. The idea of self-similarity on images can be explained as follows. [55] showed that given two images of a certain object, the most relevant correlations between them are not necessarily the raw values of pixels (or an underlying model describing those pixel values) but the internal organization of self-similarities of local regions at similar relative geometric positions. Given two images of the same object, the relation between these local self-similarities tend to be more preserved than the similarities between the images.

3.2 Self-similarity and reproducibility

As said, existing methods for measuring the reproducibility of Hi-C experiments compute correlations or distances between normalized interaction frequencies of loci pairs, which is error-prone due to technology-driven and sequence-specific biases. Here we show that this comparison can be done indirectly by using self-similarity.

When we compare two contact maps that are expected to be similar, e.g. for two technical replicates of the same biological experiment, we expect to have similar internal layout of interactions. More precisely, given two contact maps A and B for two replicates, if we observe more chromatin interactions in block α than in block β in contact map A, we expect to have more chromatin interactions in block α than block β in contact map B as well, for several local choices of α, β. In other words, to measure similarity we do not need to depend on the absolute number of interactions in each contact map, rather we can rely on pairwise comparison between many local interactions. Here we claim that the

Boolean vectors representing binary comparison between local interactions encode enough information to define a similarity measure that can be used to quantify reproducibility for contact maps.

Henceforth, a Hi-C contact map is a $N \times N$ matrix where entry (i, j) in the matrix denotes the frequency of interaction between locus (or *bin*) i and locus (or *bin*) j in the genome. First, we slide a square block of size $N/k \times N/k$ along the main diagonal of the contact map using a stride of $N/2k$ so each pair of adjacent blocks overlap by half of their size. For each position of the sliding block, we compute the sum of interaction frequencies inside the block. We store these sums in vector B, which has $2k$ components. Then, we compare all $\binom{2k}{2}$ pairs of block sums, and set the matrix $C(s, t) = \mathbf{I}(B_s > B_t)$ for all choices of $(s, t) \in \{1, \ldots, 2k\} \times \{1, \ldots, 2k\}$, where $\mathbf{I}$ is the indicator function.

We claim that the matrix C is a compact representation of the interaction distribution along the main diagonal of the contact maps, which is robust to noise and biases (thus does not require normalization) because it relies on comparing entities that belong to the same contact map, and not across maps. We compute the similarity $\mathcal{S}(A, B)$ between contact map A and B as follows

$$\begin{aligned} \mathcal{S}(A, B) &= e^{-c\|C_A - C_B\|_2} \\ &= e^{-c\sqrt{\sum_{(s,t)\in\{1,\ldots,2k\}^2}[C_A(s,t) - C_B(s,t)]^2}} \end{aligned}$$

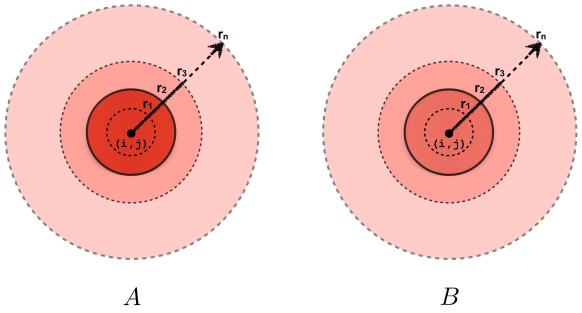
where c is a constant, C_A is the Boolean matrix for contact map A, C_B is is the Boolean matrix for contact map B. The value of k should be chosen so that the size of the resulting blocks N/k is sufficient large to enclose important chromatin structures (e.g., TADs). Parameters c and k are determined experimentally.

3.3 Self-similarity and differential chromatin interactions

For the accurate detection of differential chromatin interactions (DCI), we need to be able to distinguish true differences (which might have biological relevance) from differences caused by biases or other artifacts in the data. Since there is no ground truth for DCIs between cell types, conditions or developmental stages, there is no possibility of learning from real examples. The only differences that can be trusted are those that significantly exceed the differences observed between biological replicates. For this reason, we can also employ our self-similarity metric in a method for finding DCIs between two Hi-C contact maps. In our self-similarity representation described below, each interaction frequency is represented by a series of comparison between its surrounding local regions.

We first observe that DCIs have *locality* properties. If contact map A and B have a DCI at coordinate (i, j), this is not only reflected in the interaction difference of $A(i, j) - B(i, j)$ but also in the neighborhood of (i, j). We call *impact region* the neighborhood affected by the DCI. We call *impact radius* the size of the neighborhood being affected, which is proportional to the magnitude of the DCI. We argue that what determines the statistical significance of a DCI in a particular location (i, j) does not only depend on the statistical significance of the difference $A(i, j) - B(i, j)$ but also on the statistical significance of the difference between their region centered at (i, j). Isolated locations that have the large interaction frequency differences are often not significant and are likely to be due to noise or other artifacts.

To incorporate locality information in our self-similarity representation, each interaction $A(i, j)$ is represented by a linear combination of its neighboring interactions. To



A B

$$\Gamma(i,j) = (G_{r_1}(i,j), G_{r_2}(i,j), ..., G_{r_n}(i,j))$$

Figure 3.1: Our self-similarity metric for representing chromatin interactions is obtained by first convolving a contact map with a set of Gaussian filters with radii $\{r_1, r_2, \ldots, r_n\}$. The shade represents the intensity of the convolution for different radii. In this example, a sharp frequency change can be observed between radius r_2 and r_3 in contact map A but not in contact map B. This difference can indicate a potential DCI.

penalize interactions which are progressively farther from (i, j), we weight these local interactions via a Gaussian filter centered at (i, j) (see Figure 3.1 for an example). By gradually increasing the size of the Gaussian filter, we capture impact regions with larger and larger radii. We denote with $G^A_{r_k}$ the matrix resulting from the convolution between a Gaussian filter with radius r_k and the contact map A. We first compute $G^A_{r_k}$ for a set of n radii $\{r_1, r_2, \ldots, r_n\}$, and collect them in vector Γ_A as follows

$$\Gamma_A(i,j) = (G^A_{r_1}(i,j), G^A_{r_2}(i,j), \ldots, G^A_{r_n}).$$

It is well-known that interactions in Hi-C contact maps are more frequent when the pairs of interacting loci are closer in genomic distance due to random polymer interactions driven by one-dimensional genome proximity. To compensate for the amplification of contact frequency due to proximity, we Z-normalize the interaction frequencies in A with respect to their genomic distances along each diagonal d as follows.

$$\hat{A}(i,j) = \frac{A(i,j) - \mu_d}{\sigma_d}$$

where $d = |j - i|$, and μ_d, σ_d are the average and the standard deviation along the diagonal d, respectively.

If (i, j) is not a DCI between A and B, we expect vectors $\Gamma_{\hat{A}}(i,j)$ and $\Gamma_{\hat{B}}(i,j)$ to exhibit similar trends along their components, because they represents aggregate interaction frequency in gradually increasing neighborhood centered at (i, j). If (i, j) is a DCI with impact radius r, we expect to observe a significant difference between the k-th Gaussian representations of that interaction, where k is the index of the radius r_k closest to r. Due to biases in the interaction frequencies across different contact maps, the difference between the two feature vectors $\Gamma_{\hat{A}}$ and $\Gamma_{\hat{B}}$ cannot be directly used to indicate the significance of a change. We address this issue by take advantage of self-similarity, i.e., by using local comparison of local regions in contact maps.

According to [44], the Gaussian filter scale r must be distributed exponentially between the inner (r_1) and outer (r_n) scale limits (impact radii) $r_n = r_0 s^n$, in order to maintain a uniform change of information between successive levels of Gaussian filtering. For 5kb resolution data, we set $r_0 = 7$, $s = 2$ and $n = 10$.

Inspired by the work of [41], instead of using Γ to define the behavior of interaction frequency across a set of impact regions, we use the first order derivative of Γ with respect to impact radius r, which can be estimated by the difference $G_{r_{k+1}} - G_{r_k}$.

$$\frac{d\Gamma}{dr}(i,j,k) \approx \Delta\Gamma(i,j,k) = G_{r_{k+1}}(i,j) - G_{r_k}(i,j)$$

By computing the first order derivative for various choices of the impact radii, we carry out a comparison of local contact map regions ($G_{r_{k+1}} - G_{r_k}$). Figure 3.2 shows the first order derivatives of $\Gamma_{\hat{A}}$ and $\Gamma_{\hat{B}}$ and the difference between them for the DCI reported later in Figure 3.8. Observe the sharp change between the derivatives at radius r_2 which corresponds to a DCI with that radius.

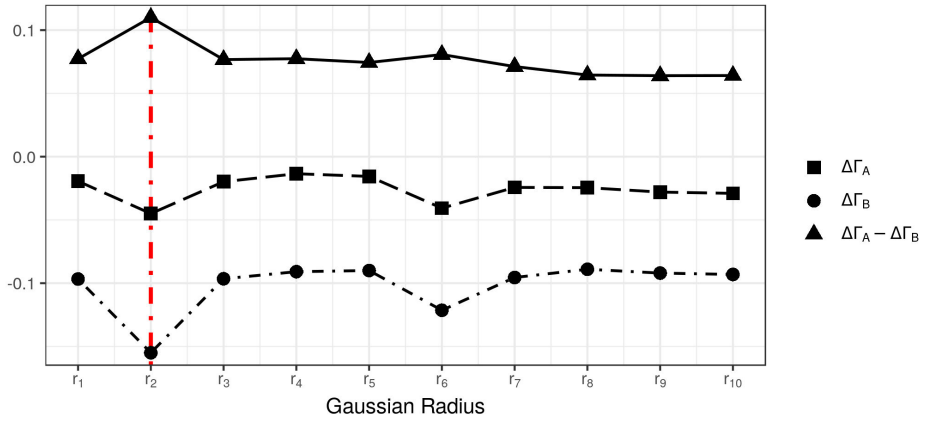


Figure 3.2: The first order derivatives $\Delta\Gamma_A$ and $\Delta\Gamma_B$ and the difference between them for the DCI reported later in Figure 3.8. A large difference between $\Delta\Gamma_A$ and $\Delta\Gamma_B$ at radius r_2 indicates the presence of a potential DCI.

In the last step of our algorithm, we compute the mean μ and standard deviation σ for the normal distribution fitted on the difference of the first order derivatives $\Delta\Gamma_A - \Delta\Gamma_B$ for each radius r_k. Then, we compute the p-value $P^k_{A,B}(i,j)$ for location (i,j) and radius r_k as follows

$$P^k_{A,B}(i,j) = \mathbf{Pr}\left(X > \left(\frac{d\Gamma_A}{dr}(i,j,k) - \frac{d\Gamma_B}{dr}(i,j,k)\right)\right)$$

where $X \approx N(\mu, \sigma)$.

From the set of k p-values for each index (i, j), we choose the smallest p-value of the difference between two contact maps A and B at that index, as follows.

$$P_{A,B}(i,j) = \min_{k \in \{1,\ldots,n\}} \{P^k_{A,B}(i,j)\}$$

These p-values $P_{A,B}(i, j)$ are finally fed into the Benjamini-Hochberg algorithm to calculate the final probabilities [2].

3.4 Results

3.4.1 Reproducibility

We evaluated our reproducibility measure on a Hi-C dataset obtained from [53] that has a variable total number of interactions and resolution. The dataset consists of five different cell types hESC (H1), Mesendoderm (MES), Mesenchymal Stem Cell (MSC), Neural Progenitor Cell (NPC) and Trophoblast-like Cell (TRO). Each cell type has two biological replicates. All experiments were carried out on a single chromosome (chromosome 1 for this work) with resolution 40 kb. According to our experience, parameter c is dependent on the particular Hi-C protocol used. Parameter k must be chosen such that the resulting blocks enclose the primary structures of contact maps which are likely to be preserved between cell types, e.g. TADs. Parameter c has to be set to the largest integer value such that the computed reproducibility for biological replicates is at least 0.9. For this dataset, we set $k = 100$ and $c = 5$. We compared our method SELFISH against two state-of-the-art reproducibility methods, namely HiCRep [65] and GenomeDISCO [61].

First, we assessed the effect of the total number of intra-chromosomal interactions captured by Hi-C experiment on different reproducibility measure. Given two biological replicates, we generated pseudo-replicates by first summing the two Hi-C matrices and then down-sampling the resulting matrix. Any pair of contact maps which are neither replicates or pseudo-replicates are called *non-replicates*. Next, each individual replicate was down-sampled to a wide range of total interactions $(10^5, 5 \times 10^5, 10^6, 2 \times 10^6, 5 \times 10^6, 10^7)$. For each of these choices, we computed the pair-wise reproducibility score.

Figure 3.3a-c illustrates the effect of the total number of interactions (which depends on the Hi-C sequencing depth) on the performance of reproducibility measures. A desirable feature for a reproducibility measure is to produce similarity scores that are invariant from the total number of interactions. Observe in Figure 3.3a-c that SELFISH is much more invariant to the total number of interactions than HiCRep and GenomeDISCO. Both these latter methods failed to report stable reproducibility scores which are independent from the sequencing depth.

In the next experiment, we evaluated the effect of binning resolution of Hi-C data on the reproducibility methods. For this experiment, we used deeply sequenced Hi-C data of cell type GM12878 from [50]. Again, a desirable property of a reproducibility score is to be robust to changes in resolution. Figure 3.3d shows that reproducibility scores for resolutions 5kb, 10kb, 25kb, 50kb and 500kb are very stable for SELFISH, whereas HiCRep and GenomeDISCO scores are resolution-dependent, in particular for GenomeDISCO.

We also tested our reproducibility measure to cluster different cell and tissue types. Contact maps of fourteen different tissues and two cell types were obtained from [53]. A

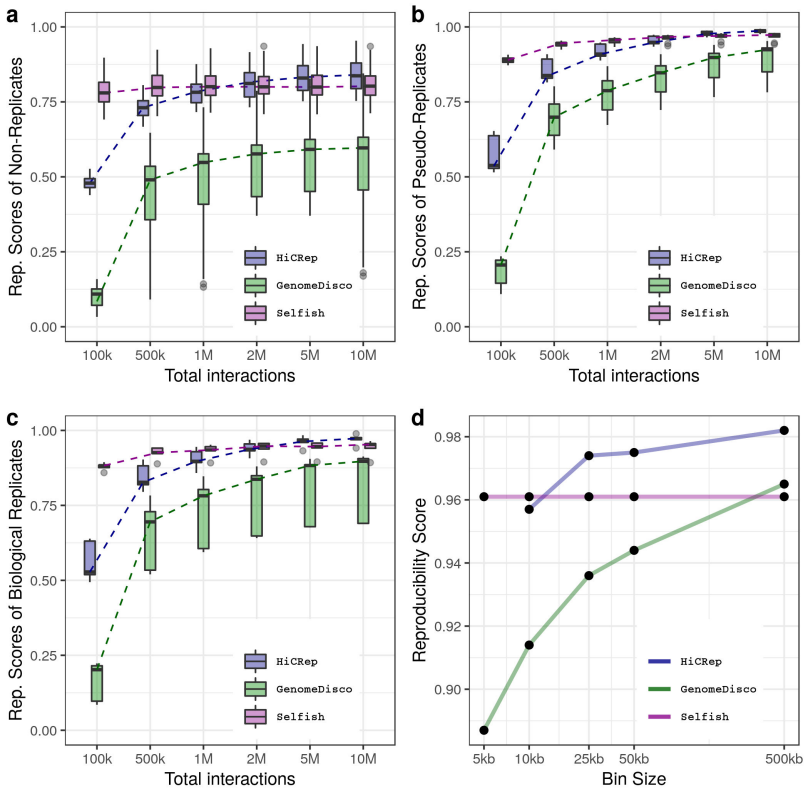


Figure 3.3: Illustrating the effect of the total number of interactions on reproducibility score of **(a)** non-replicates, **(b)** pseudo-replicates and **(c)** biological replicates. Panel **(d)** illustrates the effect of data resolution (bin size) on reproducibility score of two replicates of cell type GM12878 from [50]

.

visual inspection of Figure 3.4 shows that our reproducibility measure can cluster similar cell and tissue types. For instance, observe that SELFISH correctly clusters the right and left ventricles, IMR90 and lung as well as hippocampus and cortex together.

These experimental results clearly indicate that SELFISH outperforms existing methods in terms of robustness to changes in sequencing depth and binning size. Both of these are very desirable features which can significantly simplify Hi-C data analysis in terms of quality control for reproducibility in replicate experiments.

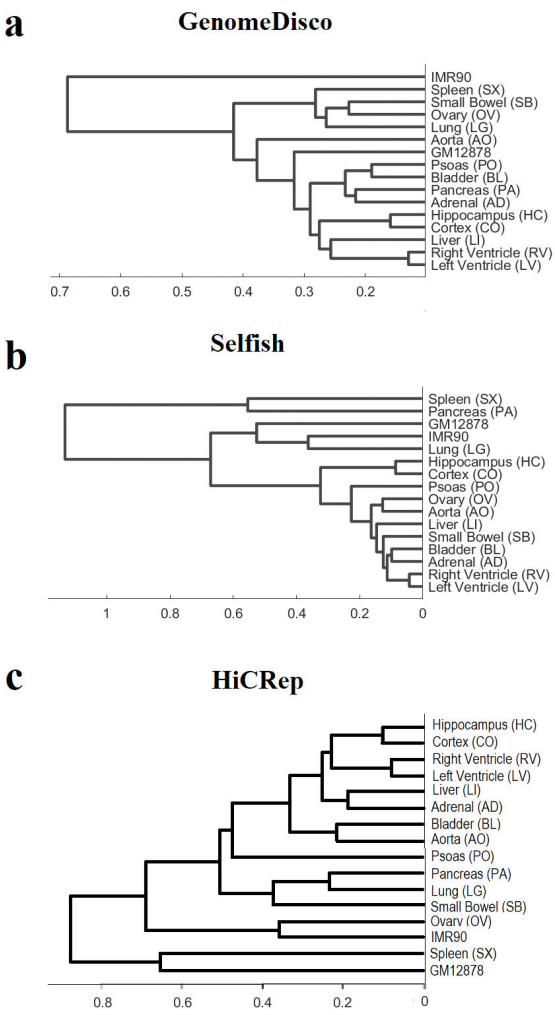


Figure 3.4: Clustering of fourteen human primary tissues and two cell lines obtained from [53]. The dendrograms are computed based on the pairwise similarity calculated using **(a)** GenomeDisco, **(b)** SELFISH and **(c)** HiCRep.

We also compared the average running time of the three methods on two replicates of GM12878 cell type from [50]. Table 3.1 show that SELFISH is by far more efficient than HiCRep and GenomeDISCO.

3.4.2 Differential Chromatin Interaction

We compared Selfish to the current state-of-the-art method for detecting differential chromatin interaction called FIND [18]. To the best of our knowledge, FIND is the only DCI detection method which works on high resolution Hi-C data by taking into account the chromatin interactions inter-dependency. Extensive experimental results in [18] show that FIND performs better than previously published methods for DCI detection.

We ran Selfish and FIND on Hi-C contact maps for cell types GM12878 and K562 obtained from [50]. We replicated some of the experiments proposed in [18] in order to make a fair comparison with FIND. First, we analyzed the enrichment of epigenetic signals in the neighborhood of detected DCIs as well as the percentage of nearby genes having significant expression fold changes. We evaluated the enrichment of four important epigenetic markers, namely the binding of CTCF, POLII and P300, as well as the presence of histone modification H3K4me3. CTCF is widely recognized as a main driver of chromatin structure [50, 60]. We computed the enrichment of CTCF differential peaks (i.e., peaks that are different between two cell types) around detected DCIs. For this part of the analysis we obtained the FIND's detected DCIs from [18].

To compute the enrichment of each marker in the neighborhood of the detected DCIs, we calculated the distance of marker peaks to their closest anchor of DCIs. Figure 3.5

Method	500kb	50kb	25kb	10kb	5kb
HiCRep	6s	636s	1045s	34479s	*–
GenomeDisco	7s	2989	4933s	30238s	61004s
Selfish	0.75s	85s	184s	345s	474s

*HiCRep fails to run on 5kb data on a server machine with 256 GB of RAM.

Table 3.1: Average running time of HiCRep, GenomeDisco and Selfish for different choices of the data resolution

shows the enrichment of epigenetic markers near DCIs. Observe that CTCF and H3K4me3 are more enriched around the SELFISH's reported DCIs than those detected by FIND, even though the number of reported DCIs for SELFISH is twice as large (30456 vs. 14131).

We also calculated the expression fold change of nearby genes for the two cell types. We first determined the set of genes which have overlap with any of the detected DCIs' anchors. Then we computed the percentage of those genes having an expression fold change of two or greater. For the set of genes overlapping FIND's DCIs, 71.46% of them were over-expressed. For SELFISH, 78.78% were over-expressed. This analysis confirmed that the differences in chromatin structure are strongly associated with the changes in gene regulation. However, the DCIs detected by SELFISH have stronger associations to differences in gene regulation than FIND. For the gene expression analysis, we used the dataset in [18]. The expression data was obtained from ENCODE, accession numbers GSE78553 and GSE78625 for cell types GM12878 and K562 respectively.

To quantify how well the Hi-C data supported the detected DCIs between two cell types GM12878 and K562, we generated a modified aggregate peak analysis (APA) plots [50, 49]. The interaction frequencies in contact maps were first Z-normalized along the diagonals as explained in Methods section. Then, for each detected DCI we calculated the interaction differences between the contact maps in a $\pm$50Kb neighborhood. By averaging over all DCIs, we computed the APA plot for differences. The differential APA score, i.e. the value of the central index in the plot compared to neighboring regions, shows how different the interactions are at reported DCIs with respect to their expected interaction frequency with that genomic distance.

Figure 4.10 shows that SELFISH produces the expected Gaussian-shaped plot around its reported DCIs while the DCIs from FIND failed to generate a similar pattern. Correspondingly, the computed APA score is higher for SELFISH (4.46) compared to FIND (3.06) suggesting a stronger detection of DCIs. Finally the peak pixel of the APA plot for FIND (score of 4.17) is not centered on the called DCI pairs suggesting that SELFISH performs better at pinpointing anchor points of chromatin interactions at high resolution.

We also compared the runtime of FIND and SELFISH on the 5kb-resolution dataset described above. FIND failed to run on the whole genome or even on large segments of contact maps. To compare the efficiency of SELFISH and FIND, we computed the average run-time required to process random contact map segments of 6Mbp. SELFISH took an average of 101.6s. FIND required an average of 12839.6s, about 120x slower than SELFISH.

To further investigate the accuracy of detected DCIs we used simulated Hi-C data generated by the method proposed in [68]. We generated 100 pairs of simulated contact maps, each of which had known location for DCIs. After running SELFISH and FIND on these simulated datasets, we obtained a p-value for each DCI location for all 100 simulated pairs of contact maps. Given the p-values and the true locations of DCIs (true positives) we computed a precision-recall curve for each simulated pair of contact maps. We used the threshold averaging method proposed by [22] to combine the 100 precision-recall curves to get the overall performance curve. We thresholded over the ratio of all indices in the contact map used for computing the precision and recall for each simulated pair. To combine the curves, we averaged all 100 calculated precision and recall values for each threshold. Figure 3.7a-c shows the performance of both methods for 2-fold, 5-fold and 10-fold DCIs.

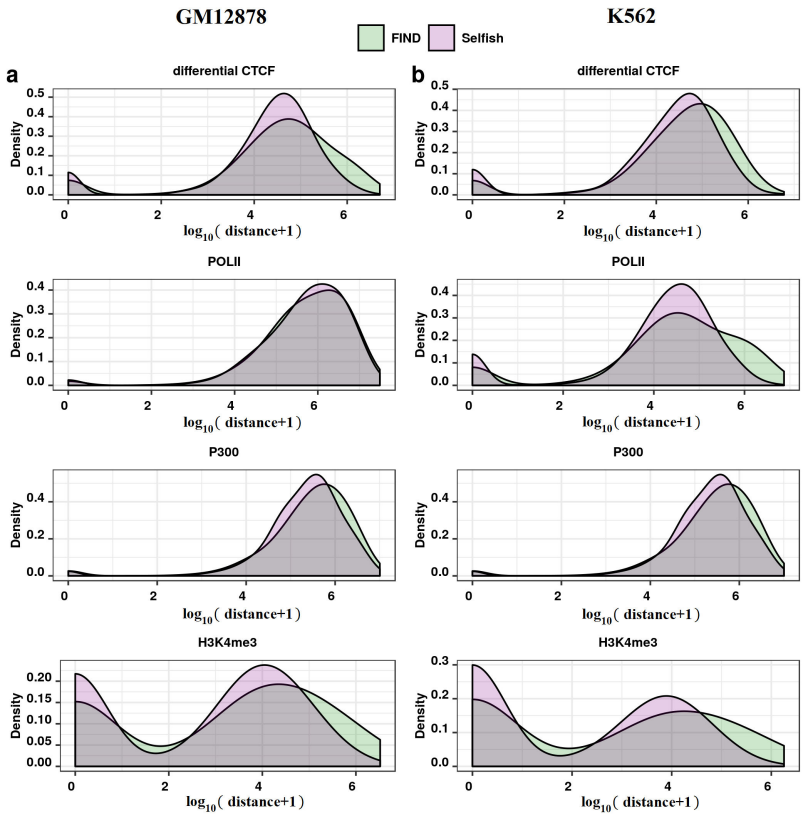


Figure 3.5: Enrichment of differential transcription factor binding and epigenetic marks (CTCF, POLII, P300 and H3K4me3) around reported DCIs for **(a)** cell type GM12878 and **(b)** cell type K562.

The vertical and horizontal bars represent the 95% confidence interval for precision and recall at that threshold respectively.

Selfish performed better than FIND on all fold change settings, confirming the stronger performance that we observed on real Hi-C data. The performance difference is most striking for small fold change values, which are more relevant for comparisons of real Hi-C datasets and yet have a very large effect on gene regulation [28]. It is also important to note that Selfish's performance is quite consistent across different samples as indicated by small confidence intervals.

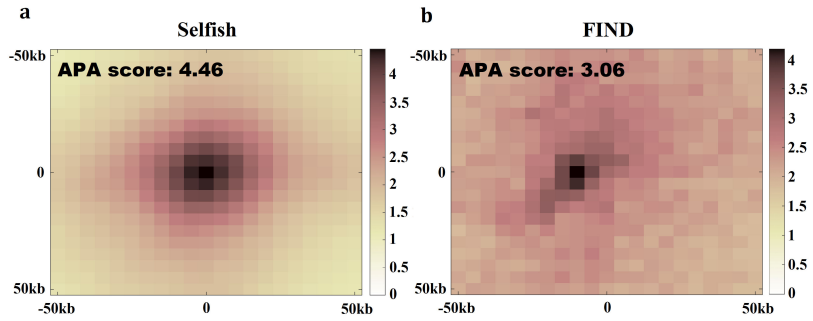


Figure 3.6: A modified APA plots for reported DCIs between two cell types GM12878 and K562 by **(a)** SELFISH and **(b)** FIND.

Figure 3.7d shows the distribution of distances between false positive DCIs produced by SELFISH and FIND to true DCIs (true positives). To generate this figure we set the number of returned DCIs equal to the number of true DCIs. Then, for each falsely detected DCI we calculated its distance to the closest true DCI. These results clearly show that most of SELFISH's false positives are located in close proximity of true DCIs confirming their relevance to true differences and their non-random distribution. FIND's false positive are instead much farther from true DCIs and are more scattered in the contact map.

In our final experiment to assess the performance of two methods, we tested SELFISH and FIND on a real test case from [5]. Figure 3.8 shows a 2-Mb region around the *Brn2* promoter (also known as *Pou3f2*) for mouse embryonic stem cells (ES) and neuronal progenitor cells (NPC). Dashed circles show the contact between the *Brn2* promoter and an NPC specific enhancer. Insets show the magnified view of this contact. Observe that the contact between the promoter and enhancer is strongly present in the NPC cell (Figure 3.8a) in contrast with the ES cell in which this interaction is weak (Figure 3.8b). The mentioned contrast shows itself as a subtle but important difference of interactions between two cell types. The highlighted regions of the epigenetic signals show the difference in the specified

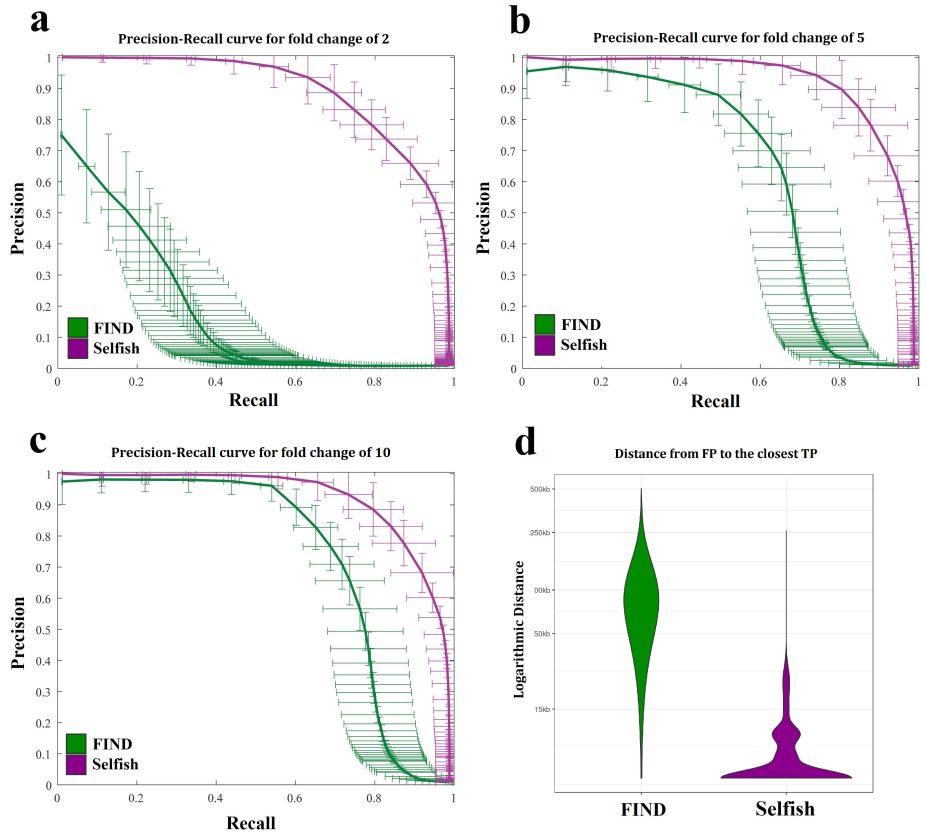


Figure 3.7: Precision-Recall curves for SELFISH (magenta) and FIND (green) for **(a)** 2-fold, **(b)** 5-fold and **(c)** 10-fold DCIs. The vertical and horizontal bars represent the 95% confidence interval for precision and recall at that threshold respectively. **(d)** The distribution of distances of FPs to closest TPs.

regions between two cell types. Detected DCIs by SELFISH and FIND for q-value $< 10^{-4}$ are shown in magenta and green squares respectively. Observe that SELFISH can identify this contact region as a DCI between two cell types, but FIND fails to detect it.

3.5 Conclusion

We presented a new approach for comparative analysis of Hi-C data using a novel self-similarity measure. We showed the utility of our measure by providing solutions to two

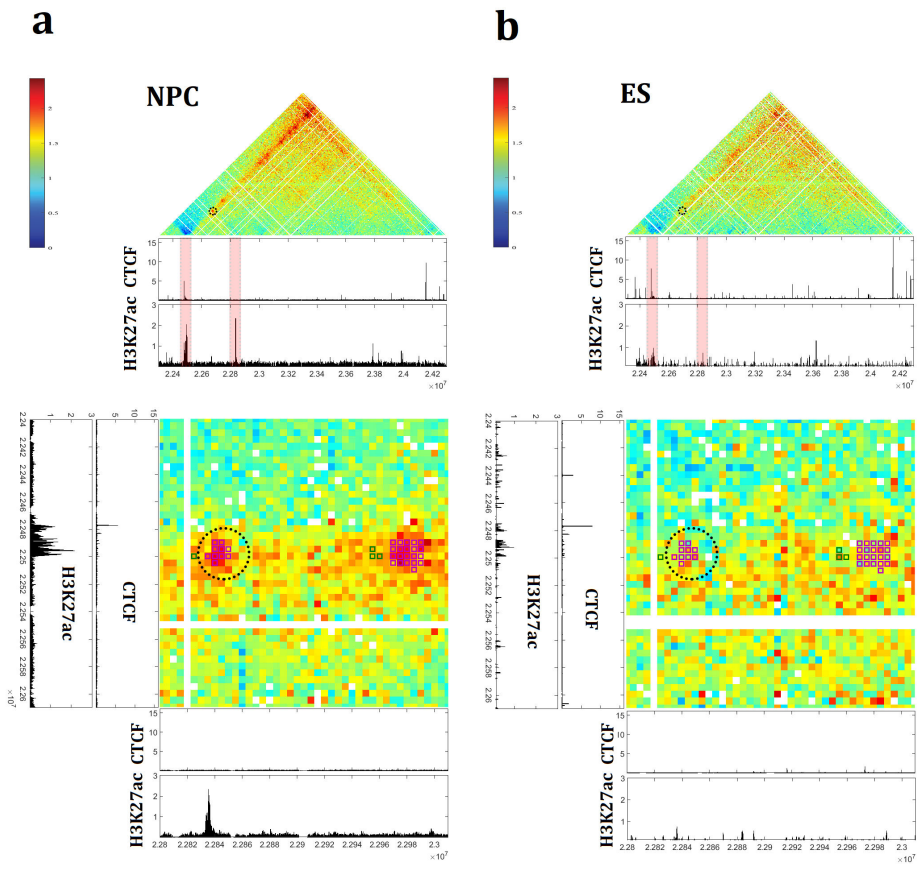


Figure 3.8: A 2-Mb region shown around Brn2 promoter of chromosome 4 of mouse neural cells **(a)** ES and **(b)** NPC. Dashed circles show the contact between the Brn2 promoter and an NPC specific enhancer. Insets show the magnified view of this contact.

important problems in the analysis of Hi-C data, namely the problem of measuring reproducibility of Hi-C replicated experiments and the problem of finding differential chromatin interactions between two contact maps.

We showed that a simple binary comparison operation between blocks in the contact maps can be used to encode the local and global features in a manner that is robust to the data resolution and sequencing depth. This encoded information is used to build a feature vector for each contact map, which in turn allows to define a simple but effective

similarity metric using the distance between their feature vectors. Experimental results showed that our self-similarity based measure outperformed two state-of-the-art methods (HiCRep and GenomeDISCO) for measuring reproducibility of replicated Hi-C experiments.

We also introduced a new method for finding differential chromatin interactions between two contact maps. SELFISH is designed based on the idea that each pairwise chromatin interaction can be represented by its neighboring interactions. Therefore, each interaction difference reveal itself as a weighted impact on the neighboring interactions. We capture this impact using a set of gradually increasing Gaussian filters. By extensively testing SELFISH on simulated and real test data we showed that it outperforms the state-of-the-art DCI detection method FIND both in accuracy and efficiency.

Chapter 4

Multi-scale Detection of Statistically Significant Interactions in Hi-C contact maps

4.1 Introduction

Several studies have proposed methods and models to detect statistically-significant chromatin interactions from Hi-C or other 3C-based experiments. Proposed methods fall in two categories.

The first group contains methods that (i) fit statistical/probabilistic models to the interaction data and (ii) assign p-values to individual interaction bins by comparing observed values to expected values based on the fitted model. For instance, Fit-Hi-C (i) uses a monotonic spline as a model for the interaction data with respect to the genomic distance

of the interacting loci, then (ii) it estimates the confidence (p-value) of each interaction using the fitted binomial distribution, and finally (iii) it corrects the computed p-values by applying a multiple hypothesis testing procedure [1]. Crucial drawbacks of such methods are that (i) the locality information in the contact map is not taken into account in the modeling, (ii) all interactions are considered independent, (iii) interactions that are in the vicinity of a significantly strong interaction are very likely to be significant as well. As a consequence of (iii), the most significant interactions are likely to cluster in a few local neighborhoods of the contact maps, making it difficult for scientists to evaluate the significant interactions genome-wide.

The second group contains methods that use peak-calling to detect statistically significant interactions by taking advantage of local information in the contact map. For instance, Rao *et al.* developed a method called Hi-C Computational Unbiased Peak Search (HiCCUPS) that detects *chromatin loops* which can be counted as a certain class of significant chromatin interactions [50]. Chromatin loops are created when pairs of genomic sites that lie far apart along the linear genome are brought into proximity by some proteins. HiCCUPS examines each pixel (interaction bin) in the contact map by comparing its interaction frequency with the interaction frequencies of its predefined neighborhoods. The algorithm identifies loops by finding "enriched" pixels, that is, loci pairs whose contact counts is significantly larger than the contact counts of its four neighborhoods, namely (1) pixels to its lower-left, (2) pixels to its left and right, (3) pixels above and below, and (4) a doughnut-shaped region surrounding the pixel of interest (see Figure 4.1). The problem with this class of methods is that they generally use a fixed size local neighborhood to

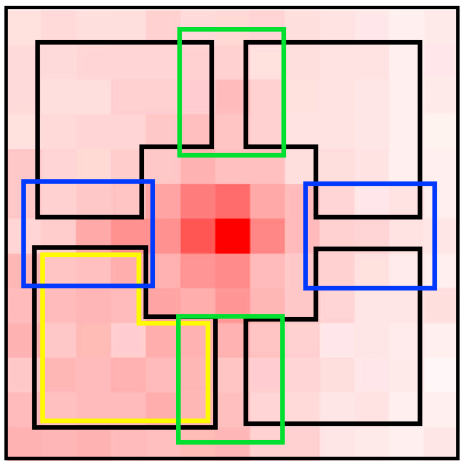


Figure 4.1: HiCCUPS significance test. Peaks are identified by detecting pixels (window's center) that are enriched with respect to four local neighborhoods of interactions shown in blue, yellow, green and black colors (source [50]).

model the background interactions. Therefore, significant interactions which are caused by proximity of larger segments of DNA (and subsequently reflected as larger enriched regions of interactions in the contact map) do not meet the local filtering conditions and will not be recovered by the method.

In this chapter, we present a new method called *Mustache* that addresses all the drawbacks mentioned above. We show that *Mustache* can detect statistically-significant interactions that are independently supported by other genomic and epigenetic data thus are likely to have biological relevance. For example, many of the detected interactions are associated with promoter-enhancer or promoter-promoter contacts, or are enriched for specific epigenetic markers or are stable across cell types.

4.2 Scale-Space modeling

Objects in real world, as opposed to idealized mathematical entities such as points or lines, are composed of a variety of structures at different scales which often makes them very difficult to detect in the absence of *a priori* knowledge about their true scales. A way of addressing this issue is to describe each object at multiple scales, making it possible to analyze each structure at its own appropriate scale. In our specific problem on contact maps, significant chromatin interactions are "blob-shaped objects" with a scale that depends on the size of the interacting DNA fragments.

Scale-space theory is a framework developed by the computer vision community for multi-scale representation of image data. In scale-space theory, each image is represented as a set of smoothed images. In order to build a scale-space representation of an image, a gradual smoothing process is conducted via a kernel of increasing width, producing a one-parameter (i.e., kernel size) family of images. As the smaller structures in finer scales are being suppressed by smoothing the image, the larger structures can be captured in coarser scales (Figure 4.2a).

The most common type of scale-space representation uses the Gaussian kernel because of its desirable mathematical properties. In particular, the causality property of Gaussian kernel guarantees that any feature at a coarse resolution scale is caused by existing feature[s] at finer resolution scales. In other words, this property makes sure that the smoothing process cannot introduce new extrema in the coarser scales of the scale-space representation of an image [38].

The Gaussian-kernel scale-space of an image $A(x, y)$ is a function $L(x, y, \sigma)$ obtained from the convolution of a variable-scale Gaussian $G(x, y, \sigma)$ with the input image, that is

$$L(x, y, \sigma) = G(x, y, \sigma) * A(x, y),$$

where $*$ represents the convolution operation in x and y, and

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}$$

(more details can be found in [41]).

Blob-shaped objects can be typically detected in an image by finding the strong responses in the application of *Laplacian of the Gaussian* operator with an image, that is

$$\nabla^2 \quad = L_{xx} + L_{yy}$$

Lindeberg showed that the normalization of the Laplacian with the factor σ^2 ($\sigma^2\nabla^2$) provides the true scale invariance required for detecting blob-shaped objects at different scales [38]. According to [40], the scale-normalized Laplacian ($\sigma^2\nabla^2$) can be accurately and efficiently estimated by the *difference-of-Gaussian* (DoG) function. Therefore, to detect the blob-shaped objects of varying scale, we can look for the scale-space maxima of the DoG function $D(x, y, \sigma)$ convolved with the image which can be computed from the difference of two nearby scales (in a scale-space representation) separated by a constant multiplicative factor k, that is

$$\begin{aligned} D(x, y, \sigma) &= (G(x, y, k\sigma) - G(x, y, \sigma)) * A(x, y) \\ &= L(x, y, k\sigma) - L(x, y, \sigma) \end{aligned}$$

4.3 Methods

Since a contact map is a special type of digital image, a set of interacting loci pairs that belong to a region of significant interactions, is a somewhat-circular (blob) structure at a specific scale in a scale-space representation. The causality property discussed above guarantees that no artifacts will be introduced due to the application of the Gaussian operator.

MUSTACHE's objective is to find blob-shaped regions of interactions with high statistical significance, i.e., regions with an average interaction significantly greater than the expected interaction. Due to random polymer interactions driven by one-dimensional genome proximity, interactions between pairs of loci that are closer in genomic distance are more frequent than interactions between loci at higher genomic distances. To account for the amplification of contact frequency due to 1D proximity, MUSTACHE performs a local z-normalization of the interaction frequencies in the contact map A with respect to their genomic distances along each diagonal d. More specifically, MUSTACHE re-scales the interactions by the logarithm of the expected interaction of the corresponding distance, as follows

$$\tilde{A}(i,j) = \frac{A(i,j) - \mu_{d_{ij}}}{\sigma_{d_{ij}}} \log(1 + \mu_d)$$

where $d = |j - i|$, μ_d is the average interaction on diagonal d, and $\mu_{d_{ij}}, \sigma_{d_{ij}}$ (not to be confused with Gaussian scale σ) are the local average and standard deviation along the diagonal d, respectively.

Then, MUSTACHE constructs the scale-space representation D of the normalized contact map $\tilde{A}$. As explained above, in order to compute $D(x, y, \sigma)$, MUSTACHE convolves

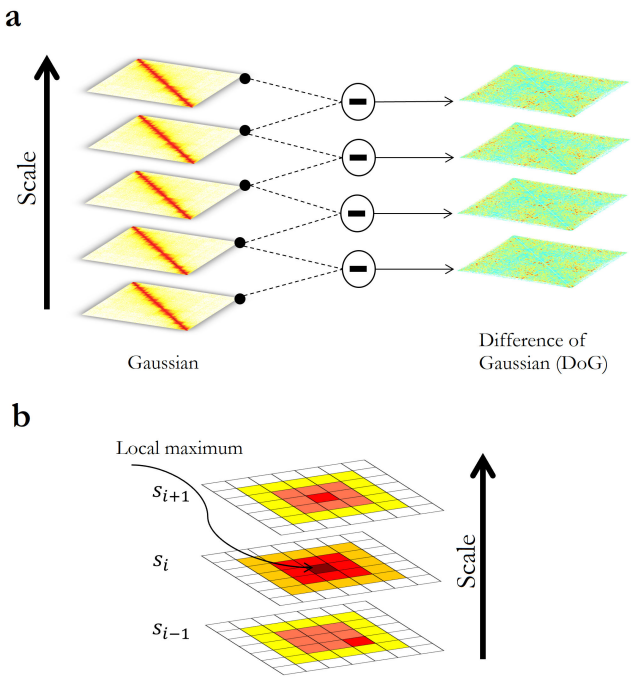


Figure 4.2: **(a)** The initial contact map is repeatedly convolved with gradually increasing Gaussians to produce a scale-space representation of the image (shown on the left). Pairwise adjacent Gaussian images are subtracted to produce the difference-of-Gaussian images (on the right); **(b)** Maxima of the difference-of-Gaussian images are detected by comparing each pixel to its 3×3×3 neighborhood in (x, y, σ) space (source [41]).

$\tilde{A}$ with gradually increasing Gaussians. This process produces a set of smoothed contact maps separated by a constant factor k in scale-space. MUSTACHE computes the difference of Gaussians by subtracting pairwise adjacent smoothed contact maps (Figure 4.2).

MUSTACHE uses two octaves of scale-space which is achieved by successive doubling of the scale parameter σ. Each octave is divided into s of intervals, such that $k = 2^{1/s}$ [41]. MUSTACHE computes the p-value $P_{\sigma_k}(x, y)$ for each pixel $D(x, y, \sigma = \sigma_k)$ by fitting a Laplace distribution on each scale of DoG, as follows.

$$P_{\sigma_k}(x, y) = \mathbf{Pr}\left(X > D\left(x, y, \sigma = \sigma_k\right)\right)$$

where X is distributed according to the Laplace distribution. MUSTACHE uses the Benjamini-Hochberg procedure to correct for multiple hypothesis testing. In case multiple candidates are available for one location (at different scales), MUSTACHE will report the one with the smallest p-value.

After computing the difference-of-Gaussian $D(x, y, \sigma)$ as explained in the previous section, MUSTACHE looks for local maxima in $D(x, y, \sigma)$. Specifically, MUSTACHE compares each pixel (x, y) to its eight neighbors in the current smoothed image and eighteen neighbors in the scale above and below (Figure 4.2b). If the value of pixel (x, y) is larger than all the neighboring pixels then it is selected as a candidate for a significant chromatin interaction region.

Detected candidates undergo a few filtering steps as described next. In the first filtering step, MUSTACHE removes candidates that are not local maximum (in 3×3 neighborhood) at least in two consecutive scales, i.e., it discards candidates that are local maximum at scale σ_i but not a local maximum at scale σ_{i-1} or σ_{i+1}. In the second filtering step, MUSTACHE computes the connected components for all candidate pixels. To compute connected components MUSTACHE uses 8-connectivity, i.e., a 3×3 neighborhood around each pixel. In each connected component, MUSTACHE keeps a single pixel with the lowest p-value.

In the third filtering step, MUSTACHE filters out candidates that are located in sparse regions of the contact map. Specifically, it discards interactions whose neighborhood (with size equal to the interaction scale) contain more than 20% unknown interactions. In

the fourth and final filtering step, candidates with interaction frequency smaller than two times the expected frequency (i.e. the interaction mean for the corresponding distance) will be discarded.

4.4 Experimental results

We compared MUSTACHE to the state-of-the-art loop-calling algorithm HiCCUPS [50]. All experiments were conducted on Hi-C data for human cell lines GM12878 and K562 obtained from [50] for interactions within 2Mb distance. For HiCCUPS, the results are directly obtained from [50].

We start by illustrating some case studies, then report on numerical evaluations. Figure 4.3 shows a few arbitrary locations on chromosome 1, 4 and 12 for GM12878. MUSTACHE's reported interactions are shown in the lower diagonal matrix using blue circles (where the radii represent the scale of interactions and their corresponding interacting fragments). HiCCUPS loops are represented by red dots on the upper diagonal matrix. Observe that MUSTACHE is calling most of the HiCCUPS loops, but it calls additional interactions which HiCCUPS fails to detect. Also observe that all detected interactions fall inside or on the boundary of topologically-associating domains (TADs), indicating that reported pairs of loci are within the same regulatory domain (consistent with the findings in [50]).

Figure 4.3-a shows the reported interactions for both methods for a region of size ∼1Mb on chromosome 1 (GM12878 cell line). Observe that MUSTACHE is calling the five chromatin loops reported by HiCCUPS as well as four additional interactions. The different radii of the circles illustrate the scales of the detected interactions, which are the products

of spatial proximity of DNA segments with different sizes. The organization of detected interactions is associated with the hierarchical organization of TADs, visually identifiable in the contact map. Figure 4.3-d shows the reported interactions for both methods for a region of size ∼1Mb on chromosome 12. Observe that MUSTACHE is detecting five significant interactions between an enhancer on the border of a TAD (upper-left corner of the shown contact map) and five additional loci spread out along the domain border, which is likely to indicate the presence of a super-enhancer. Of these five locations, HiCCUPS detects only one. Observe that there is one interaction called only by HiCCUPS on the boundary of the second TAD.

In Figure 4.4 and Figure 4.5 we take a closer look at the panels (a) and (d) of Figure 4.3. In these figures the contact maps are rotated 45 degrees, so that the main diagonal is horizontal and the interactions are represented as heat maps. Below the contact map we reported: (i) genomics coordinates, (ii) gene annotations (genes on the negative strand are shown in red color), (iii) CTCF motifs and their orientation, (iv) epigenetic signals SMC3, CTCF, RAD21, H3K4me3 and H3K27ac (v) and MUSTACHE's detected interactions shown as arcs (connecting two loci). Interactions detected by MUSTACHE and HiCCUPS are shown in green. Interactions detected only by MUSTACHE shown in blue. Interactions detected only by HiCCUPS are shown in red.

Observe in Figure 4.4 that most of the interactions detected only by MUSTACHE (blue) connect peaks for structural element signals (SCMC3, CTCF and RAD21). As an example, the MUSTACHE-only interaction denoted by '*' is connecting two loci that (i) have strong peaks in all three structural signals and (ii) correspond to convergent CTCF.

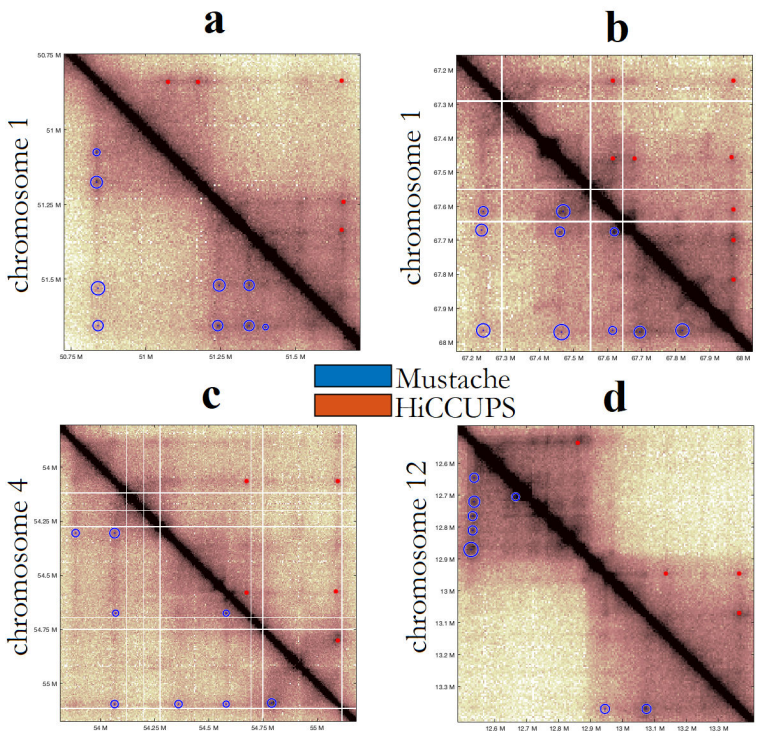


Figure 4.3: A few examples of MUSTACHE's and HiCCUPS' reported interaction. MUSTACHE's interactions are shown by blue circles (lower triangular) while HiCCUPS loops are shown by red dots (upper triangular). The outputs of two methods are shown for four different regions of GM12878 cell line, namely, **(a)** 50.75Mb-51.75Mb on chromosome 1, **(b)** 67.2Mb-68Mb on chromosome 1, **(c)** 53.8Mb-55.2Mb on chromosome 4, **(d)** 12.5Mb-13.4Mb on chromosome 12.

Taken together, these are strong evidence to support the existence of a chromatin loop [50]. In Figure 4.5, observe the interaction between the very first locus (on the left) with the five downstream loci. Among these interactions, only one is detected by HICCUPS. Four of these loci are in proximity of structural signal peaks and two in proximity of histone modification H3K4me3 and H3K27ac peaks, indicating the presence of a super-enhancer.

To measure the robustness of MUSTACHE and HiCCUPS and the reproducibility of their detected interactions, we ran the tools on two replicates of GM12878 cell line.

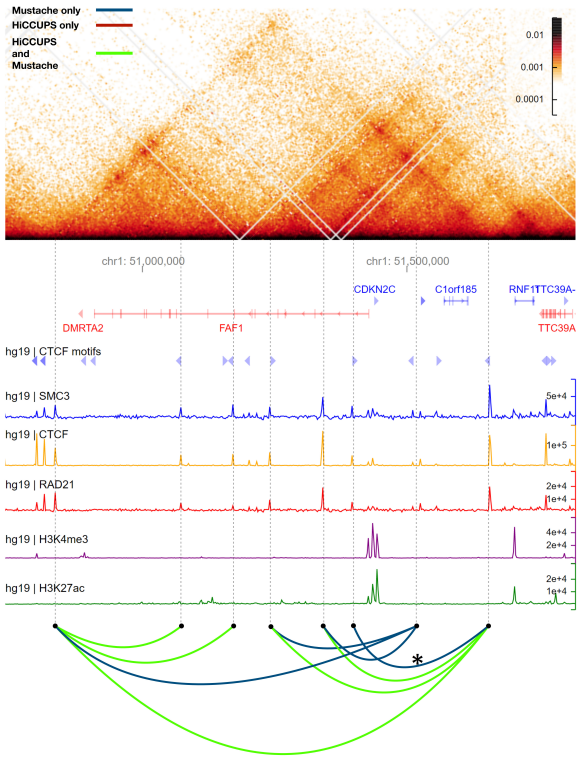


Figure 4.4: A comparison between MUSTACHE's and HiCCUPS' reported chromatin interactions in a region of chromosome 1 for the GM12878 cell line (50.75Mb−51.75Mb). The contact map is shown on the top. Below the contact map, the figure shows gene annotations for positive (blue) and negative (red) strands, CTCF motifs (and their orientation), epigenetic signals SMC3, CTCF, RAD21, H3K4me3 and H3K27ac. Arcs indicate interaction detected by both MUSTACHE and HiCCUPS (green), only MUSTACHE (blue), only HiCCUPS (red)

We ran MUSTACHE twice: once by fixing the p-value threshold at $10^{-1.3}$, and once by fixing the number of reported interactions to match the number reported by HiCCUPS. We compared the two output lists by defining two interactions to *match* if their 7.5kb×7.5kb neighborhoods overlapped. The overlap between the outputs for the replicates is shown Figure 4.6a-c. Figure 4.6a is for HiCCUPS, Figure 4.6b is for MUSTACHE (with the same number of reported interaction as HiCCUPS), Figure 4.6c is for MUSTACHE (using p-value threshold of $10^{-1.3}$). Observe that MUSTACHE is significantly more consistent than HiC-

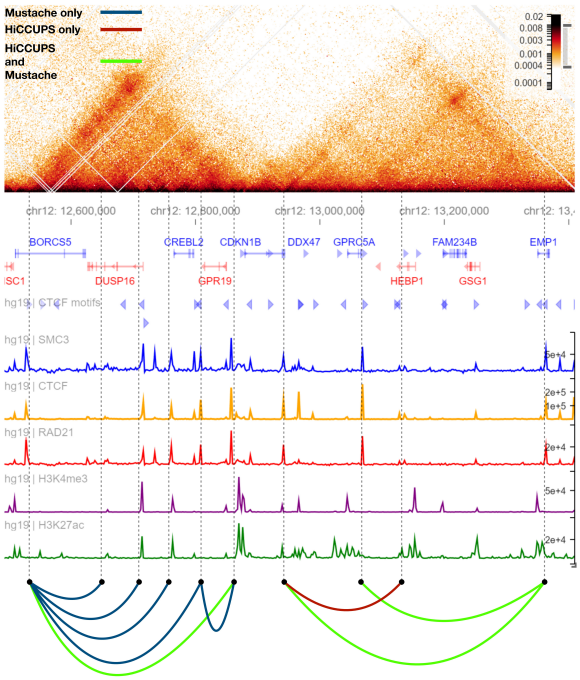


Figure 4.5: A comparison between MUSTACHE's and HiCCUPS' reported chromatin interactions in a region of chromosome 1 for the GM12878 cell line (12.5Mb−13.3Mb). The contact map is shown on the top. Below the contact map, the figure shows gene annotations for positive (blue) and negative (red) strands, CTCF motifs (and their orientation), epigenetic signals SMC3, CTCF, RAD21, H3K4me3 and H3K27ac. Arcs indicate interaction detected by both MUSTACHE and HiCCUPS (green), only MUSTACHE (blue), only HiCCUPS (red)

CUPS in terms of the reported calls between replicates, indicating a more reliable and reproducible detection methodology. For comparison, we carried out a similar analysis on two different cell lines, namely K562 and GM12878. Figure 4.6d-f illustrate the overlap of reported interactions that are conserved across the two cell lines.

Figure 4.7 illustrates the overlap between MUSTACHE's and HiCCUPS' reported interactions. For MUSTACHE, p-value thresholds of $10^{-1.5}$ and 10^{-1} are used for GM12878 and K562 cell lines, respectively. MUSTACHE recovered 70% of HiCCUPS's interactions in K562 (Figure 4.7a) and 83% HiCCUPS's interactions in GM12878 (Figure 4.7b), but

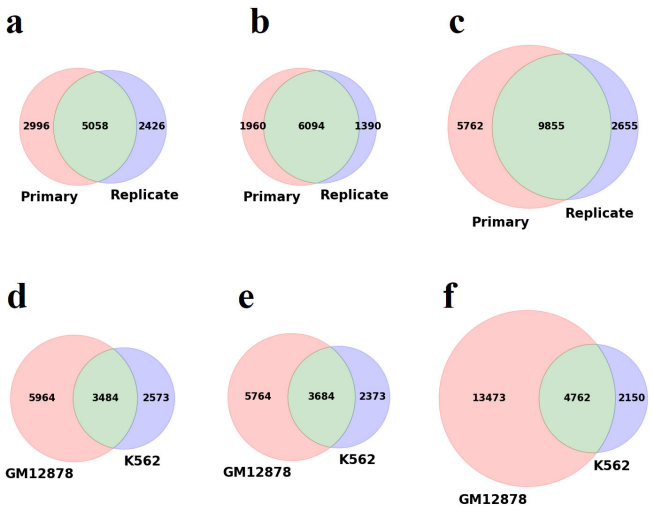


Figure 4.6: Reported chromatin interactions on two replicates for cell line GM12878 (a-c), and between two cell lines K562 and GM12878 for methods Mustache and HiCCUPS (d-f). **(a)** HiCCUPS's reported interactions on the two replicates; **(b)** MUSTACHE's reported interactions on the two replicates (with the same number of interactions as in **(a)**); **(c)** MUSTACHE's reported interactions on the two replicates (p-value threshold of $10^{-1.3}$); **(d)** HiCCUPS's reported interactions on GM12878 and K562, **(e)** MUSTACHE's reported interactions on GM12878 and K562 (with the same number of interactions as in **(d)**) **(f)** MUSTACHE's reported interactions on GM12878 and K562 (p-value threshold of $10^{-1.3}$).

reported many additional chromatin interactions.

To determine whether MUSTACHE's detected interactions were supported by other sources of evidence, we carried out a series of experiment using ChIA-PET, HiChIP, ChIP-seq and ChromHMM data types. In the first experiment, we computed what percentage of MUSTACHE's interactions are connecting a known promoter to a known enhancer, as annotated by ChromHMM [21]. More specifically we counted for how many of the interacting pair of loci, one locus overlaps with a promoter region, and the locus other overlaps with an enhancer region.

Figure 4.8-a show that 38.2% of MUSTACHE's interactions in GM12878 and 32.6% in K562 connect a promoter to an enhancer. In contrast, 37.8% of HiCCUPS' interactions

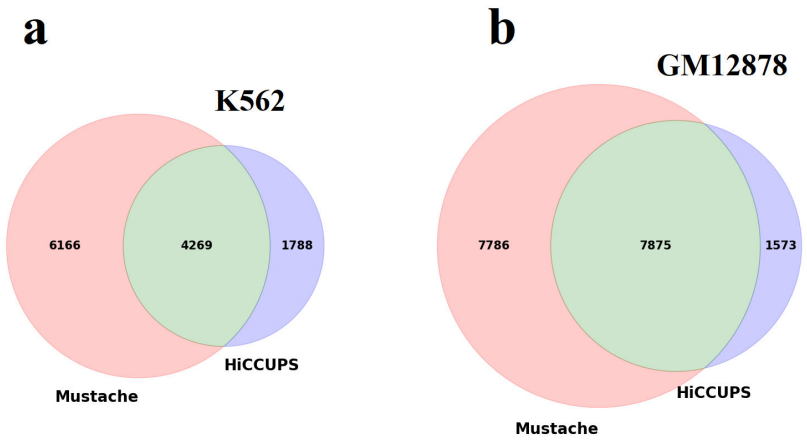


Figure 4.7: MUSTACHE's and HiCCUPS's reported chromatin interactions in K562 and GM12878 cell lines.

in GM12878 and 32.3% in K562 connect a promoter to an enhancer. Observe however that the absolute number of interactions by MUSTACHE is significantly higher. Similarly, Figure 4.8-b shows that 16.3% of MUSTACHE's interactions in GM12878 and 12.6% in K562 connect a promoter to another promoter. In contrast, 16.8% of HiCCUPS' interactions in GM12878 and 11.7% in K562 connect promoter to another promoter. But again, the absolute number of interactions reported by MUSTACHE is significantly higher for those reported by HiCCUPS.

Next, we compared the performance of MUSTACHE and HiCCUPS using ChIA-PET and HiChIP data. In this experiment, we assumed that the ChIA-PET/HiChIP interactions for GM12878 cell line were the ground truth, and computed the number of these interactions recovered by MUSTACHE and HiCCUPS by running them on Hi-C data for GM12878 cell line. We used the same matching criteria as in the consistency testing experiment above. Figure 4.9-a shows the recovery plot for the CTCF ChIA-PET interactions obtained from [29]. The x-axis represents the number of Hi-C chromatin loops called

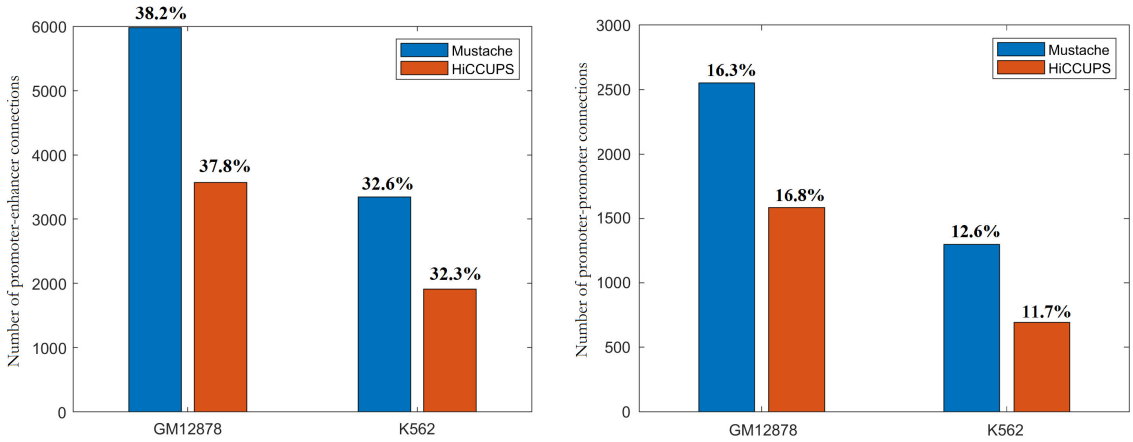


Figure 4.8: The number of chromatin interactions detected by MUSTACHE and HiCCUPS that connect promoters to enhancers and promoters to promoters (according to ChromHMM chromatin states). The percentage of all interactions called by each method, is reported above each bar. **(a)** The number of chromatin interactions detected by MUSTACHE and HiCCUPS that connect promoters to enhancers in cell lines GM12878 and K562. **(b)** The number of chromatin interactions detected by MUSTACHE and HiCCUPS that connect promoters to promoters in cell lines GM12878 and K562.

by MUSTACHE (blue) and HiCCUPS (red) sorted by their significance. For HiCCUPS the significance was the median of the q-values for all four local filters. For MUSTACHE, the significance was the reported p-value. The y-axis represents the percentage of the CTCF ChIA-PET interactions recovered by each method. HiCCUPS calls 9,484 chromatin loops in GM12878 cell line. Observe that at the point in which MUSTACHE calls the same number of loops, it recovers a greater percentage of CTCF ChIA-PET interactions than HiCCUPS. Also observe that (i) MUSTACHE can continue calling more loops and recovering more and more CTCF ChIA-PET interactions, (ii) the slope of MUSTACHE's recovery curve is far from saturation at the threshold of 9,448 interactions. Figure 4.9b-d shows the recovery plots of MUSTACHE and HiCCUPS for GM12878 Cohesin HiChIP HiCCUPS loops [45], GM12878 H3K27ac HiChIP Fithichip loops [3] and GM12878 RAD21 ChIA-PET interactions [29].

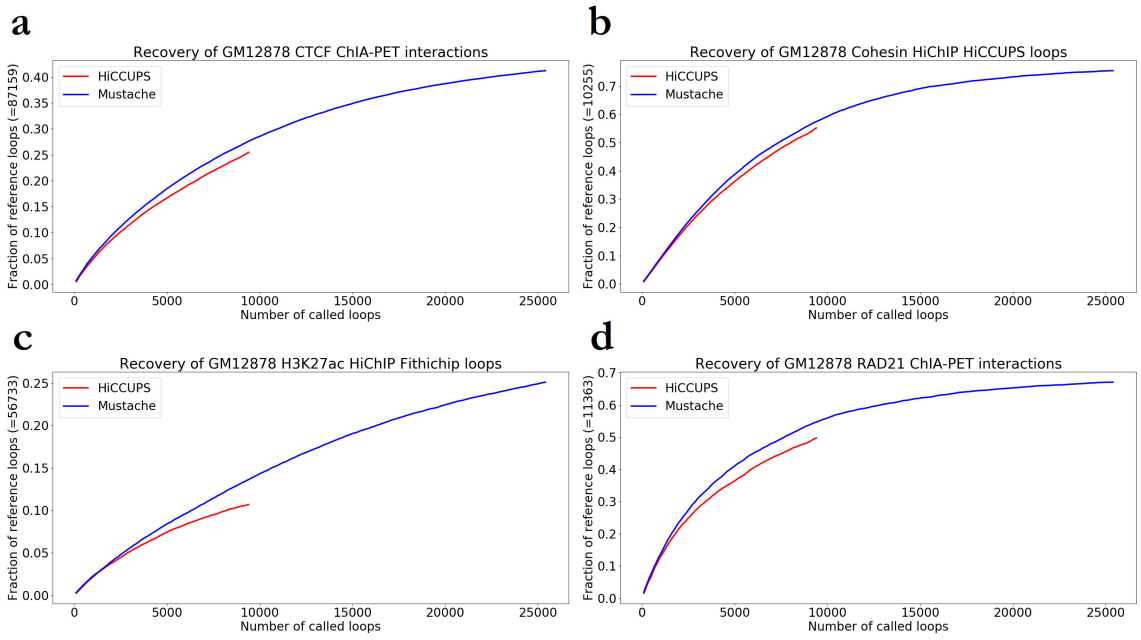


Figure 4.9: MUSTACHE and HiCCUPS recovery plots for **(a)** GM12878 CTCF ChIA-PET interactions **(b)** GM12878 Cohesin HiChIP HiCCUPS loops, **(c)** GM12878 H3K27ac HiChIP Fithichip loops, **(d)** GM12878 RAD21 ChIA-PET interactions.

Again, it is very clear from Figures 4.9b-d that MUSTACHE recovers a much higher fraction of validated interactions.

Next, we compared the enrichment of structural markers CTCF, RAD21 and SMC3 on the anchors of detected interactions for both methods. First, we identified the unique set of interacting loci for each method. Then, we determined the ratio of the interacting loci that contain one of structural marker listed above. When MUSTACHE was ran to produce the same number of detected interactions of HiCCUPS, 84.34% of the MUSTACHE's loci in GM12878 were enriched with RAD21, 84.12% with CTCF and 82.35% with SMC3. In comparison, the enrichment for HiCCUPS's interaction was 86.88% for RAD21, 86.93% for CTCF and 85.14% with SMC3. When we ran MUSTACHE loops with a p-value

threshold of $10^{-1.3}$ it produced almost two times the number of interactions, and 74.13% of the MUSTACHE's loci in GM12878 were enriched with RAD21, 74.19% with CTCF and 70.67% with SMC3.

To quantify how well the MUSTACHE significant interactions were supported by the Hi-C data, we used aggregate peak analysis (APA) [50, 49]. To generate APA plots, we aggregated interaction counts over all detected pairs of loci in a $\pm$50kb neighborhood. The result is illustrated as a 21×21 heatmap (at 5kb resolution) in which darker color indicates higher interaction count. A strong dark pixel at the center of the heatmap indicates that the number of chromatin interactions are much higher at detected loci compared with the neighboring loci. The sharper is the transition between dark and light pixels, the faster the interaction frequency decreases as we move away from the loop. For each plot we computed the APA score, which is the ratio between (a) the value of the center pixel and (b) the mean of pixels $15 - 30$kb downstream of the upstream loci and $15 - 30$kb upstream of the downstream loci [49]. Since the APA score is computed as the interaction ratio between the center pixel with a 3×3 neighborhood in the left-bottom corner which contains interactions with smaller genomic distance than the center's, the APA score would have a value smaller than one if the evaluating loci were picked by random. The plots are generated for significant interactions between loci separated by 150kb to 1Mb.

Figure 4.10 shows APA plots and APA scores on GM12878 and K562 cell lines for MUSTACHE (p-value $10^{-1.3}$) and HiCCUPS. Observe most interactions occur in the central region of the heatmap with a strong peak at the center, which indicates that HiCCUPS' and MUSTACHE's loops are enriched in Hi-C interaction. Observe that while HiCCUPS'

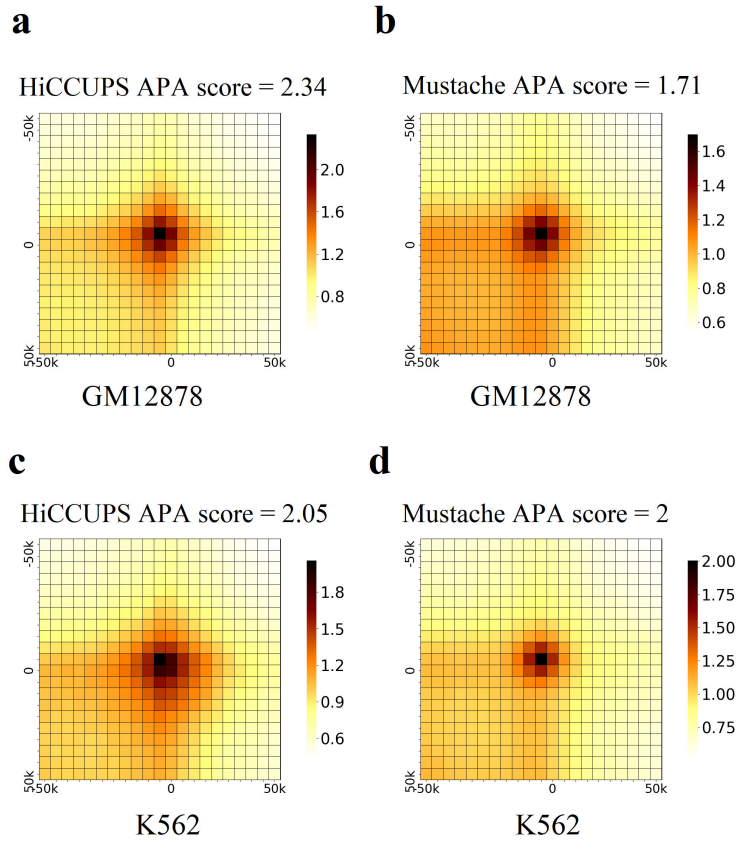


Figure 4.10: **(a)** APA plot for HiCCUPS in GM12878, **(b)** APA plot for Mustache in GM12878, **(c)** APA plot for HiCCUPS in K562 and **(d)** APA plot for Mustache in K562.

APA scores are higher than MUSTACHE, the APA plot for K562 produced for MUSTACHE is sharper.

Figure 4.11 shows the genomic distance distribution between the loci of chromatin interactions detected by MUSTACHE and HiCCUPS. Observe while MUSTACHE detects more interactions compared with HiCCUPS, the distributions have similar shapes.

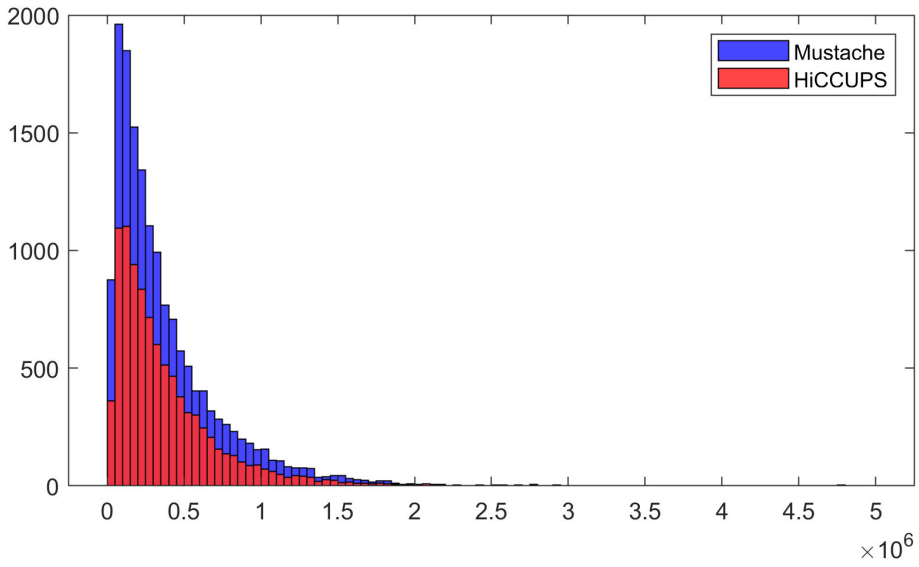


Figure 4.11: Genomic distance distribution between the loci of chromatin interactions detected by MUSTACHE and HICCUPS.

4.5 Conclusion

We presented a new method called MUSTACHE for detecting significant chromatin interactions in Hi-C data. By taking advantage of a scale-space representation of the contact map, MUSTACHE can capture chromatin interactions at different scales. Extensive experimental results show that MUSTACHE outperforms HiCCUPS in many aspects, including speed, the total number of reported interactions, and the fraction of interactions confirmed by other types of biological evidence.

Chapter 5

Conclusions

This dissertation introduced new efficient and accurate algorithms for the analysis of high throughput chromosome conformation capture data to provide insights into the functional aspects of the 3D genome organization.

In Chapter 2, we presented an efficient algorithm called EAST, to accurately identify topological associating domains in contact maps obtained from Hi-C experiment. The framework we presented for TAD identification is based on fast 2D-convolution of Haar-like features. EAST can be downloaded from https://github.com/ucrbioinfo/EAST. We performed a comparative evaluation of EAST on Hi-C data for human stem cells, mouse stem cells and mouse cortex cells. We showed that EAST extracts TADs as accurately as the state-of-the-art methods. TADs identified by EAST provides substantial enrichment of various epigenetic modification factors at their boundaries, confirming similar findings in previous studies. We also showed that EAST is very time efficient compared to other published methods. EAST is easy to use and for a given Hi-C dataset: the only parameter that

might need to be tuned by the user is the normalization factor for which we have provided some guidance.

In Chapter 3, we presented a new approach for comparative analysis of Hi-C data using a novel self-similarity measure. We showed the utility of our measure by providing solutions to two important problems in the analysis of Hi-C data, namely the problem of measuring reproducibility of Hi-C replicated experiments and the problem of finding differential chromatin interactions between two contact maps.

We showed that a simple binary comparison operation between blocks of interactions in the contact maps can be used to encode the local and global features in a manner that is robust to the data resolution and sequencing depth. This encoded information is used to build a feature vector for each contact map, which in turn allows to define a simple but effective similarity metric using the distance between their feature vectors. Experimental results showed that our self-similarity based measure outperformed two state-of-the-art methods (HiCRep and GenomeDISCO) for measuring reproducibility of replicated Hi-C experiments. We also introduced a new method for finding differential chromatin interactions between two contact maps. Selfish is designed based on the idea that each pairwise chromatin interaction can be represented by its neighboring interactions. Therefore, each interaction difference reveal itself as a weighted impact on the neighboring interactions. We capture this impact using a set of gradually increasing Gaussian filters. By extensively testing Selfish on simulated and real test data we show that it outperforms the state-of-the-art method FIND both in accuracy and efficiency. Selfish is publicly available at https://github.com/ucrbioinfo/Selfish.

In Chapter 4, we introduced a new method Mustache for detecting statistically significant chromatin interactions. We compared Mustache with state-of-the-art method HiCCUPS which is widely accepted as the standard approach for detecting chromatin loops. Based on the experimental results, we showed that Mustache not only recovers the majority of chromatin loops reported by HiCCUPS but also discover many more. We showed that these additional interactions are supported by several other independent epigenetic marks, thus are likely to be true interaction.

We showed that Mustache can recover more interactions validated by other data types, such as ChIA-PET and HiChIP. Moreover, Mustache is able to call more chromatin loops due to its scale invariance, i.e. it can detect significant interactions with varying effect size in their neighborhood, and robustness to small noises. Furthermore, reported interactions are shown to be associated with contacts between promoters and enhancers/promoters and conserved between cell types.

Finally, I would like to conclude with some thoughts about the future directions of my research. In our work on the 3D genome, we have introduced new approaches to detect local structures of chromatin in different scales as well as a new way of performing comparative analysis in Hi-C contact maps. Given the necessary tools for analyzing the Hi-C data, we believe the next step will be studying the link between chromatin folding and gene regulation. There have been a handful of studies using the 3D chromatin structure to discover the regulatory effects of a limited number of risk loci. We believe by exploiting the abundance of extracted information from Hi-C experiment and its successors about how the chromatin folds, as well as taking advantage of newly introduced experiments such

as perturb-seq [15], which evaluates the effect of individual CRISPR based perturbations on gene function, we can design and implement a general model that incorporates all this information to estimate the association of chromatin structures with gene regulation.

Bibliography

[1] Ferhat Ay, Timothy L Bailey, and William Stafford Noble. Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. *Genome Res.*, 24(6):999–1011, June 2014.

[2] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Stat. Methodol.*, 57(1):289–300, 1995.

[3] Sourya Bhattacharyya, Vivek Chandra, Pandurangan Vijayanand, and Ferhat Ay. Identification of significant chromatin contacts from hichip data by fithichip. *Nature communications*, 10(1):4221, September 2019.

[4] Boyan Bonev and Giacomo Cavalli. Organization and function of the 3D genome. *Nat. Rev. Genet.*, 17(11):661–678, October 2016.

[5] Boyan Bonev, Netta Mendelson Cohen, Quentin Szabo, Lauriane Fritsch, Giorgio L Papadopoulos, Yaniv Lubling, Xiaole Xu, Xiaodan Lv, Jean-Philippe Hugnot, Amos Tanay, and Giacomo Cavalli. Multiscale 3D genome rewiring during mouse neural development. *Cell*, 171(3):557–572.e24, October 2017.

[6] Jonathan Cairns, Paula Freire-Pritchett, Steven W Wingett, Csilla Várnai, Andrew Dimond, Vincent Plagnol, Daniel Zerbino, Stefan Schoenfelder, Biola-Maria Javierre, Cameron Osborne, Peter Fraser, and Mikhail Spivakov. CHiCAGO: robust detection of DNA looping interactions in capture Hi-C data. *Genome Biol.*, 17(1):127, June 2016.

[7] Yaqiang Cao, Xingwei Chen, Daosheng Ai, Zhaoxiong Chen, Guoyu Chen, Joseph McDermott, Yi Huang, and Jing-Dong J. Han. Accurate loop calling for 3d genomic data with cloops. *bioRxiv*, 2018.

[8] Giacomo Cavalli and Tom Misteli. Functional implications of genome topology. *Nat. Struct. Mol. Biol.*, 20(3):290–299, 5 March 2013.

[9] Haiming Chen, Jie Chen, Lindsey A Muir, Scott Ronquist, Walter Meixner, Mats Ljungman, Thomas Ried, Stephen Smale, and Indika Rajapakse. Functional organization of the human 4D nucleome. *Proc. Natl. Acad. Sci. U. S. A.*, 112(26):8002–8007, 30 June 2015.

[10] Jie Chen, Alfred O Hero, 3rd, and Indika Rajapakse. Spectral identification of topological domains. *Bioinformatics*, 32(14):2151–2158, 15 July 2016.

[11] Axel Cournac, Hervé Marie-Nelly, Martial Marbouty, Romain Koszul, and Julien Mozziconacci. Normalization of a chromosomal contact map. *BMC Genomics*, 13:436, August 2012.

[12] Emily Crane, Qian Bian, Rachel Patton McCord, Bryan R Lajoie, Bayly S Wheeler, Edward J Ralston, Satoru Uzawa, Job Dekker, and Barbara J Meyer. Condensin-driven remodelling of X chromosome topology during dosage compensation. *Nature*, 523(7559):240–244, 9 July 2015.

[13] Franklin C Crow. Summed-area tables for texture mapping. In *Proceedings of the 11th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '84, pages 207–212, New York, NY, USA, 1984. ACM.

[14] Job Dekker, Karsten Rippe, Martijn Dekker, and Nancy Kleckner. Capturing chromosome conformation. *Science*, 295(5558):1306–1311, February 2002.

[15] Atray Dixit, Oren Parnas, Biyu Li, Jenny Chen, Charles P Fulco, Livnat Jerby-Arnon, Nemanja D Marjanovic, Danielle Dionne, Tyler Burks, Raktima Raychowdhury, Britt Adamson, Thomas M Norman, Eric S Lander, Jonathan S Weissman, Nir Friedman, and Aviv Regev. Perturb-Seq: Dissecting molecular circuits with scalable Single-Cell RNA profiling of pooled genetic screens. *Cell*, 167(7):1853–1866.e17, December 2016.

[16] Jesse R Dixon, Inkyung Jung, Siddarth Selvaraj, Yin Shen, Jessica E Antosiewicz-Bourget, Ah Young Lee, Zhen Ye, Audrey Kim, Nisha Rajagopal, Wei Xie, Yarui Diao, Jing Liang, Huimin Zhao, Victor V Lobanenkov, Joseph R Ecker, James A Thomson, and Bing Ren. Chromatin architecture reorganization during stem cell differentiation. *Nature*, 518(7539):331–336, 19 February 2015.

[17] Jesse R Dixon, Siddarth Selvaraj, Feng Yue, Audrey Kim, Yan Li, Yin Shen, Ming Hu, Jun S Liu, and Bing Ren. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376–380, 17 May 2012.

[18] Mohamed Nadhir Djekidel, Yang Chen, and Michael Q Zhang. FIND: differential chromatin INteractions detection using a spatial poisson process. *Genome Res.*, February 2018.

[19] Josée Dostie and Job Dekker. Mapping networks of physical interactions between genomic elements using 5C technology. *Nat. Protoc.*, 2(4):988–1002, 2007.

[20] Sandrine Dudoit, Yee Hwa Yang, Matthew J Callow, and Terence P Speed. Statistical methods for identifying differentially expressed genes in replicated cdna microarray experiments. *Stat. Sin.*, 12(1):111–139, 2002.

[21] Jason Ernst and Manolis Kellis. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods*, 9(3):215–216, February 2012.

[22] Tom Fawcett. An introduction to ROC analysis. *Pattern Recognit. Lett.*, 27(8):861–874, June 2006.

[23] Darya Filippova, Rob Patro, Geet Duggal, and Carl Kingsford. Multiscale identification of topological domains in chromatin. In *Algorithms in Bioinformatics*, pages 300–312. Springer, Berlin, Heidelberg, 2 September 2013.

[24] James Fraser, Carmelo Ferrai, Andrea M Chiariello, Markus Schueler, Tiago Rito, Giovanni Laudanno, Mariano Barbieri, Benjamin L Moore, Dorothee C A Kraemer, Stuart Aitken, Sheila Q Xie, Kelly J Morris, Masayoshi Itoh, Hideya Kawaji, Ines Jaeger, Yoshihide Hayashizaki, Piero Carninci, Alistair R R Forrest, FANTOM Consortium, Colin A Semple, Josée Dostie, Ana Pombo, and Mario Nicodemi. Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation. *Mol. Syst. Biol.*, 11(12):852, 23 December 2015.

[25] Melissa J Fullwood, Mei Hui Liu, You Fu Pan, Jun Liu, Han Xu, Yusoff Bin Mohamed, Yuriy L Orlov, Stoyan Velkov, Andrea Ho, Poh Huay Mei, Elaine G Y Chew, Phillips Yao Hui Huang, Willem-Jan Welboren, Yuyuan Han, Hong Sain Ooi, Pramila N Ariyaratne, Vinsensius B Vega, Yanquan Luo, Peck Yean Tan, Pei Ye Choy, K D Senali Abayratna Wansa, Bing Zhao, Kar Sian Lim, Shi Chi Leow, Jit Sin Yow, Roy Joseph, Haixia Li, Kartiki V Desai, Jane S Thomsen, Yew Kok Lee, R Krishna Murthy Karuturi, Thoreau Herve, Guillaume Bourque, Hendrik G Stunnenberg, Xiaoan Ruan, Valere Cacheux-Rataboul, Wing-Kin Sung, Edison T Liu, Chia-Lin Wei, Edwin Cheung, and Yijun Ruan. An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature*, 462(7269):58–64, November 2009.

[26] Feiran Gong, Luan Sun, Zongdan Wang, Junfeng Shi, Wei Li, Sumeng Wang, Xiao Han, and Yujie Sun. The BCL2 gene is regulated by a special AT-rich sequence binding protein 1-mediated long range chromosomal interaction between the promoter and the distal element located within the 3'-UTR. *Nucleic Acids Res.*, 39(11):4640–4652, June 2011.

[27] David U Gorkin, Danny Leung, and Bing Ren. The 3D genome in transcriptional regulation and pluripotency. *Cell Stem Cell*, 14(6):762–775, 5 June 2014.

[28] William W Greenwald, He Li, Paola Benaglio, David Jakubosky, Hiroko Matsui, Anthony Schmitt, Siddarth Selvaraj, Matteo D'Antonio, Agnieszka D'Antonio-Chronowska, Erin N Smith, and Kelly A Frazer. Integration of phased Hi-C and molecular phenotype data to study genetic and epigenetic effects on chromatin looping. July 2018.

[29] Nastaran Heidari, Douglas H Phanstiel, Chao He, Fabian Grubert, Fereshteh Jahanbani, Maya Kasowski, Michael Q Zhang, and Michael P Snyder. Genome-wide map of regulatory interactions in the human genome. *Genome Res.*, 24(12):1905–1917, December 2014.

[30] Denes Hnisz, Abraham S Weintraub, Daniel S Day, Anne-Laure Valton, Rasmus O Bak, Charles H Li, Johanna Goldmann, Bryan R Lajoie, Zi Peng Fan, Alla A Sigova, Jessica

Reddy, Diego Borges-Rivera, Tong Ihn Lee, Rudolf Jaenisch, Matthew H Porteus, Job Dekker, and Richard A Young. Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science*, 351(6280):1454–1458, March 2016.

[31] Jonas Ibn-Salem. *Genome folding in evolution and disease*. PhD thesis, Universitätsbibliothek Mainz, 2018.

[32] Maxim Imakaev, Geoffrey Fudenberg, Rachel Patton McCord, Natalia Naumova, Anton Goloborodko, Bryan R Lajoie, Job Dekker, and Leonid A Mirny. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat. Methods*, 9(10):999–1003, October 2012.

[33] Daniel Jost, Cédric Vaillant, and Peter Meister. Coupling 1D modifications and 3D nuclear organization: data, models and function. *Curr. Opin. Cell Biol.*, 44:20–27, 2017.

[34] Philip A Knight and Daniel Ruiz. A fast algorithm for matrix balancing. *IMA J. Numer. Anal.*, 33(3):1029–1047, July 2013.

[35] Galih Kunarso, Na-Yu Chia, Justin Jeyakani, Catalina Hwang, Xinyi Lu, Yun-Shen Chan, Huck-Hui Ng, and Guillaume Bourque. Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat. Genet.*, 42(7):631–634, 6 June 2010.

[36] Erez Lieberman-Aiden, Nynke L van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R Lajoie, Peter J Sabo, Michael O Dorschner, Richard Sandstrom, Bradley Bernstein, M A Bender, Mark Groudine, Andreas Gnirke, John Stamatoyannopoulos, Leonid A Mirny, Eric S Lander, and Job Dekker. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–293, 9 October 2009.

[37] Yin C Lin, Christopher Benner, Robert Mansson, Sven Heinz, Kazuko Miyazaki, Masaki Miyazaki, Vivek Chandra, Claudia Bossen, Christopher K Glass, and Cornelis Murre. Global changes in the nuclear positioning of genes and intra- and interdomain genomic interactions that orchestrate B cell fate. *Nat. Immunol.*, 13(12):1196–1204, December 2012.

[38] Tony Lindeberg. *Scale-Space Theory in Computer Vision*. Springer, Boston, MA, 1994.

[39] Xiong Liu, Xueping Yu, Donald J Zack, Heng Zhu, and Jiang Qian. TiGER: a database for tissue-specific gene expression and regulation. *BMC Bioinformatics*, 9:271, June 2008.

[40] D G Lowe. Object recognition from local scale-invariant features. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 2, pages 1150–1157 vol.2, September 1999.

[41] David G Lowe. Distinctive image features from Scale-Invariant keypoints. *Int. J. Comput. Vis.*, 60(2):91–110, November 2004.

[42] Darío G Lupiáñez, Malte Spielmann, and Stefan Mundlos. Breaking TADs: How alterations of chromatin domains result in disease. *Trends Genet.*, 32(4):225–237, 1 April 2016.

[43] Yiqin Ma, Kiriaki Kanakousaki, and Laura Buttitta. How the cell cycle impacts chromatin architecture and influences cell fate. *Front. Genet.*, 6:19, 3 February 2015.

[44] Krystian Mikolajczyk. *Detection of local features invariant to affine transformations: application to matching and recognition.* PhD thesis, Grenoble INPG, 2002.

[45] Maxwell R Mumbach, Ansuman T Satpathy, Evan A Boyle, Chao Dai, Benjamin G Gowen, Seung Woo Cho, Michelle L Nguyen, Adam J Rubin, Jeffrey M Granja, Katelynn R Kazane, Yuning Wei, Trieu Nguyen, Peyton G Greenside, M Ryan Corces, Josh Tycko, Dimitre R Simeonov, Nabeela Suliman, Rui Li, Jin Xu, Ryan A Flynn, Anshul Kundaje, Paul A Khavari, Alexander Marson, Jacob E Corn, Thomas Quertermous, William J Greenleaf, and Howard Y Chang. Enhancer connectome in primary human cells identifies target genes of disease-associated DNA elements. *Nat. Genet.*, 49(11):1602–1612, November 2017.

[46] Justin M O'Sullivan, Michael D Hendy, Tatyana Pichugina, Graeme C Wake, and Jörg Langowski. The statistical-mechanics of chromosome conformation capture. *Nucleus*, 4(5):390–398, September 2013.

[47] Bhavita Patel, Changwang Deng, Michael Litt, Mariana St. Just Riberio, Kairong Cui, Yuanyuan Kang, Yi Qiu, Keji Zhao, and Suming Huang. CTCF mediated enhancer and promoter interaction regulates differential expression of TAL1 oncogene in normal and malignant hematopoiesis. *Blood*, 120(21):281–281, November 2012.

[48] T Pederson. Chromatin structure and the cell cycle. *Proc. Natl. Acad. Sci. U. S. A.*, 69(8):2224–2228, August 1972.

[49] Douglas H Phanstiel, Alan P Boyle, Nastaran Heidari, and Michael P Snyder. Mango: a bias-correcting ChIA-PET analysis pipeline. *Bioinformatics*, 31(19):3092–3098, October 2015.

[50] Suhas S P Rao, Miriam H Huntley, Neva C Durand, Elena K Stamenova, Ivan D Bochkov, James T Robinson, Adrian L Sanborn, Ido Machol, Arina D Omer, Eric S Lander, and Erez Lieberman Aiden. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159(7):1665–1680, 18 December 2014.

[51] Abbas Roayaei Ardakany and Stefano Lonardi. Efficient and accurate detection of topologically associating domains from contact maps. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik GmbH, Wadern/Saarbruecken, Germany, 2017.

[52] Gil Ron, Yuval Globerson, Dror Moran, and Tommy Kaplan. Promoter-enhancer interactions identified from Hi-C data using probabilistic models and hierarchical topological domains. *Nat. Commun.*, 8(1):2237, December 2017.

[53] Anthony D Schmitt, Ming Hu, Inkyung Jung, Zheng Xu, Yunjiang Qiu, Catherine L Tan, Yun Li, Shin Lin, Yiing Lin, Cathy L Barr, and Bing Ren. A compendium of chromatin contact maps reveals spatially active regions in the human genome. *Cell Rep.*, 17(8):2042–2059, November 2016.

[54] Tom Sexton, Eitan Yaffe, Ephraim Kenigsberg, Frédéric Bantignies, Benjamin Leblanc, Michael Hoichman, Hugues Parrinello, Amos Tanay, and Giacomo Cavalli. Three-dimensional folding and functional organization principles of the drosophila genome. *Cell*, 148(3):458–472, 3 February 2012.

[55] E Shechtman and M Irani. Matching local Self-Similarities across images and videos. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, June 2007.

[56] Yin Shen, Feng Yue, David F McCleary, Zhen Ye, Lee Edsall, Samantha Kuan, Ulrich Wagner, Jesse Dixon, Leonard Lee, Victor V Lobanenkov, and Bing Ren. A map of the cis-regulatory sequences in the mouse genome. *Nature*, 488(7409):116–120, 2 August 2012.

[57] Hanjun Shin, Yi Shi, Chao Dai, Harianto Tjong, Ke Gong, Frank Alber, and Xianghong Jasmine Zhou. TopDom: an efficient and deterministic method for identifying topological domains in genomes. *Nucleic Acids Res.*, 44(7):e70, 20 April 2016.

[58] Marieke Simonis, Petra Klous, Erik Splinter, Yuri Moshkin, Rob Willemsen, Elzo de Wit, Bas van Steensel, and Wouter de Laat. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4c). *Nat. Genet.*, 38(11):1348–1354, November 2006.

[59] John Stansfield and Mikhail G. Dozmorov. Hiccompare: a method for joint normalization of hi-c datasets and differential chromatin interaction detection. *bioRxiv*, 2017.

[60] Zhonghui Tang, Oscar Junhong Luo, Xingwang Li, Meizhen Zheng, Jacqueline Jufen Zhu, Przemyslaw Szalaj, Pawel Trzaskoma, Adriana Magalska, Jakub Wlodarczyk, Blazej Ruszczycki, Paul Michalski, Emaly Piecuch, Ping Wang, Danjuan Wang, Simon Zhongyuan Tian, May Penrad-Mobayed, Laurent M Sachs, Xiaoan Ruan, Chia-Lin Wei, Edison T Liu, Grzegorz M Wilczynski, Dariusz Plewczynski, Guoliang Li, and Yijun Ruan. CTCF-Mediated human 3D genome architecture reveals chromatin topology for transcription. *Cell*, 163(7):1611–1627, December 2015.

[61] Oana Ursu, Nathan Boley, Maryna Taranova, Y X Rachel Wang, Galip Gurkan Yardimci, William Stafford Noble, and Anshul Kundaje. GenomeDISCO: a concordance score for chromosome conformation capture experiments using random walks on contact map graphs. *Bioinformatics*, 34(16):2701–2707, August 2018.

[62] P Viola and M Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pages I–511–I–518 vol.1, 2001.

[63] Junbai Wang, Xun Lan, Pei-Yin Hsu, Hang-Kai Hsu, Kun Huang, Jeffrey Parvin, Tim H-M Huang, and Victor X Jin. Genome-wide analysis uncovers high frequency, strong differential chromosomal interactions and their associated epigenetic patterns in e2-mediated gene regulation. *BMC Genomics*, 14:70, January 2013.

[64] Eitan Yaffe and Amos Tanay. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat. Genet.*, 43(11):1059–1065, October 2011.

[65] Tao Yang, Feipeng Zhang, Galip Gürkan Yardımcı, Fan Song, Ross C Hardison, William Stafford Noble, Feng Yue, and Qunhua Li. HiCRep: assessing the reproducibility of Hi-C data using a stratum-adjusted correlation coefficient. *Genome Res.*, 27(11):1939–1949, November 2017.

[66] Galip Gürkan Yardımcı, Hakan Ozadam, Michael E G Sauria, Oana Ursu, Koon-Kiu Yan, Tao Yang, Abhijit Chakraborty, Arya Kaul, Bryan R Lajoie, Fan Song, Ye Zhan, Ferhat Ay, Mark Gerstein, Anshul Kundaje, Qunhua Li, James Taylor, Feng Yue, Job Dekker, and William S Noble. Measuring the reproducibility and quality of Hi-C data. *bioRxiv*, page 188755, February 2018.

[67] Hui Zheng and Wei Xie. The role of 3D genome organization in development and cell differentiation. *Nat. Rev. Mol. Cell Biol.*, 20(9):535–550, September 2019.

[68] Xiaobei Zhou, Helen Lindsay, and Mark D. Robinson. Robustly detecting differential expression in rna sequencing data using observation weights. *Nucleic Acids Research*, 42(11):e91, 2014.

[69] Marie Zufferey, Daniele Tavernari, Elisa Oricchio, and Giovanni Ciriello. Comparison of computational methods for the identification of topologically associating domains. *Genome Biol.*, 19(1):217, December 2018.

[70] Marie Zufferey, Daniele Tavernari, Elisa Oricchio, and Giovanni Ciriello. Comparison of computational methods for the identification of topologically associating domains. *Genome Biology*, 19(1):217, Dec 2018.