UNIVERSITY OF CALIFORNIA
RIVERSIDE

Statistical Differential Analyses of Hi-C Contact Maps

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

Doctor of Philosophy

in

Applied Statistics

by

Huiling Liu

June 2021

Dissertation Committee:

Dr. Wenxiu Ma, Chairperson
Dr. Xinping Cui
Dr. Weixin Yao

Copyright by
Huiling Liu
2021

The Dissertation of Huiling Liu is approved:

Committee Chairperson

University of California, Riverside

Acknowledgments

Foremost, I would like to express my deep and sincere gratitude to my advisor Dr. Wenxiu Ma, for her continuous guidance and support of my Ph.D study and research. Her enthusiasm, sincerity, patience and profession have deeply impressed me. She helped me come up with the research topics about differential analyses on Hi-C data and shared immense knowledge and valuable experience throughout my work and the writing of this dissertation. It was a great honor to study and work under her guidance.

Besides my advisor, my sincere thank goes to the rest of my dissertation committee: Dr. Xinping Cui and Dr. Weixin Yao, for their constructive ideas and insightful comments, as well as their encouragement throughout my studies.

In addition, I also would like to thank all the members in Dr. Ma's lab: Yangyang Hu, Tiantian Ye, Luke Klein, Jinli Zhang, Li Ma and Biswanath Chowdhury, for their helpful feedback and suggestions.

Last but not least, I am extremely grateful to my family: my parents, for giving birth to me and sacrificing for raising and educating me for my future; and my husband, for the love and understanding throughout my PhD study and my life in general.

To my parents and husband for all the support.

ABSTRACT OF THE DISSERTATION

Statistical Differential Analyses of Hi-C Contact Maps

by

Huiling Liu

Doctor of Philosophy, Graduate Program in Applied Statistics
University of California, Riverside, June 2021
Dr. Wenxiu Ma, Chairperson

Recent advances in Hi-C techniques have allowed us to map genome-wide chromatin interactions and uncover higher-order chromatin structures, thereby shedding light on the principles of genome architecture and functions. However, statistical methods for detecting changes in chromatin organizations are still in the early stage. In this dissertation, we proposed two statistical methods, namely DiffGR and scHiCDiff, for differential analysis in Hi-C contact maps.

The first method DiffGR detects differentially interacting genomic regions at the scale of topologically-associating domains (TADs) between two Hi-C contact maps. Specifically, we utilized the stratum-adjusted correlation coefficient (SCC) to measure similarity of local TAD regions. We then developed a non-parametric approach to identify statistically significant changes of genomic interacting regions. Through simulation studies, we demonstrated that DiffGR can robustly and effectively discover differential genomic regions under various conditions. Furthermore, we successfully revealed cell type-specific changes in genomic interacting regions using real Hi-C datasets.

The second method scHiCDiff focuses on detecting differential chromatin interactions (DCIs) in single-cell Hi-C data. The three-dimensional genome organization constructed from the conventional bulk Hi-C protocol represents an ensemble based on thousands to millions of nuclei, but not the actual genome organizations in individual cells. Unlike bulk Hi-C, single-cell Hi-C enables the exploration cell-specific chromosomal structures. However, interpretation and analysis of single-cell Hi-C data is at very early stage. To characterize the significant changes between different cells at the single-cell level, we built scHiCDiff which applied non-parametric tests and parametric models in distinguishing DCIs from single-cell Hi-C data. Our evaluation proved that these methods, especially the zero-inflated negative binomial (ZINB) and negative binomial hurdle(NBH) models, can effectively detect reliable and consistent DCIs of single cells between different conditions, which better capture cell type-specific variations of chromosomal structures.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Recent developments of chromatin conformation capture (3C)-based techniques—including 4C [61], 5C [18], Hi-C [41, 19, 32], ChIA-PET [39], and Hi-ChIP [49]—have allowed high-throughput characterization of pairwise chromatin interactions in the cell nucleus, and provided an unprecedented opportunity to investigate the higher-order chromatin structures and to elucidate their roles in nuclear organization and gene expression regulation. Among these techniques, Hi-C and its variants [46, 56, 53] are of particular interest because of their ability to map chromatin interactions at a genome-wide scale.

The main steps in a typical Hi-C protocol [45] are as follows: (1) Samples of nuclear DNA are cross-linked; (2) These chromatins are digested with a restriction enzyme; (3) The resulting sticky ends of two restriction fragments are filled with biotin-labeled nucleotides; (4) These sticky ends are subjected to proximity ligation; (5) After shearing the loop, only chimeric fragments with biotin label will be pulled down; and (6) These DNA fragments are sent to high-throughput paired-end sequencing (Figure 1.1A).

For the pre-processing of the Hi-C sequencing data, the genome can be partitioned into a sequence of continuous non-overlapping equal-size bins. For instance, when using a bin size of 1 mega-base (Mb), each bin represents a chromatin fragment of length 1 Mb. Then a symmetric chromatin interaction frequency matrix (Y) with certain resolution (bin size) can be constructed and visualized as a heatmap (Figure 1.1B). Typically, the contact count Y_{ij} is measured by the number of read pairs between the bin pair i and j in the cells, and therefore denotes as the interaction frequency between these two chromatin fragments i and j. Thus, larger count number $Y_{i,j}$ indicates higher contact frequency between chromatin regions i and j and therefore implies the closer spatial proximity between these two loci.

Because of the complex experimental steps in Hi-C protocol, various sources of biases (driven from the Hi-C techniques and DNA sequencing platforms) can be introduced in Hi-C raw data, which make the data analysis challenging. Hence, several normalization methods were developed to reduce biases in the data. Explicit-factor correction algorithms (e.g. HiCNorm [30]) corrected known systematic bias in terms of fragments length, GC content and mappability while matrix-balancing methods (e.g. ICE [31] and KR [34]) assumed uniform visibility for all genomic loci and assured equal row and column sums for correcting both known and unknown biases.

A particularly important characteristic of Hi-C contact matrices is the presence of the topologically-associating domains (TADs), which are functional units of chromatin with higher tendency of intra-domain interactions dixon2012topological. TADs are largely conserved across cell types and species[16, 59]. Moreover, CTCF and other chromatin binding proteins are enriched at the TAD boundaries, indicating that TAD boundary

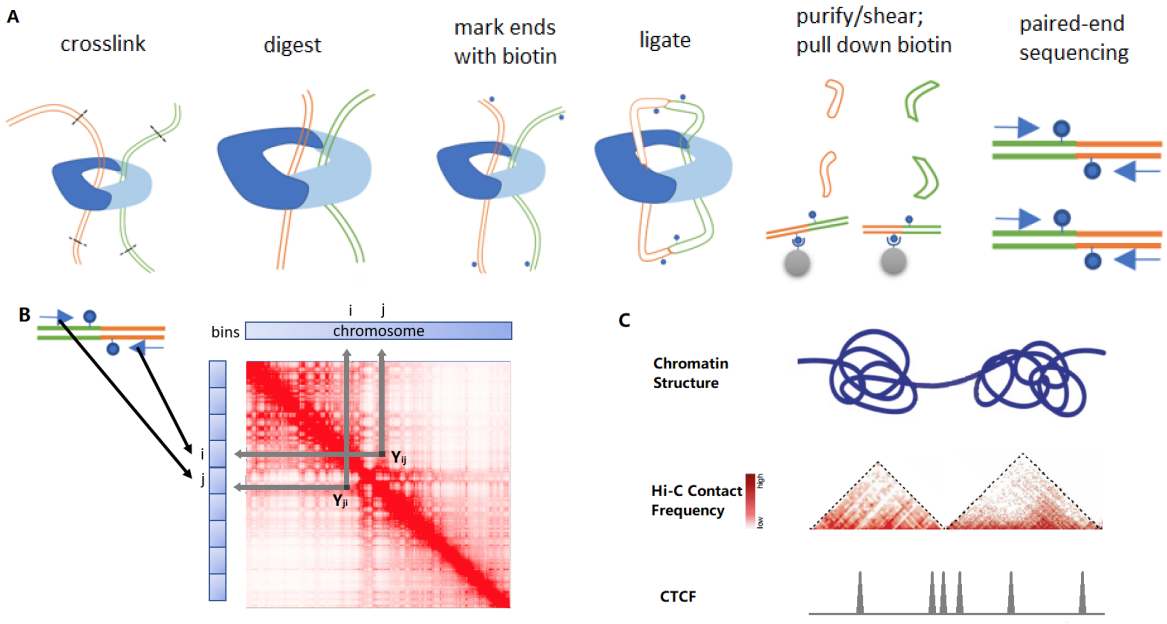

Figure 1.1: **Illustration of Hi-C Features.** (A). Main steps in Hi-C protocol: DNA is cross-linked and cleaved by a restriction enzyme. The sticky ends of restriction fragments are filled in with biotin-labelled nucleotides. Ligate ends of the fragments to form a loop. After shearing the loop, biotin-labelled ligation products are pulled down for paired-end sequencing. (B).Visualization of Hi-C contact matrix: Interaction frequency of each bin pair is visualized by the gradation of color. The higher the interaction contact, the darker the color. (C). Diagram of chromatin architecture represents a chromatin fiber (blue) spanning two adjacent TADs. Interaction frequency visualizes HiC contact counts of these two adjacent TADs. It is shown that the topological boundary region form chromatin loops and are enriched with CTCF.

regions form chromatin loops and play an essential role in gene expression regulation dixon2012topological, fudenberg2016formation(Figure 1.1C).

Several computational methods have been developed to detect TADs in Hi-C contact maps. These methods can be categorized into two groups: one-dimensional (1D) statistic-based methods and two-dimensional (2D) contact matrix-based methods [5]. Of these, 1D statistic-based methods often take a sliding window approach along the diagonal of Hi-C contact matrix and compute a 1D statistic for each bin to detect TADs and/or TAD boundaries. For instance, Dixon et al. 2012 [16] introduced a statistic named directionality index (DI) to quantify whether a genomic locus preferentially interacts with upstream or downstream loci and developed a hidden Markov model to call TADs from DIs. Later, Crane et al. 2015 [10] proposed a novel TAD detection method, which computes an insulation score (IS) for each genomic bin by aggregating chromatin interactions within a square sliding through the diagonal and then searches for the minima along the IS profile as TAD boundaries. Unlike the 1D statistic-based methods which calculate statistics using local information, the 2D contact matrix-based methods utilize global information on the contact matrix to capture TAD structures. For example, the Armatus algorithm [21] identifies consistent TAD patterns across different resolutions by maximizing a quality scoring function of domain partition using dynamic programming. In addition, Levy et al. 2014 [38] proposed a TAD boundary detection method named HiCseg, which performs a 2D blockwise segmentation via a maximum likelihood approach to partition each chromosome into its constituent TADs. Recently, several review papers have quantitatively compared the performances of the aforementioned TAD-calling methods and demonstrated that HiCseg

detects a stable number of TADs against changes of sequencing coverage and maintains the highest reproducibility among Hi-C replicates across all resolutions when compared with other TAD-calling methods [24, 28].

With the fast increasing of Hi-C datasets, a key statistical challenge in 3D chromatin analyses is to assign reliable statistical measures to compare differences in chromatin structures. The differences may arise from different developmental stages, cell lines, disease states or individuals. Further, chromatin structures also exhibit multi-scale differences among different Hi-C maps in the entire chromosomes, compartments, TADs, and chromatin loops and interactions. Surveying the literature, we noticed that several computational tools (see section 2.1) have been developed for comparative Hi-C analysis, but the majority of them focused on the similarity comparison between Hi-C maps and the detection of differential chromatin interactions (DCIs). As TADs are strongly linked to cell type-specific gene expression [16], appropriate statistical methods for detecting differentially interacting regions at the TAD level are in high demand.

In the dissertation, we firstly introduced a novel statistical method, DiffGR, for detecting differentially interacting genomic regions at TAD level between two Hi-C contact matrices. Briefly, DiffGR utilizes the stratum-adjusted correlation coefficient (SCC) instead of the standard Pearson correlation coefficient (Pearson CC) to measure the similarity of local genomic regions between two Hi-C contact maps and then applies a non-parametric permutation test on those SCC values to assess the statistical significance of differences in local genomic regions. We demonstrated that DiffGR can effectively and robustly identify

differentially interacting genomic regions at TAD level in both simulated data and real Hi-C data from different cell types.

The conventional Hi-C studies discovered some commonly shared structural features in 3D genome organization, but it is worth noting that, because of the dynamic nature of the chromatin fiber and the variability of nuclear processes among single cells, the geometry of genome organization varies from cells to cells [22]. Actually, the conventional Hi-C (named ensemble Hi-C/ bulk Hi-C) contact maps are an ensemble based on millions of cells and the 3D genome organization constructed from them denotes the average but not the actual genome organizations of individual cells.

In recent years, several single-cell Hi-C (scHi-C) experiments were developed to capture the single-cell Hi-C contact counts for thousands of cells simultaneously [50, 54, 23]. These scHi-C profiles provide promising opportunities to investigate cell-specific genomic structures, but the research of single-cell Hi-C data is still at very early stage. Current scHi-C studies focus more on clustering cells into constituent cell types. For instance, reproducibility methods [58, 67, 70, 71] coupled with multidimensional scaling (MDS) were applied to scHi-C data to evaluate similarity among single cells; among them, HiCRep with MDS yielded reasonable embedding of the single cells [42]. Later, scHiCluster [74] was developed as a single-cell clustering algorithm based on imputations using linear convolution and random walk and scHiCTools [40] implemented the clustering by a novel InnerProduct approach.

However, few of the aforementioned methods pay attention to studying variations in chromosome structures between different biological conditions (cell types or cell states),

especially on identifying significant changes in the interaction intensity (i.e., differential interactions) of Hi-C maps between two or more biological conditions at single-cell level. To fill in the blanks on singe-cell Hi-C differential interaction detection, in this dissertation, we later implemented a new tool named scHiCDiff, which applied two non-parametric tests (Cramer-von Mises test and Kolmogorov–Smirnov test) and three parametric models (Negative Binomial [NB], zero-inflated Negative Binomial[ZINB], and Negative Binomial Hurdle [NBH]) to distinguish the bin pairs with significant changes in counts. Specifically, ZINB and NBH regression models took the sparsity effect in sciHi-C data into consideration and performed a rigorous statistical test (likelihood ratio test: Chi-square test). We demonstrated, through simulation studies and real data analysis, scHiCDiff can effectively reveal differentially interaction bin pairs between different conditions.

Chapter 2

DiffGR: Detecting Differentially Interacting Genomic Regions From Hi-C Contact Maps

2.1 Introduction

With the fast accumulation of Hi-C datasets, there has been a growing interest in performing differential analysis of Hi-C contact matrices. To date, several computational tools have been developed for comparative Hi-C analysis, but the majority of them focused on the similarity comparison between Hi-C maps and the identification of differential chromatin interactions (DCIs), which represent different chromatin looping events between two Hi-C contact maps. In reproducibility assessment, HiCRep[71] smoothed Hi-C contact matrix with 2D mean filter and measured the similarity of two Hi-C contact matri-

ces by stratum-adjusted correlation coefficient(SCC) according to genomic distance, while QuASAR[58] computed the interaction correlation matrix weighted by interaction enrichment. HiC-Spector[70] transformed the Hi-C matrices to a Laplacian matrix and summarized the Laplacian by matrix decomposition. Later, GenomeDISCO[67] utilized random walks on the network defined by Hi-C map to smooth data before calculating similarity. As to the detection of DCIs, in early studies, the most common strategy was to use the fold change values between two Hi-C contact maps. For instance, Wang et al.2013 [68] used a simple fold-change strategy to detect the influence of estrogen treatment on chromatin interactions in MCF-7 Hi-C samples. Additionally, Dixon et al.2015 [15] utilized the fold change values of chromatin interactions to train a random forest model to discover the epigenetic signals that were more predictive of changes in interaction frequencies. In addition to these fold change-based approaches, another commonly utilized method for detecting DCIs was the binomial model implemented by the HOMER software [29]. In contrast, in more recent studies, count-based statistical methods, such as edgeR [57] and DESeq [44], have been adopted to identify pairwise chromatin interactions that show significant changes in contact frequencies. Among them, Lun et al.2015[45] presented a tool named diffHic for rigorous detection of differential interactions by leveraging the generalized linear model (negative binomial regression) of edgeR, and demonstrated that edgeR outperformed the binomial model. Later, Stansfield et al.2018[64] introduced MD normalization and performed a Z-test to detect statistically significant DCIs. While all these methods assumed independence among pairwise interactions, which holds true only in coarse-resolution Hi-C maps, Djekidel et al.2018[17] presented a novel method, named FIND, that takes into account the

dependency of adjacent loci at finer resolutions. Briefly, FIND utilizes a spatial Poisson process model to detect DCIs that show significant changes in interaction frequencies of both themselves and their neighborhood bins. Lastly, Cook et al. 2020[7] introduced ACCOST to identify differential chromatin contacts by extending the DESeq model used in RNA-seq analysis and repurposing the "size factor" to account for the notable genomic-distance effect in Hi-C contact matrices.

In the cell nucleus, chromatin is organized at multiple levels, ranging from active and inactive chromosomal compartments and sub-compartments (on a multi-Mb scale) [41, 56], to TADs (0.5–2 Mb on average) [16] and fine-scale chromatin interacting loops [56, 46]. Chromatin structures also exhibit multi-scale differences among different cell types in their compartments, TADs, and chromatin loops. Among these, changes in TAD organizations are of particular interest as TADs are strongly linked to cell type-specific gene expression [16]. For example, Taberlay et al.2016[65] have shown that genomic rearrangements in cancer cells are partly guided by changes in higher-order chromatin structures, such as TADs. They discovered that some large TADs in normal cells are further segmented into several smaller TADs in cancer cells, and these changes are tightly correlated with oncogene expression levels. Current differential analyses of TAD structures between different cell types and conditions are limited to the detection of TAD boundary changes. Recently, Chen et al. 2018 [5] proposed a TAD boundary detection approach named HiCDB, which is constructed based on local measures of relative insulation and multi-scale aggregation. In addition to calling TAD boundaries in single Hi-C sample, HiCDB also provides differential TAD boundary detection using the average values of relative insulation across multiple

samples. Later, Cressswell et al.2020 [11] developed TADCompare, which uses a spectral clustering-derived metric named eigenvector gap to identify differential and consensus TAD boundaries and track TAD boundary changes over time. The HiCDB and TADCompare methods focused on detecting changes in TAD boundaries rather than changes in chromatin organization within TADs. However, differential TAD boundaries do not necessarily indicate differential chromatin conformation within those regions. First, Hi-C contact matrices are often sparse and noisy, which might lead to unstable detection of TAD boundaries. Second, chromatin interactions within a TAD could be strengthened or weakened in another Hi-C sample, which would suggest different patterns of chromatin organization within the same TAD region. Therefore, appropriate statistical methods for detecting differentially interacting regions at the TAD level are in demand.

To tackle this problem, we developed a novel statistical method, DiffGR, for detecting differential genomic regions at TAD level between two Hi-C contact maps. Briefly, DiffGR utilizes the stratum-adjusted correlation coefficient (SCC), which can effectively eliminate the genomic-distance effect in Hi-C data, to measure the similarity of local genomic regions between two contact matrices, and then applies a non-parametric permutation test on those SCC values to detect genomic regions with statistically significant differential interactions. We demonstrated, through simulation studies and real data analysis, that DiffGR can effectively and robustly identify differentially interacting genomic regions at TAD level.

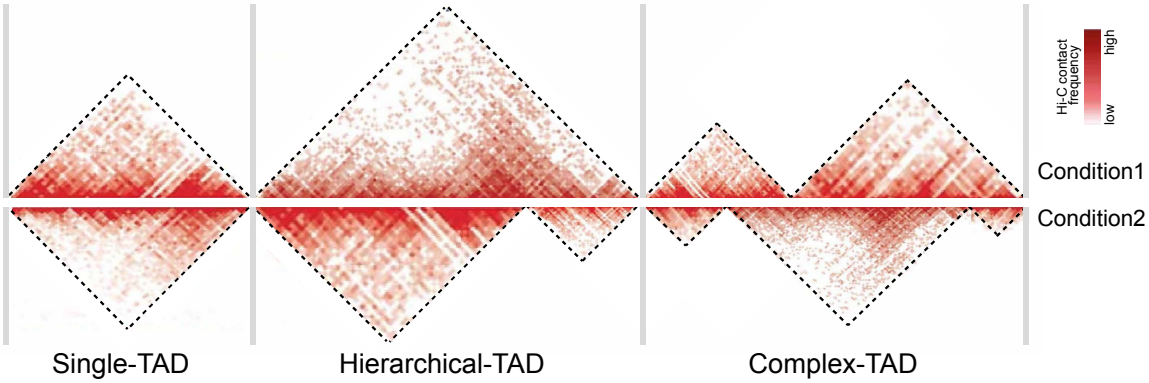

Figure 2.1: **Illustration of three candidate types of differential genomic regions.**

2.2 Methods

The DiffGR method detects differentially interacting genomic regions in three steps, as described below (Sections 2.2.1-2.2.3). In addition, the simulation settings are outlined in Section 2.2.4 and real data pre-processing and analyses are described in Section 2.2.5.

2.2.1 Identifying candidate genomic regions

Suppose we have two sets of Hi-C data and their corresponding contact frequency matrices as the input. First, we detect the TAD boundaries in each Hi-C data, separately. Specifically, we apply HiCseg [38] to the raw contact matrices and obtain the corresponding TAD boundaries. Note that in this step one can exchange HiCseg with another TAD caller, whose detected TADs satisfy the non-overlapping and continuous properties. We choose HiCseg because it has been shown that HiCseg produces more robust TAD boundaries than other TAD-calling methods [24]. We next combine the TAD boundaries from both

Hi-C contact maps to identify the candidate genomic regions for subsequent analyses. TAD boundaries within three-bin distance are considered to be a common boundary shared by both Hi-C datasets and replaced by the middle bin locus. We then partition the genome into non-overlapping candidate regions using the common TAD boundaries, and categorize these candidate regions into the following three groups: (1) single-TAD candidate regions, (2) hierarchical-TAD candidate regions, and (3) complex-TAD candidate regions, as illustrated in Figure 2.1.

2.2.2 Measuring similarity of candidate regions between two Hi-C contact maps

In the second step, we evaluate the candidate regions one at a time. Suppose a candidate genomic region is bounded by two common TAD boundaries shared by both Hi-C maps, and contains m unique TAD boundaries in either one of the two Hi-C maps. In the single-TAD candidate region, $m = 0$; in the hierarchical-TAD or complex-TAD candidate regions, $m >= 1$. For each candidate region, we consider all $\binom{m+2}{2}$ possible (sub)TADs, which are separated by any pair of TAD boundaries within that region, as potential differential TADs. For each potential differential TAD, we calculate the stratum-adjusted correlation coefficient (SCC) [71] to measure the similarity of intra-TAD chromatin interactions between two Hi-C samples.

The SCC metric was introduced by Yang et. al. 2017 [71] as a measure of similarity and reproducibility between two Hi-C contact matrices. To account for the pronounced distance-dependence effect in Hi-C contact maps, chromatin contacts are first stratified into K stratum according to the genomic distances of the contacting loci pairs, and the

correlation coefficients of contacts within each stratum are calculated between two samples. These stratum-specific correlation coefficients are then aggregated to compute the SCC value using a weighted average approach, where the weights are derived from the Cochran-Mantel-Haenszel (CMH) statistic [1]. That is, the SCC ρ is calculated as

$$\rho = \sum_{k=1}^{K} \left(\frac{N_k r_{2k}}{\sum_{k=1}^{K} N_k r_{2k}} \right) \rho_k,$$

where N_k is the number of elements in the k-th stratum, r_{2k} is the product of standard deviations of the elements in the k-th stratum of both samples, and ρ_k denotes the correlation coefficient of the k-th stratum between two samples.

The original SCC metric is computed using the intra-chromosomal contact matrices with a predefined genomic distance limit. The resulting value has a range of $[-1, 1]$ and can be interpreted in a way similar to the standard correlation coefficient. Here we use SCC as a local similarity measurement to evaluate each potential differential TAD between two Hi-C samples. In the SCC calculation, an upper limit of genomic distance is set to 10 Mb because TADs are commonly smaller than 10 Mb and distal interactions over a genomic distance larger than 10 Mb are often sparse and highly stochastic. In addition, as the sparsity of Hi-C matrices might affect the precision of SCC values, the loci pairs with zero contact frequencies in both samples are excluded from the calculation.

Hi-C contact maps are often sparse due to sequencing coverage limits and contain various systematic biases. To solve these issues, when pre-processing the Hi-C contact matrices, we first smooth each contact map by a 2D mean filter [71], which substitutes the contact count observed between each bin pair by the average of all contact counts in its neighborhood. This smoothing process improves the contiguity of the TAD regions with

elevated contact frequencies, thereby enhancing the domain structures. Next, we utilize the Knight-Ruiz (KR) normalization [34] on the smoothed matrices to remove potential biases.

2.2.3 Detecting statistically significant differential regions

In the third step, we identify differential genomic regions by first finding differential TADs within these candidate regions. In each candidate genomic region, we calculate the SCC values for all potential differential TADs as described above. Then we develop a non-parametric permutation test to estimate the p-values for these local SCC values (Section 2.2.3). Additionally, we propose a quantile regression strategy to speed up the permutation test (Section 2.2.3). Finally, we consider a candidate region to be a differentially interacting genomic region, if at least one TAD within that region exhibits a statistically significant difference between the two samples and the size of the largest differential TAD meeting this criterion is greater than one third of the length of the entire candidate region.

Permutation test to compute p-values of local SCCs

Since the local SCC values are calculated for all potential differential TADs of various sizes, we perform the following non-parametric permutation test for each unique TAD size.

Suppose s is a potential differential TAD whose length is l_s and SCC value between two Hi-C samples is ρ_s. To assess the statistical significance of the observed SCC value ρ_s, the null distribution of SCC values for TADs of the same size is estimated via the following permutation procedure. To generate a random TAD with length l_s, we first randomly select l_s positions from main diagonal of Hi-C contact matrix, then $l_s - 1$ position from the 1st off-

diagonal, ..., and lastly 1 position from the $(l_s - 1)$-th off-diagonal. We subsequently extract contact counts of these randomly selected positions from the two Hi-C contact matrices to construct the permuted TAD pair and calculate its SCC value. We repeat the above random TAD generation step N times ($N = 2000$) and obtain the corresponding SCC values $\{\rho_i^{l_s}\}$, $i = 1, \cdots, N$. Then the p-value of the observed SCC value ρ_s can be computed as:

$$p_s = \frac{\sum_{i=1}^{N} I(\rho_i^{l_s} < \rho_s)}{N},$$

where $I(\cdot)$ is the indicator function. Lastly, we compare the p-values with a pre-defined significance level α (by default $\alpha = 0.05$) to determine differential TADs meeting the significance threshold. Note that the permutation framework accounts for the multiple testing correction using the Benjamini-Hochberg procedure [3].

The permutation framework always detect significant differential TADs even when two sample are very similar (like biological-replicates from same experiments). This is because the high similarity between biological replicates would cause their corresponding random TAD patterns tending to have rather high SCC values and then result in some non-differential TADs with relatively low SCC values falsely being detected as differential ones. In order to reduce the number of false positives, we provide the option to filter the final p-values $p_s.adj$ by an empirical or automatically calculated threshold. This option allows us to pre-specify the meaningful SCC between the two Hi-C datasets that must be reached in order to call a differential TAD truly significant.

$$p_s.adj = \begin{cases} 0.5 & if\ p_s < \alpha\ and\ \rho_s > \theta \\ p_s & otherwise \end{cases}$$

The threshold θ normally can be defined as 0.85 (A clear margin separated non-replicates from biological/pseudo replicates in the whole chromosome similarity comparison between multiple cell lines[72]) or can be calculated automatically as $\theta = \frac{\rho_{nr}^{ls} + \rho_{br}^{ls}}{2}$, where ρ_{nr}^{ls} represents the mean α quantile of SCCs between non-replicate data and ρ_{br}^{ls} is the average of α quantile of SCCs between their corresponding biological/pseudo-replicate data. Here, we call matrices from the same cell type as biological replicates, matrices from different cell lines (non-replicates) and matrices sampled from pooled biological replicates (pseudo-replicates).

Speed-up algorithm

In the previously described permutation test, we need to generate N random TAD pairs for each of the unique TAD sizes. However, such permutation procedure would be very time-consuming, especially for fine-resolution Hi-C datasets. To speed up the permutation process, we adopt a non-parametric regression approach to estimate the quantiles of the SCC values. As shown in Figure 2.2, we can clearly observe that there is a consistent pattern of the critical values (quantiles) of SCCs that exist for different quantiles and in different datasets. When the TAD size is relatively small, the quantile of SCC values increases dramatically with the TAD size; eventually when the TAD size is large, the quantile of SCC values levels off. One possible explanation of this observed pattern is that small-size TADs contain insufficient amount of information to produce reliable local SCC values. As a result, the SCCs of randomly generated small TAD pairs are often low, which would result in low quantile values. As the TAD size increases, sufficient interaction information is obtained

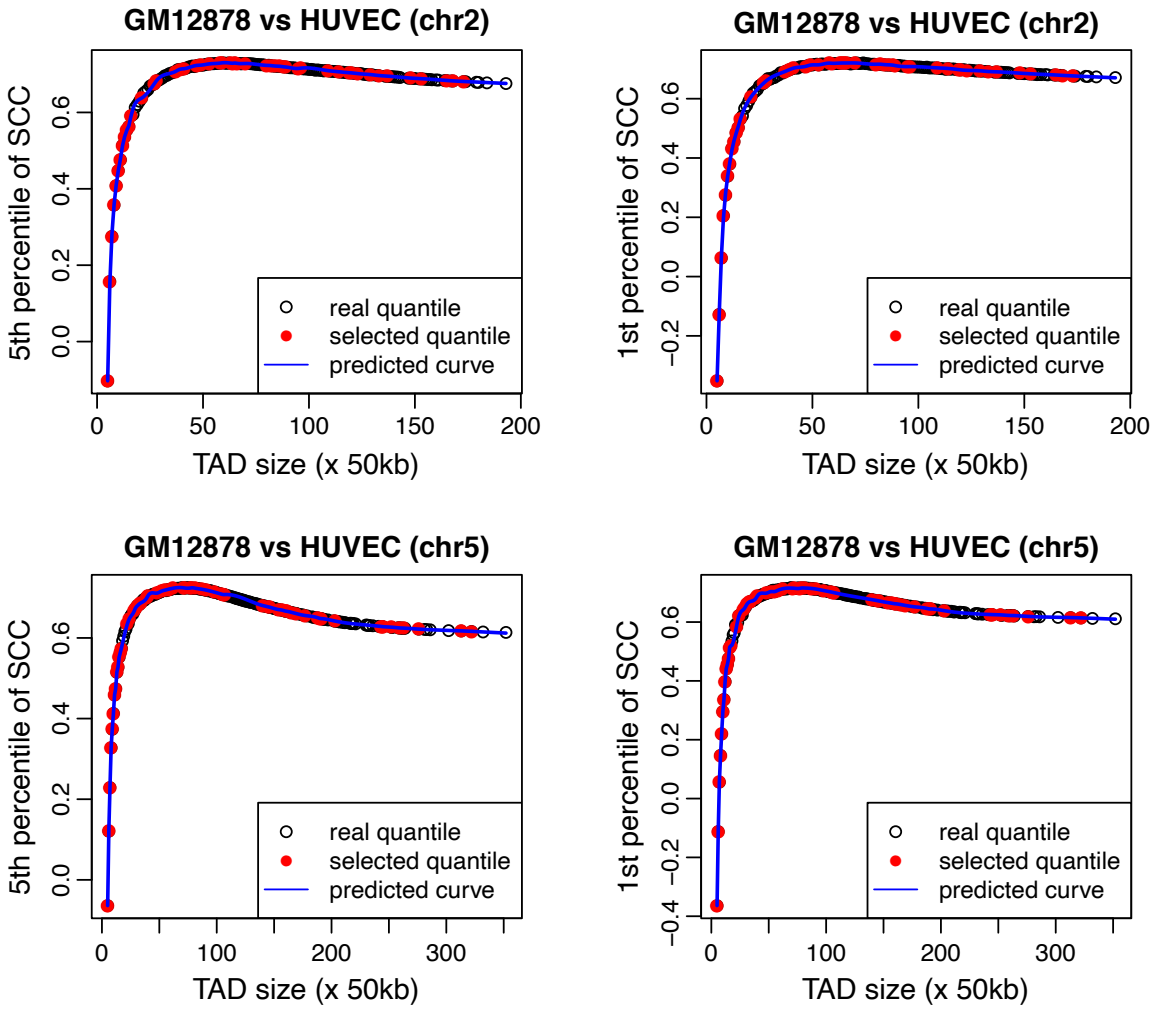

Figure 2.2: **Quantiles of local SCC values computed by permutation.** The x-axis indicates the TAD size (bin size = 50 kb), the y-axis is the corresponding 5th (or 1st) percentile of local SCC values computed from the comparison between GM12878 and HUVEC cells. The open black circles represent real quantile values; the red points denote the randomly selected points by the speed-up algorithm to fit the regression line; the blue line is the predicted quantile curve by a smooth spline.

from the data. Therefore, the corresponding SCC values would be stabilized leading to relatively reliable and steady quantiles.

To facilitate the permutation process, we adopt a non-parametric regression strategy to approximate the SCC quantiles (Figure 2.2). Specifically, instead of performing the permutation procedure for all unique TAD sizes, we randomly select a subset of m TAD sizes. For each selected TAD size, we generate N random TAD pairs, compute their local SCCs, and identify a series of quantiles (α_1-th percentile, α_2-th percentile, ..., α_J-th percentile) of the SCCs accordingly. Therefore, for a particular quantile (for example, the α_j-th percentile), we would have m quantile values; one for each of the selected TAD sizes. Based on these m data points, we fit a curve between the α_j-th percentile and the TAD size via a smoothing spline. Following this regression procedure, for any given TAD size, we can predict a series of quantiles of the SCCs which would be utilized to estimate the p-values as previously described in Section 2.2.3.

As to the selection of TAD sizes, we typically choose m to be 25% of the number of unique TAD sizes that are larger than 15 bins. In addition, we also include all TAD sizes from 1 to 15 bins to obtain a better fitting at the beginning of the quantile curve.

2.2.4 Simulation settings

To evaluate the performance of the DiffGR method, we conducted a series of simulation experiments by varying the proportion of altered TADs, proportion of TAD alternation, noise level, and sequencing coverage level. Specifically, we utilized the published chromosome 1 contact matrix of K562 cells at 50-kb resolution [56] as the original Hi-C data and simulated the altered Hi-C contact matrices as described below.

Single-TAD alternation

As TADs are conserved genomic patterns and TAD boundaries are relatively stable across cell types and even across species [16], our simulations primarily focused on the scenarios of single-TAD alternations. Suppose we had an original Hi-C contact matrix M and its identified TAD boundaries. Each of our simulated Hi-C matrices contained two components: the signal matrix S and the noise matrix N, with a certain signal-to-noise ratio.

First, to construct the signal matrix S, we randomly selected a subset of TADs from original contact matrix to serve as the true differential TADs. Then we replaced a certain portion of contact counts in each selected TAD by randomly sampling contact counts from the corresponding diagonals of the contact matrix. Second, we simulated the noise matrix N which represents the random ligation events in Hi-C experiments. Briefly, we generated these contacts by randomly choosing two bins, i and j, and adding one to the entry N_{ij} in the noise matrix. The probability of sampling each bin in the bin pair was set proportional to the marginal count of that bin in the original matrix. The sampling process was repeated C times, where C was the total number of contacts in the original Hi-C contact matrix M. The resulting random ligation noise matrix N contained the same number of contacts as the original contact matrix M.

To summarize, we had the following parameters in our single-TAD simulations.

- proportion of altered TADs. Using HiCseg, we detected 189 TADs with a mean size of 1.2 Mb in the original K562 chromosome 1 contact matrix (Figure 2.3). By default, we set the proportion of altered TADs to be 50%, which can vary from 20% to 70%.

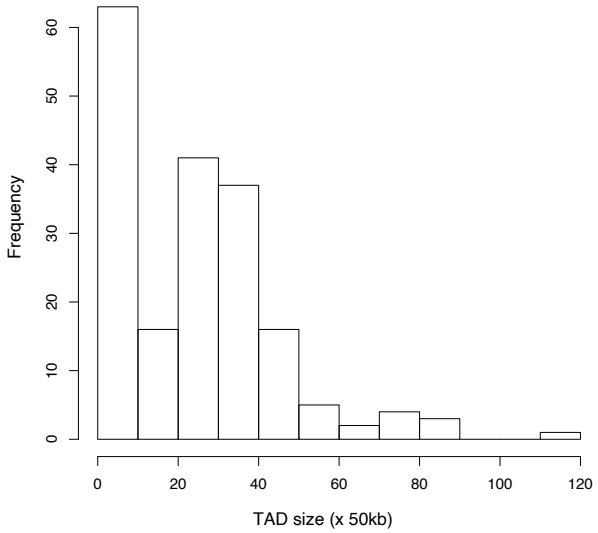

Figure 2.3: **Histogram of TAD size of all HiC-seg identified TADs on Chromosome 1 of K562 cells at 50-kb resolution.**

- proportion of TAD alternation. In the default setting, we substituted all contact counts in the selected TADs by random counts permuted from the matching diagonals in Hi-C maps. To reduce the degree of intra-TAD alternation, we gradually decreased the proportion of randomly substituted intra-TAD contacts from 100% to 10%.

- noise level, i.e., the ratio between the noise and signal matrices. The noise level was set to 10% by default, and varied from 1% to 80%.

For each simulation parameter setting, we generated 100 altered Hi-C contact matrices to compare against the original contact matrix. To evaluate the accuracy of the detection results, we used the false detection rate which defines as inaccurate percentage and is computed as $1 - Accuracy = \frac{\mathrm{FP+FN}}{\mathrm{N}}$, where FP denotes the falsely detected differential

regions, FN represents the the falsely detected non-differential regions, and N is the total number of candidate regions being tested.

Hierarchical-TAD alternation

To simulate the alternation pattern of hierarchical TADs, we randomly selected 50% of the large TADs whose size was greater than 10 bins in the signal matrix to serve as the true differential TADs. For each of the selected large TAD, we chose a random subTAD boundary to split it into two smaller subTADs (each with size > 5 bins). We then replaced all inter-subTAD contact counts by randomly sampled counts from Hi-C maps. Next, we validated the performance of DiffGR under the hierarchical-TAD condition with respect to different noise levels similar to the single-TAD simulations. Because the complex-TAD condition has complicated TAD boundaries between two samples and occurs less frequently in real data, we did not generate simulation data for this condition.

Simulating low-coverage contact matrices

Low sequencing depth of Hi-C experiments would lead to low-coverage and sparse contact matrices, thus it could potentially affect the performance of the detection of differentially interacting regions. To simulate low-coverage contact matrices, we started with a deep-sequenced Hi-C contact map obtained from GM12878 human cells [56], and down-sampled the contact counts to generate lower-coverage matrices. Specifically, for each non-zero contact count M_{ij} in the original matrix, we assumed that the simulated contact count follows a binomial distribution $M'_{ij} \sim \text{Binomial}(M_{ij}, p)$, where the binomial param-

eter $p = \{0.2, 0.4, 0.6, 0.8, 1.0\}$ represents the relative coverage level of the down-sampled contact matrix M'. In addition, 10% noise were added to the down-sampled matrices.

2.2.5 Real data pre-processing steps

In our real data analysis, we used the published Hi-C datasets by Rao et. al. 2014 [56] (GEO accession number: GSE63525), which include five human cell types: B-lymphoblastoid cells (GM12878), mammary epithelial cells (HMEC), umbilical vein endothelial cells (HUVEC), erythrocytic leukemia cells (K562), and epidermal keratinocytes (NHEK). The GM12878 dataset contains two replicates, which were pooled together in cell type-specific compairson analyses. We applied DiffGR to detect differential genomic regions between each pair of cell types at 25-kb, 50-kb, and 100-kb resolutions. Since some of these Hi-C datasets were not deeply sequenced, the local variations introduced by low sequencing coverage made it challenging to capture large domain structures, especially in fine-resolution analyses. Therefore, to enhance the domain structures, all contact matrices were first pre-processed by a 2D mean filter smoothing and then normalized by the KR method to eliminate potential biases.

In addition to Hi-C contact maps, ChIP-seq and RNA-seq data were also applied in real data result analysis. CTCF and histone modification (including H3K4me1, H3K4me2, H3K27me3, and H3K36me3) ChIP-seq datasets from five cell lines were obtained from ENCODE project [6] (https://www.encodeproject.org/). The ChIP-seq peak files were in narrowpeak/broadpeak BED format. The ChIP-seq peaks were aggregated into fixed-size bins with the same resolution as the Hi-C data, and the bin-wise peak counts were normalized by the total number of peaks in each ChIP-seq dataset. The absolute mean

differences of the normalized bin-wise peak counts were calculated for each pair of cell lines for the subsequent analyses.The RNA-seq datasets were obtained from the ENCODE project (The ENCODE Project Consortium 2012) with accession numbers GSE78553 for GM12878 cells and GSE78625 for K562 cells. The expression value of genes from samples were in read count format.

2.3 Results

2.3.1 DiffGR accurately detected single-TAD differences in simulated datasets

To validate the accuracy and efficiency of our DiffGR method, we first generated pairs of original and simulated Hi-C contact matrices, where a given proportion of TADs in the simulated contact matrices were altered (see Methods). We used the intra-chromosomal contact matrix of chromosome 1 in K562 cells at 50-kb resolution to serve as the original contact matrix. At the default setting, we altered 50% of the original TADs by completely replacing the intra-TAD contact counts by randomly sampled counts outside the TAD regions. In addition, we added 10% random-ligation noise into the altered contact matrices.

We first simulated Hi-C matrices with various proportions of altered TADs (20%, 30%, 40%, 50%, 60%, and 70%). With each proportion setting, we completely mutated the intra-TAD counts and added 10% noise, and repeated this simulation procedure 100 times. As expected, the performance of the DiffGR method depended on the proportion of altered TADs. As shown in Figure 2.4a and Supplementary Table S2.1, when the proportion of

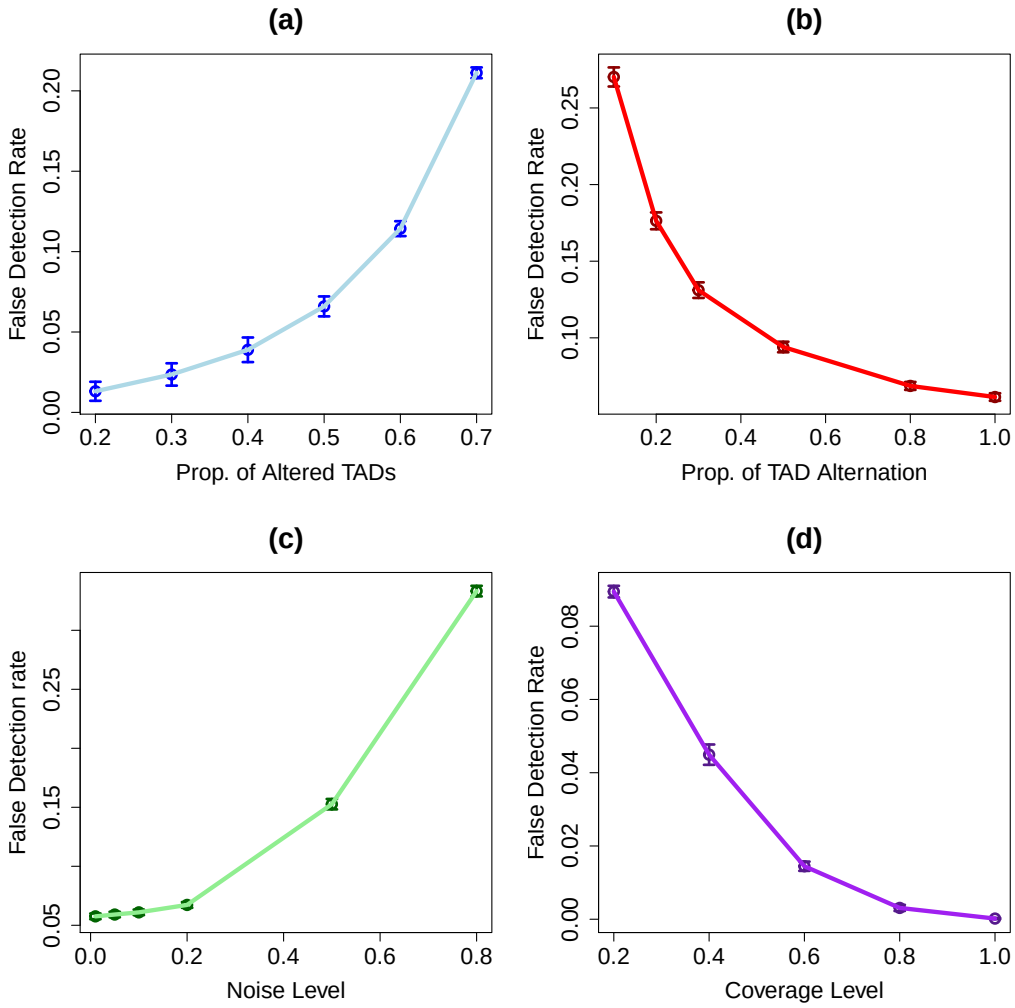

Figure 2.4: **Performance of single-TAD simulations.** The curves display the mean false detection rates at different levels of (a) proportion of altered TADs, (b) proportion of TAD alternation, (c) noise, and (d) sequencing coverage. Vertical bars represent 95% confidence intervals.

altered TADs changed from 20% to 70%, the false detection rate increased from 0.01 to 0.21. One possible explanation of this observed trend is that when the majority of TADs were altered, the large differences between the original and altered matrices would affect the permutation test and therefore lead to inaccurate detection. However, differential TADs rarely exist in large proportion in real data. The false detection rates of our method remained below 0.07 when the proportion of altered TADs was smaller than or equal to 50%, which demonstrated that our method can accurately and reliably detect single-TAD differences under these conditions.

In the default simulation setting, we completely altered the selected TADs by substituting all intra-TAD contact counts by randomly sampled counts from the matching diagonals outside the TADs. To investigate the influence of the degree of TAD alternation on the DiffGR performance, we generated a series of simulated contact matrices, in which half of original TADs were altered and the proportion of intra-TAD alternation varied from 10%, to 20%, 30%, 50%, 80%, and 100%. In theory, TADs with higher degrees of alternation are easier to identify, whereas TADs with minor changes remain difficult to be detected. As illustrated in Figure 2.4b and Supplementary Table S2.2, the performance of DiffGR improved resulting in higher accuracy as the percentage of randomly substituted counts in altered TADs increased. Even with the most challenging case where only 10% of the intra-TAD counts were altered, the accuracy of our method was 0.73, suggesting that DiffGR can effectively detect subtle TAD differences.

2.3.2 DiffGR performed stably against changes in noise and coverage levels

Next we sought to evaluate the robustness of our method under various noise and sequencing coverage conditions. In the earlier simulations, we added 10% noise to the simulated differential contact matrices. To evaluate the performance of our method under different noise levels, we fixed the proportion of altered TADs at 50% and the proportion of intra-TAD alternation at 100%, and simulated the differential contact matrices with a wide range of noise levels (1%, 5%, 10%, 20%, 50%, and 80%). Intuitively, a good detection method should easily discover the differential regions in the less noisy matrices, and it becomes more challenging to detect the differential regions in the noisier cases. Our results demonstrated that DiffGR was able to correctly rank the simulated datasets. We observed a monotonic increasing trend of the false detection rate and a decreasing tendency of other precision measures as the noise levels raised (Figure 2.4c and Supplementary Table S2.3). With moderate noise levels that were not greater than 20%, the accuracy of DiffGR remained above 0.93, indicating that our method can correctly detect differential TAD regions in such noisy cases.

The sequencing coverage of the Hi-C contact maps is another major factor that could affect the performance of our method. Considering two Hi-C replicates that have the same underlying TAD structures but different sequencing coverage levels, we questioned whether our DiffGR method can correctly categorize them as non-differential. In other words, we intended to estimate the false positive rates caused by low-coverage and sparse Hi-C data. To directly investigate the influence of the sequencing coverage on the detection

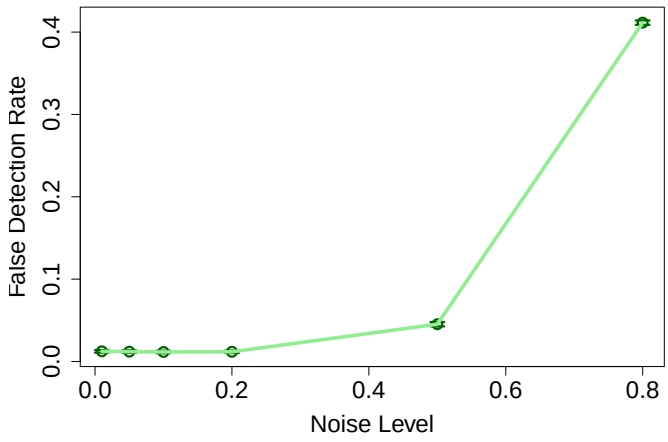

Figure 2.5: **Performance of hierarchical-TAD simulations.** The curve shows the mean false detection rates at various noise levels. Vertical bars represent 95% confidence intervals.

of differential regions, we utilized the GM12878 chromosome 1 contact matrix as the original matrix, and generated a series of down-sampled contact matrices with lower coverage levels (20%, 40%, 60%, 80%, and 100%). Figure 2.4d and Supplementary Table S2.4 shows that the average false detection rates remained below 0.05 for most coverage levels, except for the lowest coverage level of 20%, demonstrating the robustness of our DiffGR method under low-coverage conditions.

2.3.3 DiffGR successfully detected hierarchical-TAD changes

In addition to single-TAD differences, hierarchical-TAD changes also exist in some genomic regions between different cell types. In these regions, one of the Hi-C contact maps exhibits a single dominant TAD structure, while the other Hi-C contact map presents two or more subTADs separated by additional boundaries in between. Hierarchical TADs are computationally challenging to detect. Although the two Hi-C maps have different

TAD boundaries, the chromatin interaction patterns within the subTADs could be very similar. Consequently, the correlation coefficients (CCs) for the strata with small genomic distances might still remain high between two contact maps. In addition, as the genomic distance increases, the weight of the corresponding stratum in the SCC calculation gradually declines. As a result, the SCC values are primarily contributed by CC values from strata with smaller genomic distances, which makes it difficult to detect differential regions in the hierarchical-TAD cases.

To evaluate the performance of DiffGR in this more challenging situation, we simulated contact matrices containing hierarchical-TAD structures with respect to varying noise levels (see Methods) and then computed the false detection rate in a similar manner as in the single-TAD simulations. As demonstrated in Figure 2.5 and Supplementary Table S2.5, the trend of the false detection rates and other measure statistics across various noise levels under the hierarchical-TAD setting was similar to the pattern observed in the single-TAD case (Figure 2.4c and Supplementary Table S2.3). Furthermore, the false detection rates maintained low (less than 0.05) when the noise level was within 50%. Taken together, these results indicated that DiffGR can reliably detect the differentially interacting genomic regions with hierarchical-TAD patterns.

2.3.4 SCC outperformed Pearson CC in measuring similarity of local TAD regions

In the proposed DiffGR method, we used SCC to measure the similarity of local TAD regions between two Hi-C contact maps. In addition to SCC, other commonly used similarity measurements for comparing Hi-C contact matrices include Pearson and

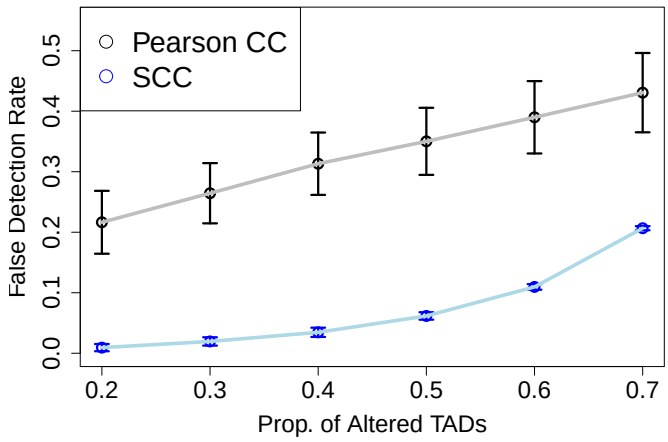

Figure 2.6: **Comparison between SCC and Pearson CC.** The curves represent the mean false detection rates at various proportions of altered TADs using either SCC (blue) or Pearson CC (black) as the local similarity metric. Vertical bars represent 95% confidence intervals.

Spearman CCs. The main advantage of SCC over the standard CCs is that SCC explicitly takes the genomic-distance effect into consideration, thereby achieving better performance in evaluating the Hi-C contact matrices [71]. Therefore, we expected SCC to serve as a good metric to compare chromatin interacting patterns at local TAD regions.

To validate our choice of the SCC similarity metric, we tested a variation of the DiffGR method that substitutes SCC with the standard Pearson CC, and evaluated its performance using the simulated Hi-C contact matrices with various proportions of altered TADs. As shown in Figure 2.6 and Supplementary Table S2.6, our method utilizing SCC evidently outperformed the other alternative employing Pearson CC. For each proportion of altered TADs, the false detection rates based on Pearson CC were significantly higher than those replying on SCC. Moreover, the variations of the false detection rates measured by SCC were much smaller than those obtained by Pearson CC. Therefore, these results

demonstrated that SCC is indeed a better similarity metric than Pearson CC in measuring local TAD patterns between Hi-C contact matrices.

2.3.5 DiffGR revealed cell type-specific genomic interacting regions

After validating our method on simulated datasets, we further applied DiffGR to published Hi-C datasets in five human cell types (GM12878, HMEC, HUVEC, K562, and NHEK) [56]. In total, we conducted one comparison of biological replicates in GM12878 and ten pairwise comparisons among five cell types.Then we identified statistically significant differential genomic regions between each comparison with FDR cutoff 0.05. In each pairwise comparison, we first applied HiCseg to identify TAD boundaries from the 50-kb contact matrix for each data and then partitioned the genome into three types of candidate regions: single-TAD candidate regions, hierarchical-TAD candidate regions, and complex-TAD candidate regions.

We first sought to evaluate the performance of our method on GM12878 biological replicates. Previous studies have shown that the high degree of similarity between bio-replicates and dominant consistence between TAD boundaries in replicate data [72, 16, 56]. As expected, majority (89.55%) of genomic regions belonged to single-TAD type and few (2.45%) differential genomic regions were detected by our method (Figure 2.7). Specifically, only 1.97% of single-TADs were identified as differential while 6.17% and 4.94% were detected in hierarchical-TADs and complex-TADs respectively, indicating that candidate genomic regions are more likely to be differential between replicate samples when some unique patterns of TAD boundaries appeared.

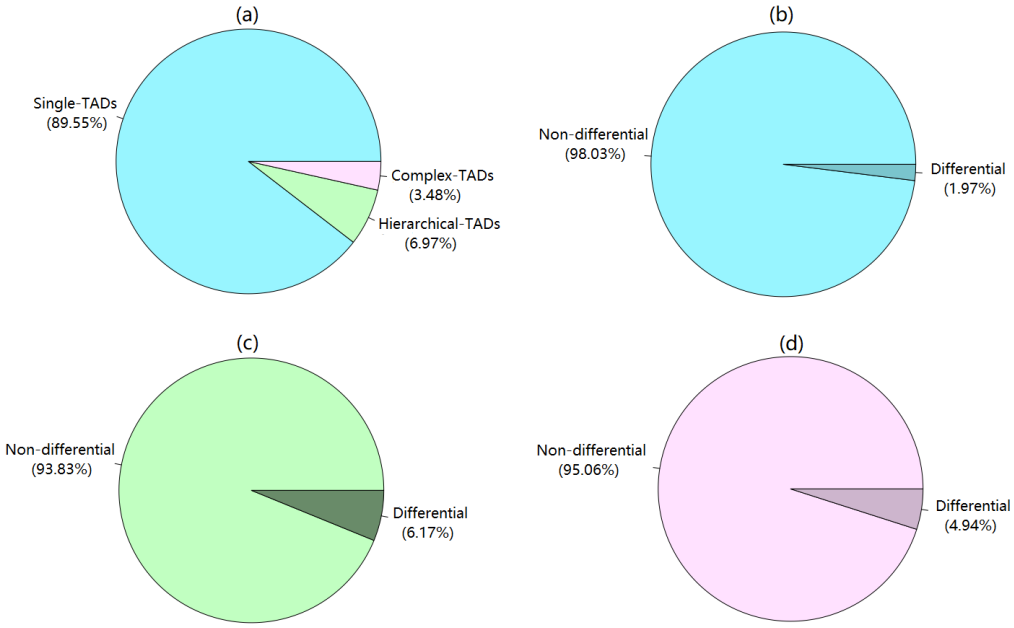

Figure 2.7: **Piecharts of DiffGR results obtained from human GM12878 bio-replicates.** Piechart(a) presents the proportions of three types of candidate regions. The rest three piecharts (b)-(d) display the proportions of detected differential genomic regions in each candidate category(blue for single-TADs, green for hierarchical-TADs and purple for complex-TADs).

Next, for the candidate regions from all ten pairwise comparisons, as illustrated in Figure 2.8, 55.57% belonged to the single-TAD category (consistent with previous observations indicating that TAD boundaries are stable across cell types [16]), 31.88% to the hierarchical-TAD category, and 12.55% to the complex-TAD category. Our DiffGR analyses showed that only 24.26% of the single-TAD candidate regions showed statistically significant differences between two samples; 59.24% of the hierarchical-TAD candidate regions were determined to be differential; while the differential proportion of the complex-TAD category was as high as 89.82%. These observations indicated that candidate genomic regions with more distinct patterns of TAD boundaries are more likely to be detected as differential be-

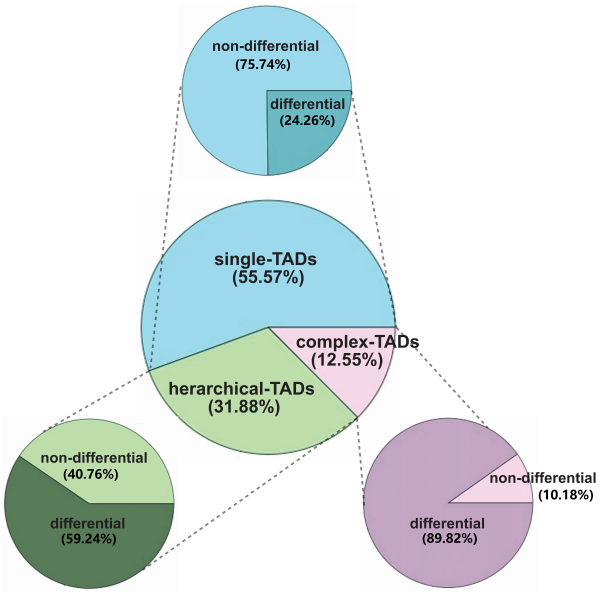

Figure 2.8: **Piecharts of DiffGR results obtained from human Hi-C datasets.** The center piechart presents the proportions of three types of candidate regions. The three outer piecharts display the proportions of detected differential genomic regions, one for each candidate category.

tween two Hi-C samples. In addition, we found that the proportion of detected differential regions varied largely across chromosomes, ranging from 0.14 to 0.76 (Figure 2.9).

In addition to partitioning the genome at 50-kb resolution, we also performed differential analyses on the five human Hi-C datasets at 25-kb and 100-kb resolutions, separately. We calculated the overlapping rate (that is, the proportion of the genome that was classified into the same differential or non-differential status) between different resolutions. Overall, we observed a high consistency between the detected differential regions across different resolutions, where the overlapping rate was 0.9856 between the detection results at 50-kb and 100-kb resolutions, and 0.9480 between those at 25-kb and 50-kb res-

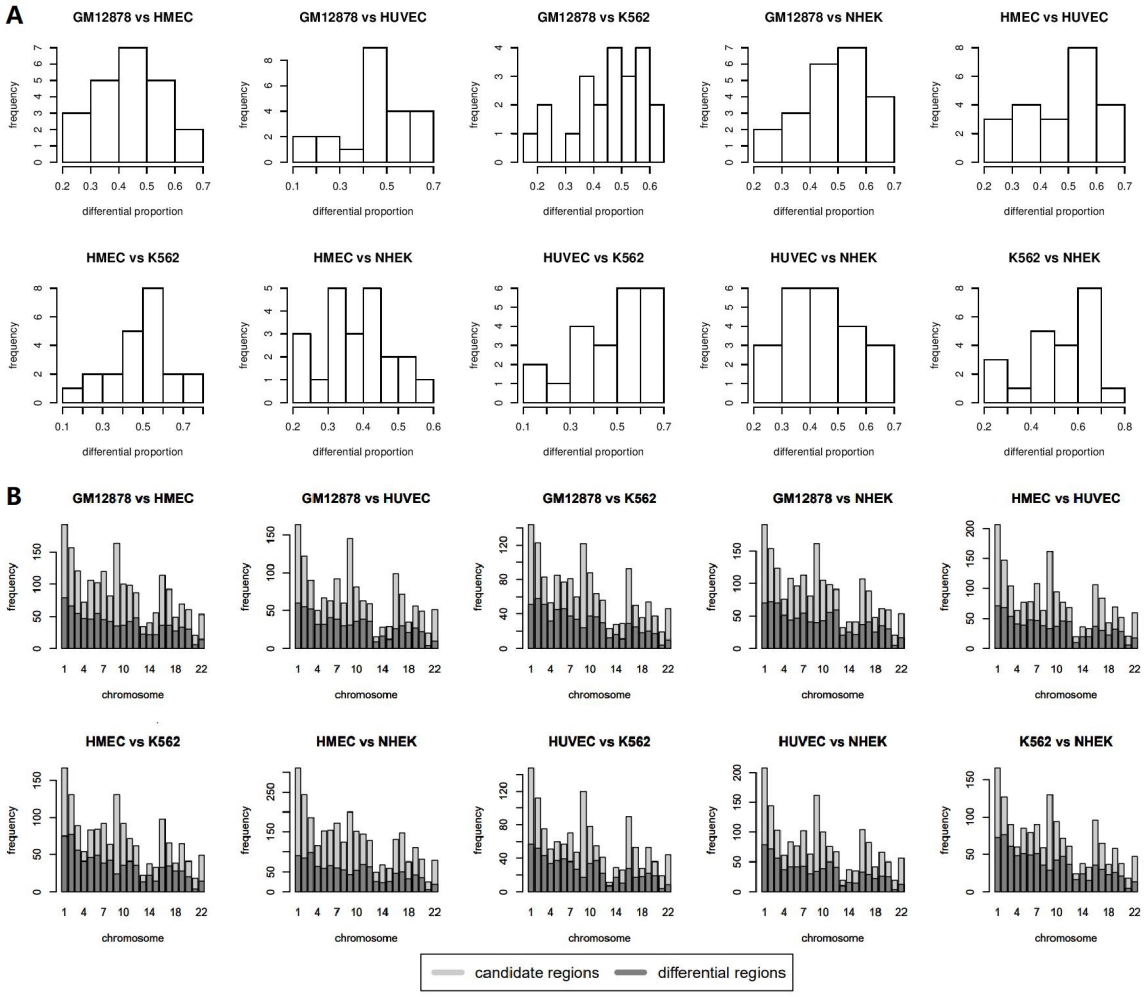

Figure 2.9: **Summary of DiffGR-detected differential genomic regions in human Hi-C datasets.** (a). Histograms of chromosome-wide proportion of deferentially interacting genomic regions for all pairwise comparisons between two cell types; (b). Barplots of the numbers of candidate regions and detected differential genomic regions per chromosome for all pairwise comparisons between two cell types.

Table 2.1: **Agreements between ChIP-seq data and DiffGR-detected differential genomic regions.**

	100 kb	50 kb	25 kb
CTCF	76 (34.55%)	124 (56.36%)	142 (64.55%)
H3K4me1	57 (25.91%)	110 (50.00%)	136 (61.82%)
H3K4me2	56 (25.45%)	91 (41.36%)	116 (52.73%)
H3K27me3	53 (24.09%)	86 (39.09%)	114 (51.82%)
H3K36me3	36 (16.36%)	72 (32.73%)	110 (50.00%)

A total of 220 t-tests (10 pairwise comparisons between cell types, 22 chromosome-wide tests for each comparison) were conducted. If the mean absolute differences of a ChIP-seq signal at the TAD boundaries in the differential regions were significantly higher than those in non-differential regions, the results were labeled significant consistent. The counts and percentages of significant consistent results were reported for each ChIP-seq dataset at each resolution.

olutions. These results demonstrated that DiffGR can robustly and consistently detect cell type-specific differential genomic regions across various resolutions.

2.3.6 Changes in CTCF and histone modification patterns were consistent with DiffGR detection results

As there is no ground truth of differential chromatin interacting regions in real data, we sought to evaluate the performance of our method by investigating the association between the changes of 1D epigenomic features and 3D genomic interaction regions. The chromatin architectural protein CTCF plays an essential role in establishing higher-order chromatin structures such as TADs. In addition, it has been shown that CTCF and many histone marks are enriched or depleted at TAD boundaries.[16] Therefore, we hypothesized that differential bindings of CTCF and histone modifications would also be present at the TAD boundaries in differential genomic interacting regions.

To test this hypothesis, we first combined TAD boundaries from both Hi-C datasets and classified them into two categories: those within the DiffGR-detected differential regions and those outside the differential regions. We then utilized the ChIP-seq datasets of CTCF and histone modifications (including H3K4me1, H3K4me2, H3K27me3, H3K36me3) from the ENCODE project [6]. For each ChIP-seq dataset, we calculated the mean absolute difference of ChIP-seq peaks between the two cell types within the neighborhood (+/- 1 bin) of each TAD boundary. We expected that if two cell lines have highly different chromatin structures in certain genomic regions, we would observe different patterns of CTCF bindings and histone modifications in these regions. Therefore, we performed the following t-test for each ChIP-seq dataset using the DiffGR detection results. In each chromosome, we evaluated whether the mean absolute differences of the ChIP-seq signal at the TAD boundaries in differential regions were significantly different from those in non-differential regions. If the ChIP-seq differences at the TAD boundaries in differential regions were significantly higher (with a significant level 0.1) than those in non-differential regions, we considered the ChIP-seq changes to be consistent with our differential detection results.

Table 2.1 summarizes the ChIP-seq analyses on the DiffGR detection results obtained from five human Hi-C datasets at 100-kb, 50-kb, and 25-kb resolutions. For each ChIP-seq dataset, we performed 220 t-tests (ten pairwise comparisons between cell types, 22 chromosome-wide tests one for each autosome) at each resolution. Overall, we observed that the agreement between the changes of ChIP-seq signal and chromatin structures was improved in finer-resolution analyses. As shown in Table 2.1, 76 out of 220 (34.55%) tests showed significantly higher absolute mean differences of CTCF values at the TAD bound-

aries in DiffGR-detected differential genomic regions than those in non-differential regions at 100-kb resolution. Whereas in the results at 25-kb resolution, 142 (64.55%) tests exhibited significantly larger changes of CTCF bindings in differential regions than non-differential ones. In addition, the histone modification datasets showed similar results in agreement with the detection results of differentially interacting regions in Hi-C contact maps. At 25-kb resolution, the majority of the t-tests showed significantly larger changes of ChIP-seq signal in differentially interacting regions for all four histone modification datasets, including H3K4me1, H3K4me2, H3K27me3, and H3K26me3. Collectively, these results indicated that the changes in CTCF bindings and histone modifications were in good agreements with the differences in genomic interacting regions. Furthermore, at finer resolution our DiffGR method produced more accurate identification of differentially interacting genomic regions in higher agreement with the CTCF and histone modification data.

We would like to point out that the cases where the changes of CTCF or histone modifications are not in significant agreement with the detection results of differentially interacting genomic regions do not necessarily suggest that these epigenomic features are inconsistent with 3D genome organization or DiffGR detection results are inaccurate. Due to the resolution limit of Hi-C contact maps, the boundaries of differential regions are usually identified with a resolution of tens to hundreds of kilobases. Aggregating ChIP-seq data with such a large bin size dilutes the signal, thereby yielding less statistical power to detect significant changes. Moreover, CTCF and histone modifications play fundamental roles in regulating chromatin structures and gene expression; their effects are not limited to TAD formations. Therefore, changes of CTCF bindings or histone modifications exist in

many genomic loci other than TAD boundaries, therefore they may not be reflected in our analyses.

2.3.7 Differential RNA-seq results were consistent with DiffGR detection results

Besides the investigation of the changes of 1D epigenomic features, we further study the relationship between quantitative changes in expression levels and 3D genomic interaction regions to better assess the performance of our method. Previous research mentioned that topological changes have a large effect on the cross-talk between enhancers and promoters that can alter gene expression[56, 15]. Thus, we expected that differential expressed genes would be probably located in differential genomic regions.

To evaluate the assumption, we first detected significant changes in expression levels between GM12878 and K562 cell by DESeq2[44]. Later we calculated the percentage of detected differential expressed genes whose loci were inside identified differential genomic regions. To check the superiority of such proportion, we randomly chose partial genes, whose number equals to the number of detected changes in expression levels, with 200 iterations, then computed their corresponding proportions located in differential regions and performed t-test for comparison.

In summary, a total number of 9120 differential expressed genes were detected by DESeq2 at the significance level of 0.05 and 79.54% (with p-value=3.72×10^{-5}) of them were located in Differential genomic regions, demonstrating that majority changes appeared in RNA-seq data happened to be consistent with DiffGR detection results.

Table 2.2: **Functional enrichment of differential genes located in differential genomic regions**

Go Term	P_value
GO:0002376 immune system process	1.7E-9
GO:0050776 regulation of immune response	5.9E-8
GO:0002757 immune response-activating signal transduction	7.8E-8
GO:0002682 regulation of response to stress	2.2E-7
GO:0080134 regulation of immune system process	2.7E-7
GO:0045321 leukocyte activation	2.8E-7

Top 2000 differential genes located within differential genomic regions at 25kb resolution were utilized in GO enrichment test. Test results were given by DAVID[60].

To further explore the potential functional roles of differential expressed genes located in differential genomic regions, we performed Gene Ontology (GO) enrichment analysis on those top 2000 genes using DAVID [60]. From Table 2.2, it was shown a high enrichment for GO terms related to the immune system, which is consistent with previous researches that many H3K4me1 peaks overlap with known autoimmune disorder SNPs in the B-lymphoblast cell line GM12878[8].

2.3.8 DiffGR detection results were supported by FIND and TADCompare results

Several previous Hi-C comparative studies mentioned that majority of the chromatin structural changes tend to couple with the formation/disappearance of topologically associated domains (TADs) [56, 15], implying that many changes in interaction counts are likely to be observed within genomic regions at TAD level. Hence, we checked differential chromatin interactions (DCIs) between GM12878 and K562 cells at 50kb resolution by

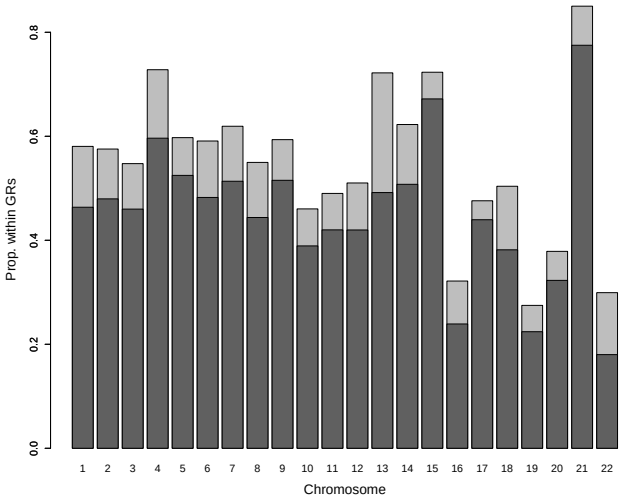

Figure 2.10: **Result Comparison between FIND and DiffGR.** Barchart of the proportions of DCIs detected by FIND located in candidate genomic regions (GRs)/Differential GRs for chromosomes between GM12878 and K562. The bars denote the proportions of DCIs detected by FIND located in candidate GRs and the dark grey bars represent the proportions of DCIs specially classified as differential GRs.

FIND[17] and contrasted FIND results with our DiffGR results. From Figure 2.10, it was shown that the percentages of DCIs detected by FIND located within candidate genomic regions were dominant in the majority of chromosomes and with 55.43% across the whole genome. In addition, 82.80% of the DCIs located in candidate genomic regions are classified into differential regions, demonstrating that DiffGR effectively detected the regions with significant changes in counts.

Next, we explored differential TAD boundaries detection for previous five human Hi-C datasets at 50kb resolution given by TADcompare[11], showing that 76.25% differential TAD boundaries displayed concordant results with DiffGR. To investigate the advantage of DiffGR compared with TADCompare, we further performed tests on changes in CTCF and histone modification patterns for those detected differential TAD boundaries

Table 2.3: **Advantageous results of differential TAD boundaries in DiffGR-detected differential genomic regions.**

	consistent number	significant consistent number
CTCF	155 (70.45%)	98 (44.55%)
H3K4me1	145 (65.91%)	89 (40.45%)
H3K4me2	133 (60.45%)	79 (45.91%)
H3K27me3	146 (66.36%)	76 (34.55%)
H3K36me3	127 (57.73%)	51 (23.18%)

A total of 220 tests (10 pairwise comparisons between cell types, 22 chromosome-wide tests for each comparison) were conducted. If the mean absolute differences of a ChIP-seq signal at the differential TAD boundaries detected by TADCompare in the differential regions were higher/significantly higher(based on t-tests) than those in non-differential regions, the results were labeled consistent/significantly consistent. The counts and percentages of consistent/significant consistent results were reported for each ChIP-seq dataset.

within differential/non-differential genomic regions. From Table 2.3, we observed that 155 out of 220 (70.45%) contrasts showed higher absolute mean differences of CTCF values at differential TAD boundaries in DiffGR-detected differential genomic regions than those in non-differential regions and 98 (44.55%) tests exhibited significantly larger changes of CTCF bindings at TADCompare-detected differential boundaries in differential regions than non-differential ones. In addition, the histone modification datasets (including H3K4me1, H3K4me2, H3K27me3, and H3K26me3) showed similar results in agreement with the advantageous results of differential bounds in differentially interacting regions.

2.4 Discussion

With the fast growth of Hi-C datasets, there has been a dramatically increasing interest in comparative analysis of Hi-C contact maps. However, most existing methods for comparative Hi-C analysis focused on the identification of differential chromatin inter-

actions, while few studies addressed the detection of differential chromatin organization at TAD scale.

To solve this problem, we developed a novel method, DiffGR, for calling differentially interacting genomic regions between two Hi-C contact maps. Taking genomic distance features of Hi-C data into consideration, our algorithm utilized the SCC metric instead of the standard Pearson CC to measure the similarity of local genomic regions between Hi-C contact maps. Furthermore, we proposed a non-parametric permutation test to assess the statistical significance of the local SCC values. Through empirical evaluations, we have demonstrated that DiffGR can effectively discover differential regions in both simulated data and real Hi-C data from different cell types. That is, DiffGR produced robust and stable detection results under various noise and coverage levels in simulated data; DiffGR detection results in real data were effectively validated by the ChIP-seq and RNA-seq data and shown consistency and advantage compared with TADCompare results.

We envision a few possible extensions and future directions based on this work. First, our method performs pairwise comparison between two Hi-C contact maps. Extending our method to differential analyses among three or more samples would require a more general statistical framework. In addition, we currently utilized pooled Hi-C maps in our analyses. One possible future direction is to incorporate multiple biological replicates to detect reproducible differences. Lastly, our method is specifically designed for bulk Hi-C data, which remains a significant challenge in identifying differential regions at single-cell level.

Software availability

The software is published under the GNU GPL $\geq$ 2 license. The main function DiffGR is explained in the Appendix A and the source code is publicly available at https://github.com/wmalab/DiffGR.

Supplementary Tables

The supplementary tables show the summary of the performance on simulated data. The performance of each condition's effect on DiffGR detection was assessed using the following metrics:

Index	Illustration
TP	True positives
FP	False positives
TN	True negatives
FN	False negatives
Sensitivity	TP/(TP +FN)
Specificity	TN/(FP +TN)
Accuracy	(TP +TN)/(TP +FP +TN +FN)
Precision	TP/(TP + FP)
F1 score	2TP/(2TP +FP +FN)
MCC	Matthews correlation coefficient: $\frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$

Supplementary Table S2.1: Evaluation of the effect of proportion of altered TADs on DiffGR detection.

	0.2	0.3	0.4	0.5	0.6	0.7
TP	36.53	53.64	69.82	82.62	92.39	92.95
FP	0.29	0.33	0.36	0.27	0.11	0
TN	150.71	131.67	112.64	94.73	75.89	57
FN	1.47	3.36	6.18	11.38	20.61	39.05
Sensitivity	0.9613	0.9411	0.9187	0.8789	0.8176	0.7042
Specificity	0.9981	0.9975	0.9968	0.9972	0.9986	1
Accuracy	0.9907	0.9805	0.9654	0.9384	0.8904	0.7934
Precision	0.9924	0.994	0.9949	0.9967	0.9988	1
F1 score	0.9764	0.9666	0.9551	0.9339	0.8988	0.8262
MCC	0.9709	0.9537	0.9292	0.883	0.8011	0.6468

Supplementary Table S2.2: Evaluation of the effect of proportion of TAD alternation on DiffGR detection.

	0.1	0.2	0.3	0.5	0.8	1
TP	43.38	61.37	69.9	77.06	81.95	83.28
FP	0	0	0	0	0	0
TN	95	95	95	95	95	95
FN	50.62	32.63	24.1	16.94	12.05	10.72
Sensitivity	0.4615	0.6529	0.7436	0.8198	0.8718	0.886
Specificity	1	1	1	1	1	1
Accuracy	0.7322	0.8274	0.8725	0.9104	0.9362	0.9433
Precision	1	1	1	1	1	1
F1 score	0.6291	0.7886	0.852	0.9006	0.9313	0.9394
MCC	0.5485	0.6975	0.7707	0.8344	0.8798	0.8925

Supplementary Table S2.3: Evaluation of the effect of noise level on DiffGR detection.

	0.01	0.05	0.1	0.2	0.5	0.8
TP	84.06	83.75	83.39	82.23	66.04	32.2
FP	0.94	0.92	0.91	0.94	0.89	1.18
TN	94.06	94.08	94.09	94.06	94.11	93.82
FN	9.94	10.25	10.61	11.77	27.96	61.8
Sensitivity	0.8943	0.891	0.8871	0.8748	0.7026	0.3426
Specificity	0.9901	0.9903	0.9904	0.9901	0.9906	0.9876
Accuracy	0.9424	0.9409	0.939	0.9328	0.8474	0.6668
Precision	0.9889	0.9891	0.9892	0.9887	0.9867	0.9652
F1 score	0.939	0.9373	0.9352	0.9281	0.8199	0.504
MCC	0.8891	0.8863	0.8829	0.8714	0.7251	0.4327

Supplementary Table S2.4: Evaluation of the effect of coverage level on DiffGR detection.

	1	0.8	0.6	0.4	0.2
TP	0	0	0	0	0
FP	0	0.19	0.5211	2.75	14.5079
TN	220	219.81	219.4789	217.25	205.4921
FN	0	0	0	0	0
Sensitivity	1	1	1	1	1
Specificity	1	0.9991	0.9976	0.9875	0.9341
Accuracy	1	0.9991	0.9976	0.9875	0.9341
Precision	1	0.81	0.5634	0	0
F1 score	1	0.81	0.5634	0	0
MCC	1	1	1	1	1

Supplementary Table S2.5: Evaluation of the effect of hierarchical setting on DiffGR detection.

	0.01	0.05	0.1	0.2	0.5	0.8
TP	91.67	91.75	91.82	91.78	85.58	16.71
FP	0	0	0	0	0.13	0.48
TN	95	95	95	95	94.87	94.52
FN	2.33	2.25	2.18	2.22	8.42	77.29
Sensitivity	0.9752	0.9761	0.9768	0.9764	0.9104	0.1778
Specificity	1	1	1	1	0.9986	0.9949
Accuracy	0.9877	0.9881	0.9885	0.9883	0.9548	0.5885
Precision	1	1	1	1	0.9985	0.9725
F1 score	0.9874	0.9878	0.9882	0.988	0.9522	0.2998
MCC	0.9757	0.9766	0.9773	0.9769	0.9133	0.2998

Supplementary Table S2.6: Evaluation of Pearson correlation coefficient performance on DiffGR detection.

	0.2	0.3	0.4	0.5	0.6	0.7
TP	7.96	5.98	1.57	0.74	0.47	0.41
FP	33.67	33.16	30.37	29.33	28.97	29.01
TN	140.13	133.04	128.23	122.07	114.83	107.19
FN	7.24	16.82	28.83	36.86	44.73	52.39
Sensitivity	0.8095	0.7049	0.6207	0.6079	0.6042	0.6031
Specificity	0.8215	0.8241	0.8388	0.8448	0.8467	0.8465
Accuracy	0.7835	0.7356	0.6868	0.6498	0.6101	0.5693
Precision	0.4321	0.4441	0.4546	0.43	0.44	0.4
F1 score	0.273	0.1805	0.0933	0.0451	0.0482	0.0062
MCC	0.8515	0.7526	0.6817	0.6463	0.6449	0.6194

Chapter 3

scHiCDiff: Detection of Single-cell Hi-C Differential Chromatin Interactions

3.1 Introduction

The stochastic nature of chromosome conformation and spatial genome organization results in variations of cell-to-cell chromatin interactions, even among cells of a functionally homogeneous population[22]. Thus, although ensemble Hi-C is a powerful tool for capturing the geometry of genome organization, it is not sufficient to employ bulk Hi-C data to illustrate the heterogeneity of higher order chromosome structures among individual cells [22]. In addition, it has been shown that single-cell studies on RNA-seq for RNA expression [66] and ATAC-seq for chromatin accessibility [12] provided deep insight into

the interplay between intrinsic cellular processes and dynamic gene expression in biological and biomedical areas. Considering the great potential value of single-cell Hi-C (scHi-C) on researches of cell-specific chromosomal architecture, several scHi-C protocols have been developed [50, 54, 23].

Similar to ensemble Hi-C data, identifying biologically interesting interactions from scHi-C maps are challenging, since scHi-C data also contains systematic biases in terms of effective length, GC content and mappability of fragment ends [69]. scHiCNorm [43] introduced zero-inflated and hurdle models to remove those systematic biases. However, when interactions specific to a certain experiment condition or cell line are being sought, the significant read counts derived after normalization may not be of scientific interest. An alternative key approach is to figure out differences in chromatin structure by identifying interaction counts that are significantly varied across two or more biological conditions at single-cell level.

Despite there have been several advanced comparative analysis methods for the detection of differential chromatin interactions (DCIs) in bulk Hi-C data (See section 2.1) [45, 64, 17, 7], considering the unique features of data sparsity of single-cell Hi-C data, it is not suitable to simply apply those approaches in single-cell Hi-C data. To tackle the problems of differential analysis specialized on scHi-C data, we looked through several methods from singe-cell RNA sequencing data for inspiration. Similar to Hi-C data, RNA sequencing data also faced high heterogeneity and excessive zero problems in single-cell data compared to bulk one. A variety of methods have recently been proposed to analyze differential expression in single cell RNA-seq (scRNA-seq) data. For instance, SCDE [36]

modeled the count frequencies of each cell as a mixture of a zero-inflated Negative Binomial distribution and a dropout component (poisson distribution). D3E [13] used non-parametric tests (Cramer-von Mises test, Kolmogorov–Smirnov test) or parametric Poisson-Beta model to compare the distributions of expression values of each gene for identifying the DE genes. Later, DEsingle [47] utilized a zero-inflated Negative Binomial(ZINB) regression model to estimate the NB parameters (mean and dispersion parameters) and the proportion of the real and drop-out zeros in the observed expression data. According to the estimators, a hypothesis test was performed to decide whether the two ZINB populations had significant difference.

Inspired by the D3E and DEsingle approaches, we designed a novel statistical algorithm scHiCDiff to detect differential chromatin interactions between two Hi-C experiments at single-cell level. In our method, we introduced both non-parametric tests and likelihood ratio tests with parametric models to capture the bin pairs showing significant changes, and demonstrated that zero-inflated Negative Binomial (ZINB) and Negative Binomial Hurdle (NBH) regression models can effectively eliminate the effects of extreme sparsity and provide reliable detection results of DCIs in scHi-C comparative analysis.

3.2 Methods

The scHiCDiff tool identifies the changes in chromatin interactions in two steps: data normalization and differential detection tests, as described below (Sections 3.2.1-3.2.3). In addition, the simulation settings are explained in Section 3.2.4 and real data pre-processing are outlined in Section 3.2.5.

3.2.1 Data Normalization

As mentioned in Section 3.1, scHi-C data also contains systematic biases in terms of effective length, GC content. and mappability of fragment ends and scHiCNorm was specially designated to eliminate those systematic biases [43]. Hence, based on the guidance from scHiCNorm [43], we first processed the data by scHiCNorm normalization with the negative binomial hurdle option.

Consider m raw single-cell contact count matrices $\tilde{C}^k \in N^{n\times n}$, for $k = 1, \cdots, m$, where $\tilde{c}^k_{ij}$ is the interaction count between loci i and j in the k-th experiment ($1 \leq i, j \leq n$ and $1 \leq k \leq m$). Here, we assume the observed count ($\tilde{c}^k_{ij}$) in the scHi-C contact matrix follows a negative binomial hurdle model. After fitting a regression model per chromosome per cell with bias features being variables, the normalized count ($\hat{c}^k_{ij}$) is calculated as the observed count ($\tilde{c}^k_{ij}$) divided by the estimated mean of the regression model ($\hat{\mu}^k_{ij}$).

Further, the genomic distance effect, whereby pairs of genomic loci that are proximal along the chromosome exhibit many more Hi-C contacts that distal pairs of loci, dominates every single-cell and bulk Hi-C matrix. Also, the form that such distance effect often varies between different Hi-C experiments. Thus, we introduce a size factor to account for the genomic distance bias. Here, we enforce that the median normalized count for pairs of bins at each given distance in each matrix is the same, by taking for $d \in [0, n-1]$:

$$\hat{s}^k_d = \underset{|i-j|=d}{median}\hat{c}^k_{ij}.$$

Therefore, the final normalized count between loci i and j in the k-th experiment is given by

$$c_{ij}^k = \frac{\hat{c}_{ij}^k}{\hat{s}_{|i-j|}^k}.$$

3.2.2 Detecting Differential Interactions by Non-parametric Tests

After data normalization, the m normalized contact count matrices $\{C^k \in N^{n\times n}, k = 1, \cdots, m\}$ were generated. Here, c_{ij}^k represents the normalized interaction count between loci i and j in the k-th experiment (for $1 \leq i, j \leq n$ and $1 \leq k \leq m$). Each experiment is done in one of two conditions $\mathbf{A} = \{condition1, condition2\}$. We denote by $\rho(k) \in \mathbf{A}$ as the condition corresponding to the k-th contact count matrix, for $1 \leq k \leq m$. For *condition* $A \in \mathbf{A}$, $A = 1$ or 2, we denote m_A as the number of contact count matrices belonging to *condition* A, in particular it holds that $m_1 + m_2 = m$.

To identify the difference of interaction counts for a specific bin pair between two conditions, it is equivalent to compare the empirical distributions of bin pair counts from different conditions. Here, we consider two non-parametric methods: Kolmogorov–Smirnov test and Cramér-von Mises test. The null hypothesis for the two tests is that the two groups are drawn from the same distribution. The premise of these non-parametric test is that when the cells are drawn from two populations with the same distribution, the test should result in a high p-value; otherwise, if two groups are drawn from different population of cells, then the resulting p-value should be low.

Kolmogorov–Smirnov test

In Kolmogorov-Smirnov (KS) test, it captured the maximum absolute difference (L1 norm) between the empirical cumulative distribution function (ECDF) of two populations. That is, for a specific test in the bin pair (i, j), given that the first condition has normalized read counts $c_{ij}^{(1,1)}, ..., c_{ij}^{(m_1,1)}$ with the ECDF of $F_1(x)$ and the second condition has normalized interactions $c_{ij}^{(1,2)}, ..., c_{ij}^{(m_2,2)}$ with the ECDF of $F_2(x)$. Define

$$D_{m_1,m_2} = \max_x |F_1(x) - F_2(x)| .$$

When m_1 and m_2 are large enough, the null hypothesis that two samples are drawn from the same distribution is rejected at significant level α if

$$D_{m_1,m_2} > c(\alpha)\sqrt{\frac{1}{m_1} + \frac{1}{m_2}}$$

where $c(\alpha) = \sqrt{-ln(\frac{\alpha}{2}) \times \frac{1}{2}}$ by [35].

Thus, the p-value associated with the null hypothesis was calculated as:

$$p(D_{m_1,m_2}) = 2e^{\frac{2m_1m_2}{m_1+m_2}D_{m_1,m_2}^2}.$$

Cramér-von Mises test

In the Cramér-von Mises (CVM) test [2], it improved on KS test using the full joint sample and compared two ECDFs by looking at the sum of the squared differences between them instead of maximum distance. That is,

$$T = \frac{m_1m_2}{m_1 + m_2} \int_{-\infty}^{\infty} (F_1(x) - F_2(x))^2 dF_{1+2}(x)$$

where $F_1(x)$ and $F_2(x)$ denote the ECDFs of normalized read counts of bin pair (i, j) in two conditions respectively and $F_{1+2}(x)$ is the ECDFs of all normalized read counts of bin pair (i, j) in both conditions.

Suppose that $r_1, ...r_l, .., r_{m_1}$ and $s_1, ..., s_k, ..., s_{m_2}$ represent the ranks of the read counts from the two conditions, in the ordered pooled sample, then the statistic T could be rewritten as:

$$T = \frac{U}{m_1 m_2 (m_1 + m_2)} - \frac{4m_1 m_2 - 1}{6(m_1 + m_2)}$$

where

$$U = m_1 \sum_{l=1}^{m_1} (r_l - l)^2 + m_2 \sum_{k=1}^{m_2} (s_k - k)^2.$$

Then the p-value associated with the null-hypothesis that two samples are drawn from the same distribution was computed as

$$p(T) = 1 - \frac{1}{\pi\sqrt{T}} \sum_{w=0}^{\infty} \frac{\Gamma(w + 0.5)}{\Gamma(0.5)w!} (4w + 1)^{0.5} e^{\frac{-(4w+1)^2}{16T}} K_{0.25} \frac{(4w + 1)^2}{16T}$$

where $\Gamma(z)$ is Euler's Gamma function, and $K_v(z)$ is a modified Bessel function of the second kind.

3.2.3 Detecting Differential Interactions by Parametric Models

Although non-parametric tests are advantageous since it allows us to detect DCIs in any single-cell dataset without prior assumption on the data distribution, parametric models are more common approaches in most comparative analysis by customizing their models suitable for specific cases.

Negative Binomial Model

According to previous studies, Negative Binomial (NB) distribution model is the most widely used for differential interaction detection in bulk Hi-C data. Thus, we apply this model in scHi-C data differential analysis as a reference. Here, we assume the read count C_{ij} of bin pair $(i,\ j)$ follows a negative binomial distribution with mean μ_{ij} and dispersion α_{ij}, then probability mass function (PMF) of C_{ij} is shown as:

$$P(C_{ij} = c_{ij}) = f(c_{ij}) = \frac{\Gamma(c_{ij} + (\alpha_{ij})^{-1})}{c_{ij}!\Gamma((\alpha_{ij})^{-1}))}\left(\frac{(\alpha_{ij})^{-1}}{\mu_{ij} + (\alpha_{ij})^{-1}}\right)^{(\alpha_{ij})^{-1}}\left(\frac{\mu_{ij}}{\mu_{ij} + (\alpha_{ij})^{-1}}\right)^{c_{ij}}$$

When it comes to identify differential interactions between two conditions, it is equivalent to testing the heterogeneity of two populations. Each population is characterized by a negative binomial model with parameters. When any of the parameters of two models has significant difference, the bin pair can be considered as differential.

Specifically, for bin pair (i, j), to test the difference of the two populations, we followed the three steps:

(1) Calculate the Maximum Likelihood Estimation (MLE) of the two NB populations' parameters $\hat{\Theta}_{ij,1} = \{\hat{\mu}_{ij,1}, \hat{\alpha}_{ij,1}, \hat{\mu}_{ij,2}, \hat{\alpha}_{ij,2}\}$ with Expectation-Maximization (EM) algorithm for each condition respectively.

(2) Calculate the constrained MLE of the two NB populations' parameters $\hat{\Theta}_{ij,0} = \{\hat{\mu}_{ij,0}, \hat{\alpha}_{ij,0}\}$ under the null hypothesis $H_0 : \mu_{ij,1} = \mu_{ij,2}, \alpha_{ij,1} = \alpha_{ij,2}$. It is equivalent to calculate the unconstrained MLE using pooled condition data together.

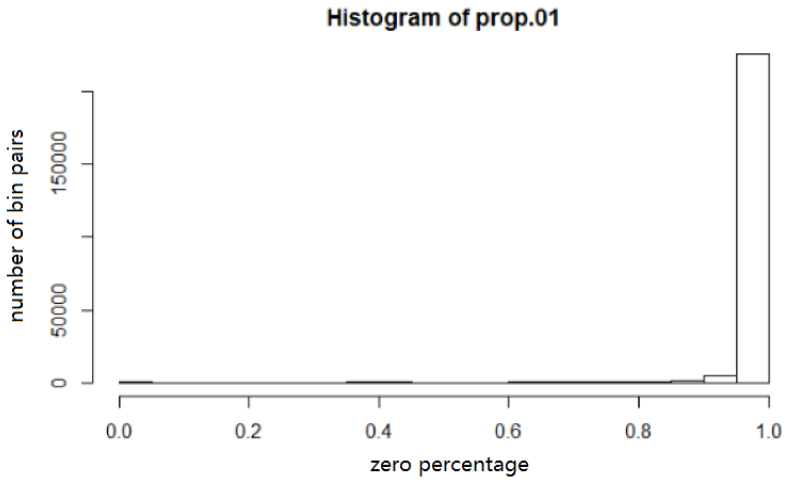

Figure 3.1: **Histogram of zero percentages of read counts for all bin pairs in a scHi-C mouse Diploid ESC dataset**

(3) Hypothesis testing of H_0. Under the null hypothesis H_0, the statistics χ^2_{LR} follows a χ^2_2 distribution,

$$\chi^2_{LR} = -2\left[l(\hat{\Theta}_{ij,0}|m) - l(\hat{\Theta}_{ij,1}|m)\right] \sim \chi^2_2,$$

where $l(\Theta_{ij}|m)$ is the log-likelihood function. Then the hypothesis testing of H_0 is conducted using the χ^2_{LR} statistics.

However, the extreme sparsity appears in most bin pairs of single-cell Hi-C data (Figure 3.1). When the read count data of interest have excessive zeros, this may result in over-dispersion problem. However, negative binomial model is not designed for dealing with excess sparse observations, and cannot fully explain those excess zeros. Therefore, we seek to adopt the zero-inflated Negative Binomial and Negative Binomial Hurdle model instead of negative binomial model to describe the read counts and the excessive zeros.

Zero-inflated Negative Binomial Model

The zero-inflated Negative Binomial (ZINB) distribution is a mixture of constant zeros and a negative binomial distribution with a mixing parameter p. That is, the zero counts of ZINB model come from two populations: always zero set and negative binomial distribution, and the non-zero reads only come from negative binomial distribution. The PMF of ZINB distribution for read counts C_{ij} of bin pair (i, j) in a group of cells is defined as:

$$P(C_{ij} = c_{ij}) = \begin{cases} p_{ij} + (1 - p_{ij})f(0) & c_{ij} = 0 \\ (1 - p_{ij})f(c_{ij}) & c_{ij} > 0 \end{cases}$$

where $f(z)$ is the density function of negative binomial model with mean parameter μ_{ij} and dispersion parameter α_{ij}, and p_{ij} is the mixing parameter of constant zeros.

Negative Binomial Hurdle Model

The Negative Binomial Hurdle (NBH) model[48] is also a two-part model, which assumes that zeros and positive read counts come from different processes. The first part models binary responses: zero and positive count data via mixture parameter p. The second part models the non-zero count data (greater than 0) with truncated negative binomial distribution.

Then the PMF of NBH model for read counts C_{ij} of bin pair (i, j) in a group of cells is defined as:

$$P(C_{ij} = c_{ij}) = \begin{cases} p_{ij} & c_{ij} = 0 \\ \frac{1-p_{ij}}{1-f(0)} f(c_{ij}) & c_{ij} > 0 \end{cases}$$

where $f(z)$ is the density function of negative binomial model with mean parameter μ_{ij} and dispersion parameter α_{ij}, and p_{ij} is the mixture parameter of zero counts.

When it comes to test on the difference of the two ZINB/NBH populations, we performed similar steps as in the NB model;

(1) Calculate the Maximum Likelihood Estimation (MLE) of the two ZINB/NBH populations' parameters $\hat{\Theta}_{ij,1} = \{\hat{p}_{ij,1}, \hat{\mu}_{ij,1}, \hat{\alpha}_{ij,1}, \hat{p}_{ij,2}, \hat{\mu}_{ij,2}, \hat{\alpha}_{ij,2}\}$ with Expectation Maximization (EM) algorithm for each condition respectively.

(2) Calculate the constrained MLE of the two ZINB/NBH populations' parameters $\hat{\Theta}_{ij,0} = \{\hat{p}_{ij,0}, \hat{\mu}_{ij,0}, \hat{\alpha}_{ij,0}\}$ under the null hypothesis $H_0 : p_{ij,1} = p_{ij,2}, \mu_{ij,1} = \mu_{ij,2}, \alpha_{ij,1} = \alpha_{ij,2}$. It is equivalent to calculate the unconstrained MLE using pooled condition data together.

(3) Hypothesis testing of H_0. Under the null hypothesis H_0, the statistics χ^2_{LR} follows a χ^2_3 distribution,

$$\chi^2_{LR} = -2\left[l(\hat{\Theta}_{ij,0}|m) - l(\hat{\Theta}_{ij,1}|m)\right] \sim \chi^2_3$$

Note that all differential detection frameworks account for the multiple testing correction using the FDR procedure[3].

3.2.4 Simulation Setting

To assess the performance of non-parametric and parametric methods in scHiCDiff, we made a variety of comparisons on simulated data. First, we used the scHi-C dataset from diploid ESCs cultured with 2i in Nagano et al. [51] as the base for data simulation. Here, we designated a series of simulation data according to the following steps:

(1). Merge the data from k-sampled single cells to conduct an ensemble Hi-C map;

(2). Generate a pair of pseudo-bulk dataset (one with a pre-specified differential contact, one similar to original pseudo-bulk dataset);

(3). Simulate k single-cell dataset by downsampling from each pseudo-bulk dataset. Specifically, in the step 2, to simulate the contact frequencies, we use the merged k-sample scHi-C contact matrix (like bulk Hi-C data) as a reference. For each pairwise interaction $(i,\ j)$, we utilize a negative-binomial distribution with a dispersion of 1000 using the R function `rnbinom`. The non-differential interactions are sampled from a negative binomial with a mean equal to the value of the corresponding pairwise interaction in the bulk count matrix, whereas the differential interactions are sampled from a negative binomial with a mean equal to the fold change of their corresponding pairwise interaction in the merged Hi-C contact map. We try to make the simulated DCIs as sparsely distributed as possible by selecting a small number of interactions to be DCIs (approximately 1%). All these DCIs show an increase in their interaction count with a given fold-change value. Then in the final step, the contact frequencies are simulated by accounting for both total read counts and respective pairwise contact frequencies of original scHi-C matrices. (Here, we set the weights of these two factors with 0.1 and 0.9 respectively to insure the similarity of simulated data compared to original one.)

Next, we conducted a series of simulation experiments by varying the factors with respect to the three aspects: (1) fold change, i.e., the discrepancy degree in read counts between two conditions. Here, we set the fold change with 2, 5 and 10; (2) resolution, which evaluate the sensitivity of our tool on data sparsity level and test from 200kb to

1Mb; and (3) sample number for each condition, which is set ranging from 20 to 100. For each simulation factor setting, we generated 20 iterations to evaluate the performance. At the default setting, we predefined the fold change of 5 in read counts of differential local pairs under the resolution of 200kb and sample size of 50 in each condition. Then, we changed one of the factors (fold change, resolution or sample number) each time for comparison.

As the ROC curve (receiver operating characteristic curve) is a widely used tool to assess the performance of a classification model by plotting the true positive rate (TPR) against the false positive rate (FPR) at various classification threshold [20], we further performed ROC analyses on simulated data to diagnose the methods in scHiCDiff by utilizing `ROCR` package in R. In this case, the true (false) positive signals are the regions simulated to be (not to be) DCIs but reported as DCIs by the algorithm while the true (false) negatives are the regions (not) DCIs in the simulation that are identified as non-DCIs.

3.2.5 Real Data Pre-processing

To assess the reliability of our algorithm on real data, we did comparative analysis on Flyamer et al. dataset [23] and Kim et al. dataset [33]. For Flyamer et al. data, it derived from mouse oocytes and zygotes (GEO accession: GSE80280) with resolution 200kb. For quality control, we firstly ruled out the cells with less than 5k non-diagonal contacts, resulting in 143 remaining cells (89 oocyte cells and 54 zygote cells). The data were then normalized using scHiCNorm. For the subsequent data processing, we only consider cells with more than 100 non-zero bin pairs for each chromosome (except for chrX) after scHiCNorm normalization, because the cells with few non-zero counts are too sparse to be reasonably adjusted by genomic distance effect. Consequently, we applied scHiCDiff

to 86 oocyte cells and 34 zygote cells for differential interactions detection and analyses. As to Kim et al. dataset, we compared the 500-kb single-cell Hi-C data getting from human H1ESC and GM12878. With the same data processing and filtering procedure, 150 H1ESC cells and 120 GM12878 cells were utilized for differential interactions analysis.

In addition to Hi-C contact maps, ChIP-seq data were applied to evaluate the results of Kim et al. data [33] analysis. CTCF and other transcription factors (including RAD21, EP300, POLR2A and H3K4me3) ChIP-seq datasets were obtained from ENCODE project [6]. The ChIP-seq peak files were in narrowpeak BED format. The ChIP-seq peaks were aggregated into bins with 500kb and the bin-wise peak counts were normalized by the total number of peaks in each ChIP-seq dataset. The absolute mean differences of the normalized bin-wise peak counts were calculated for each pair of cell lines for the subsequent analysis.

3.3 Results

3.3.1 scHiCDiff successfully detected differential chromatin interactions in simulated data

We first tested simulated data under the default setting with respect to their performance on two different data pre-processing procedures. From Figure 3.2, compared with results with scHiCNorm normalization only, normalization with additional genomic distance adjustment showed equivalent results in non-parametric tests, while parametric

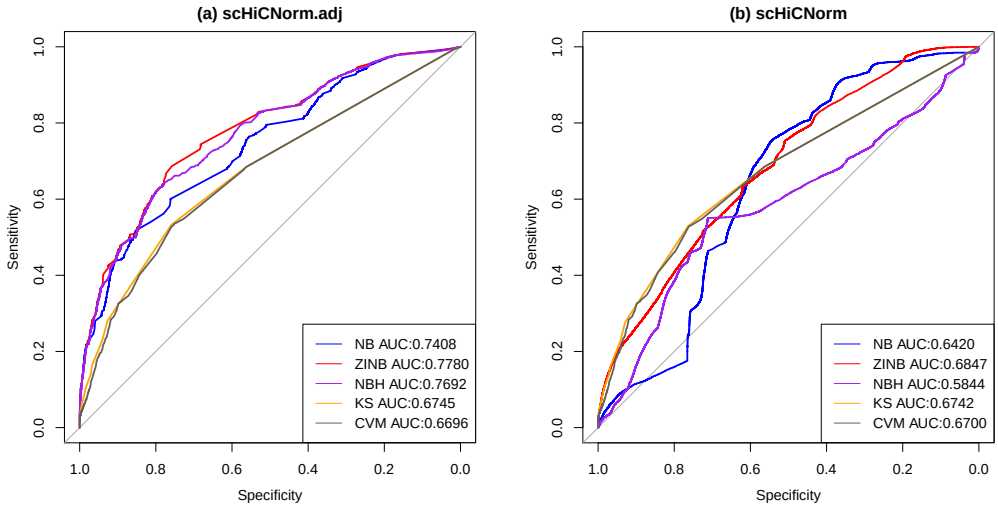

Figure 3.2: **Comparison of ROC curves on the simulated scHi-C data with two different normalization pre-processing.** ROC curves of 5 differential analysis methods on the simulated scHi-C data with two different normalization preprocessing ways: (a) scHiCNorm with genomic distance adjustment, (b) scHiCNorm only. The corresponding AUC (area under the ROC curve) values of ROC curves were shown at the bottom right corner of plots. 20 simulations were generated with fold change=5, resolution=200kb and sample size per condition=50.

models with additional genomic distance adjustment was superior in performance, implying that the necessity of eliminating the effect of genomic distance on scHi-C data.

Further, to validate the efficiency of our scHiCDiff algorithm, we evaluated their performance on three different aspects: fold change, resolution and sample number. Overall, the ROC curve analysis showed that parametric models outperformed non-parametric tests (Figure 3.3). One possible explanation of this phenomenon is that non-parametric tests were lacking of consideration on the special properties of scHi-C data. In addition, the curves demonstrated the superiority of the models specifically designed for zero-inflated Hi-C data (ZINB and NBH) in improving the power of detecting differential chromatin interactions in single-cell Hi-C data (Figure 3.3). Typically, Hi-C data at finer resolution (smaller size of

chromatin regions tested for counts) have a higher proportion of zero interaction frequencies (sparsity). The benefits of ZINB and NBH models were more pronounced at higher resolutions with more obvious discrepancy on AUC values(Figure 3.3(g)-(i)), confirming our observation of the poor performance of Negative Binomial model in handling excessive sparsity problem appearing in scHi-C data. When it came to fold change, we simulated scHi-C matrices with various levels of count discrepancy (2, 5 and 10) in predefined DCI locations between two conditions. As expected, scHiCDiff was able to detect the majority of the introduced differences with relatively low numbers of false positives, and the power of detecting differential interactions increased dramatically as the fold change increased (Figure 3.3 (a)-(c)). In addition, we checked the influence of sample size on the detection of DCIs with different methods in scHiCDiff. The ROC curves demonstrated the increase in power in identifying differential chromosome interactions as the number of sample per experimental condition increased from 20 to 50 (Figure3.3 (d)-(e)). Surprisingly, no significant improvement of performance on these models appeared when sample size per condition increases from 50 to 100 (Figure3.3 (e)-(f)). This is likely due to the sufficiency of data contained in both conditions leading to the stability of outputs on differential interaction detection.

3.3.2 scHiCDiff revealed cell type-specific differential chromatin interactions

After assessing our algorithm on simulated datasets, we further applied scHiCDiff to two published scHi-C datasets: Flyamer et al. dataset[23] and Kim et al. dataset[33].

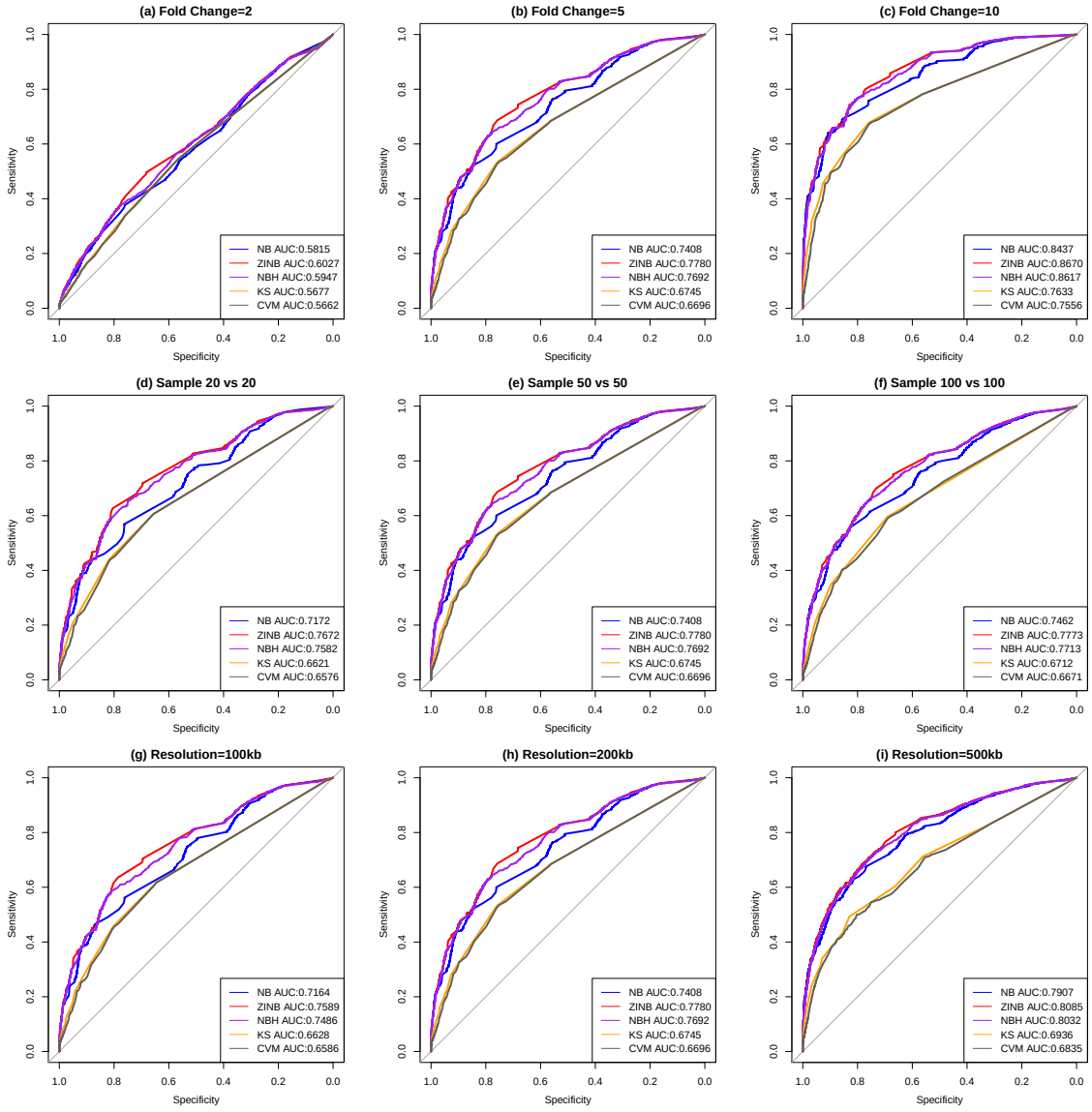

Figure 3.3: **ROC curves and AUCs of 5 differential analysis methods on the simulated scHi-C data.** The AUC of each model is listed on bottom right corner of each graph. In the default setting, each set generated 20 simulations with fold change=5, resolution=200kb and sample size for each condition=50. Then, one of the factors is altered each time for comparison (The altered factor is annotated above each graph). (a)-(c) Comparison across different fold changes. (d)-(f) Comparison across different sample sizes. (g)-(i) Comparison across different resolutions.

Table 3.1: Total number of detected differential contact interactions in oocyte and zygote cells comparison.

	KS	CVM	NB	ZINB	NBH
oocyte (NSN vs SN)	0	0	349	18	46
oocyte vs zygote	11626	31649	36633	17410	17878

In Flyamer et al. dataset, we conducted one comparison of oocyte cells within different conditions and the other comparison between oocyte and zyogote cells. We first sought to evaluate the performance of our algorithm between 23 active immature oocytes (non-surrounded nucleolus, NSN) and 60 transcriptionally inactive mature (surrounded-nucleolus, SN) oocytes. In theory, high degree of similarity between cells can be observed in cells from the same cell type. As expected, few differential chromatin interactions were detected by our algorithm (Table 3.1). Specifically, none of differential chromatin interactions were identified by two non-parametric tests, while 349, 18 and 46 bin pairs were detected as differential (with P-value < 0.05) across the whole genome (except for ChrX) by NB, ZINB and NBH models respectively, indicating that significant changes in interactions seldom appear between single cells from the same cell type. Next, for the interactions from oocyte-zygote cell comparisons, as illustrated in Table 3.1, 11626 and 31649 interactions were identified by KS and CVM tests respectively. As to our parametric models, it was shown that 17410 bin pairs detected by ZINB showed statistically significant differences between two conditions; 17878 of the NBH-identified interaction regions were determined to be differential; while the number of DCIs discovered by NB model almost doubled as high as 36633. These results confirmed that scHiCDiff can effectively detect differential interactions between different cell types.

Table 3.2: Total number of detected differential contact interactions in H1ESC and GM12878 comparison.

	KS	CVM	NB	ZINB	NBH
H1ESC (BioR1 vs BioR2)	0	0	473	0	13
H1ESC vs GM12878	3403	6639	6821	2198	2904

Besides the detection in Flyamer et al. dataset, we also performed similar tests in Kim et al. dataset. Here, we also conducted one comparison between biological replicates in H1ESC and another comparison between H1ESC and GM12878 cells. Similar features of detection results (Table 3.2) were also observed in Kim et al. dataset, demonstrating that all methods in scHiCDiff efficiently captured substantively more DCIs between different cell types than between replicates of the same cell-type.

3.3.3 Consistent detection results were found in scHiCDiff methods

Although scHiCDiff could identify many differential interactions between different cell types, the number of detected DCIs varied from model to model. To further investigate the relevance of detection results, we did ten pairwise consistency comparisons among five models in scHiCDiff. Here, we measured two types of consistent percentage for each model pair: The first proportion is calculated by the number of DCIs simultaneously appearing in the top 1% smallest P-value list of both comparison models over the number of DCIs with the top 1% P-values; the second rate is the ratio of the number of intersect over the number of union of detected DCIs whose P-values < 0.05 in two comparison models.

Overall, several consistent detection results have been shown among different model comparisons. Specifically, in Flyamer et al. dataset, the overlapping rate of de-

tected DCIs in top 1% list among parametric model comparisons were extremely high (all over 80%) and this consistent rate in KS versus CVM comparison was also considerably high (76%) (See Table 3.3). With respect to the criteria based upon DCIs with P-values < 0.05, the detected bin pairs in ZINB model was highly consistent with those identified by NBH model, while the overlapping rates for the rest model comparisons were relatively low (See Table 3.3). Due to the low coverage of Kim et al. data, the consistency performance of detection results between H1ESC and GM12878 was generally worse than those shown in oocyte and zygote comparison. However, high concordant rates between ZINB and NBH models were still witnessed in H1ESC and GM12878 comparison and similar detection results appeared between KS and CVM tests with respect to the top 1% P-value criteria (See Table 3.4). A reasonable explanation for these observed features was that 1) KS and CVM tests originated from the same idea but were measured with different distance norms in practice, which leaded to the high overlapped rate of detected DCIs in top 1% list between two non-parametric tests; 2) NB model were more likely to identify bin pairs with relatively small difference in counts, whereas ZINB and NBH could more efficiently distinguish the significant changes by taking excessive zeros into consideration.

Table 3.3: Average proportions of common detected DCIs in oocyte and zygote comparison.

	common DCIs with top 1% Pvalue rank	common DCIs with Pvalue <0.05
KS vs CVM	0.7597	0.4987
KS vs NB	0.6164	0.4328
KS vs ZINB	0.6752	0.6246
KS vs NBH	0.6790	0.6256
CVM vs NB	0.4572	0.6115
CVM vs ZINB	0.4970	0.4833
CVM vs NBH	0.5015	0.4944
NB vs ZINB	0.8158	0.4733
NB vs NBH	0.8206	0.4829
ZINB vs NBH	0.9701	0.9601

Table 3.4: Average proportions of common detected DCIs in H1ESC and GM12878 comparison.

	common DCIs with top 1% Pvalue rank	common DCIs with Pvalue <0.05
KS vs CVM	0.8002	0.4178
KS vs NB	0.4093	0.3947
KS vs ZINB	0.6291	0.4117
KS vs NBH	0.7067	0.4966
CVM vs NB	0.3358	0.3596
CVM vs ZINB	0.5075	0.2819
CVM vs NBH	0.5865	0.3414
NB vs ZINB	0.5262	0.3748
NB vs NBH	0.5298	0.3396
ZINB vs NBH	0.8623	0.7261

3.3.4 Changes in CTCF and other transcription factors were consistent with scHiCDiff detection results

As there is no ground truth of differential chromatin interactions in real data, we sought to assess the performance of scHiCDiff by investigating the association between the changes of 1D epigenomic features and 3D genomic interaction regions. As CTCF is a master controller of the chromatin architecture and other transcription factors also play important roles in gene regulation [56], we expect the DCIs are more likely to be located in the neighborhoods of these differential ChIP-seq peaks. To test this hypothesis, we compared the differential ChIP-seq peaks at the bin sites appearing within or without detected DCIs. Here, in each chromosome, we evaluated whether the mean absolute differences of the ChIP-seq signal at the bin loci appearing within detected DCIs were significantly higher (with a significant level 0.1) or higher than those appearing without detected DCIs.

A total of 22 chromosome-wide tests (except for ChrX) were conducted for models in scHiCDiff between H1ESC and GM12878 cells. As shown in Table 3.5, 13 out of 22 (59.09%) tests were significantly higher differences of CTCF values at bins appearing in DCIs detected by ZINB and NBH while such number in KS and CVM tests are 12 and that of NB model was 8. Consistent with transcription factors results, better performance of ZINB/NBH models could also be observed in the comparison of EP300 and H3K4me3 results (Table 3.5). On the strength of superior performance and high consistency on models considering excessive zeros (ZINB and NBH), we utilized the result getting from ZINB model for subsequent analysis.

Table 3.5: Agreements between CTCF data and differential chromatin interactions.

	CTCF	RAD21	EP300	POLR2A	H3K4me3
KS	12/21	12/20	13/20	7/11	6/13
CVM	12/21	12/20	13/20	7/11	6/13
NB	8/17	6/15	7/16	4/9	5/8
ZINB	13/18	11/17	14/18	7/10	7/12
NBH	13/18	11/17	14/18	7/10	7/12

A total of 22 chromosome-wide tests were conducted for each model.The number of chromosomes, whose mean absolute differential transcription peaks at the bins appearing in DCIs were significantly higher(based on t-tests with P-value<0.1) or higher than those not appearing in DCIs were shown in each block, were recorded.

3.3.5 Stable detection results were conducted by scHiCDiff

Since the differential contact interactions are inherently associated with the experimental conditions being studied, they are likely to be largely constant between conditions. To investigate the stability of scHiCDiff in differential interaction detection between conditions, we randomly selected partial cells for test. To be specific, considering that the number of qualified cells in Flaymer et al. dataset are 86 and 34 for oocytes and zygotes respectively and the simulation results indicates that the detection results are more reliable when sample size for each condition is great than 50, we retained all 34 zygote cells and repeatedly randomly chose 70 of 86 cells from oocytes with 20 times for stability test. Similar to previous consistency test, we compared the randomly selected test-data detection results with the original full-data detection results with two criteria: common detected DCIs with the top 1% smallest P-values and those with P-value < 0.05.

From Table 3.6, it was shown that the average proportion of consistent results between test datasets and original full dataset were extremely high (with 88.69% in top 1% P-value comparison and 73.25% in P-value<0.05 comparison), revealing that our ZINB

Table 3.6: Mean numbers and proportions of common detected differential contact interactions for detection stability verification.

	oocyte(70) vs oocyte(86)	
	Mean Num.	Mean Prop.
common DCIs in top 1% Pvalue rank	725.55	0.8869
common DCIs with P-value <0.05	746.13	0.7325

model could steadily output significant differential contact interactions between different cell types.

3.3.6 scHiCDiff detection results were supported by TAD and DiffGR results

Several bulk Hi-C comparative studies indicated that the majority of the chromatin structural changes strongly correlated to topologically associated domains (TADs) [56, 15, 62]. Additionally, existing scHi-C papers mentioned that variable contact clusters averaged into population TADs when pooled together and the locations of detected TAD borders were generally unchanged in the pools of single-cell data from different cell cycles ([23, 51]).

In Flaymer et al. dataset, because of the lack of bulk Hi-C data for oocytes and zygotes, we merged their single cells separately and called for TAD boundaries of their pooled matrices by HiCseg[38]. Then, we measured the proportion of DCIs located within TADs (Fig 3.4). The results indicated that majority of the DCIs detected by scHiCDiff ZINB model were located within TADs (87.00% and 54.18% of detected DCIs inside TADs of oocyte and zygote respectively). Additionally, considering chromatin structures exhibit

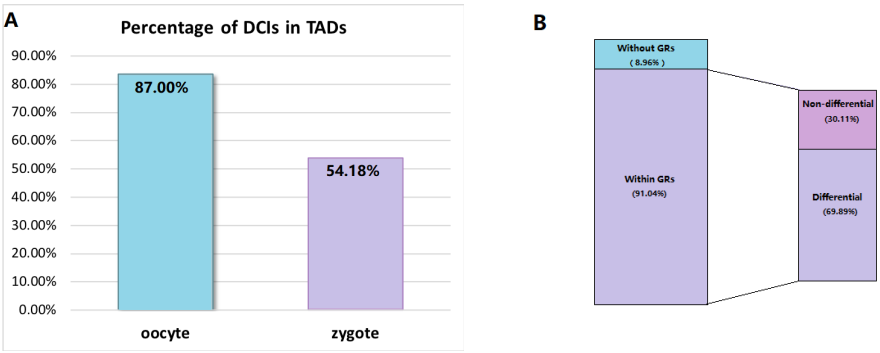

Figure 3.4: **Performance of ZINB model on DCI Detection of single-nucleus oocyte and zygote cells.** (A) The proportion of the DCIs located inside population TADs. (B) The proportion of the DCIs within genomic regions at TAD level and the percentage belonging to differential genomic regions.

differences between different cell types of real data in their TADs, some DCIs outside TADs were still worthy of note. Thus, we further figured out genomic regions between pooled oocyte and zygote maps at TAD levels by DiffGR. It was shown that 91.04% of single-cell DCIs were situated within candidate genomic regions and 69.89% among them were classified into differential genomic regions.

Later, bulk Hi-C data for H1ESC [16] and GM12878 [56] were utilized to assign TAD boundaries and genomic regions. Similarly, we calculated the proportion of DCIs located within TADs, showing that 87.64% in H1ESC TADs and 86.77% in GM12878 TADs. The following DiffGR results are also consistent with those getting from Flyamer et al. data, reporting that 90.70% of single-cell DCIs were within genomic regions and 57.69% among them belonged to differential genomic regions. All of these results suggested that scHiCDiff can reliably detect DCIs at single-cell level.

Table 3.7: Functional enrichment of genes located within the DCI sites

Go Term	P_value
GO:0042100 B cell proliferation	7.9E-6
GO:0002285 Lymphocyte activation involved in immune response	1.8E-3
GO:0030183 B cell differentiation	4.1E-3
GO:0001959 Regulation of cytokine-mediated signaling pathway	7.2E-3
GO:0002366 Leukocyte activation involved in immune response	4.8E-2

3.3.7 GO term enrichment analysis confirmed the function of DCIs identifying by scHiCDiff

To further explore the potential functional roles of the changes in interactions on gene expression regulation, we attempted to pick out the most influenced bin sites, which appeared more than two times in DCIs detected by scHiCDiff ZINB model between H1ESC and GM12878 comparison, and then performed Gene Ontology (GO) enrichment analysis on the genes located within these loci (1145 genes getting from 656 unique bin sites) using DAVID [60].

The functional analysis of the 1145 genes showed a high enrichment for Biological Process (BP) GO terms of B cell proliferation and differentiation, lymphocyte activation and leukocyte activation involved in immune response, and regulation of cytokine-mediated signaling pathway (Table 3.7), which all related to the immune system. In addition, these findings were concordant with previous researches that GM12878 cell line belongs to B lymphoblast cells with known autoimmune disorder SNPs [8], further supporting the reliability of our scHiCDiff detection results.

3.4 Discussion

The increasing availability of scHi-C data opens the door to investigate the principles that govern the spatial organization of the chromatin between different species and cell types at single-cell level. However, with the lack of single-cell differential chromatin interaction detection tools, it is hard to make significant conclusions.

Here, we have presented scHiCDiff, a tool applying both non-parametric tests(CVM/ KS) and parametric models(NB/ ZINB/ NBH) to identify DCIs between two conditions at single-cell Hi-C level. Specially, we introduced zero-inflated Negative Binomial(ZINB) /Negative Binomial Hurdle (NBH) regression models to fit the feature of excessive zeros in scHi-C matrices and performed a rigorous likelihood ratio test to figure out the bin pairs showing significant changes in contact counts.

To evaluate the performance of scHiCDiff, we applied it on both simulated and real data. In simulation studies, we showed that ZINB/NBH models outperformed other three approaches with respect to different fold changes, sample sizes and resolutions. The superiority and reliability of ZINB/NBH detection results in real data were also validated by the ChIP-seq data and GO term analysis and shown highly consistency and stability compared to non-parametric tests and NB model. To summarize, ZINB and NBH models produced more accurate and stable detection of differential chromatin interactions, while commonly used NB model and non-parametric tests might be prone to errors in detection without taking the extreme sparsity characteristic of scHi-C contact maps into account.

However, more room remains for improvement; for example, the tool can be extended to allow comparisons among more than two conditions. It would also be beneficial

to be able to subdivide the types of differential contact interactions for further analysis of their functions.

Software availability

The software is published under the GNU GPL ≥ 2 license. The main functions in scHiCDiff are explained in the Appendix B and the source code is publicly available at https://github.com/wmalab/scHiCDiff.

Chapter 4

Conclusions

In this dissertation, we proposed two statistical comparative analyses on Hi-C contact maps: DiffGR focused on detecting differentially interacting genomic regions at the TAD level between bulk Hi-C data and scHiCDiff distinguished differential chromatin interactions from single-cell Hi-C matrices.

Unlike the existing approaches for comparative analyses at TAD level approaches, which concentrated on the detection of TAD boundary changes, DiffGR figured out the changes in chromatin organization within TADs. Specifically, the stratum-adjusted correlation coefficient (SCC) measures the similarity of local candidate genomic regions at TAD level and non-parametric tests on SCCs are developed to identify statistically differential genomic interacting regions. The main advantages of DiffGR are listed as the follows:

- Taking genomic distance features of Hi-C data into consideration, DiffGR utilizes the SCC metric instead of the standard Pearson CC to measure the similarity of local genomic regions between Hi-C contact maps.

- In contrast to the parametric approaches that were used by most Hi-C data analysis methods, our non-parametric approach does not have a set of predefined assumptions about the nature of the null distribution and, therefore, is more robust and can be applied to more diverse data from real cases.

- A non-parametric smoothing spline regression is applied to facilitate the permutation test and it was shown the speed-up algorithm could steadily conduct consistent outputs.

In our simulation studies, we conducted a series of simulation experiments on single-TAD and hierarchical TAD conditions, and evaluated how the performance of DiffGR was impacted by various factors, including the proportion of altered TADs, proportion of TAD alternation, noise level, and sequencing coverage level. Overall, we demonstrated that DiffGR can robustly and effectively discover differential genomic regions under various conditions. In real data analyses, DiffGR revealed cell type-specific changes in genomic interacting regions, which were effectively validated by the ChIP-seq and RNA-seq data and were concordant with the results of FIND and TADCompare. To summarize, DiffGR provided a statistically rigorous method for the detection of differentially interacting genomic regions in Hi-C contact maps from different cells and conditions, therefore would facilitate the investigation of their biological functions.

Currently, only a few limitations can be attributed to our DiffGR algorithm. First, owing to the application of permutation tests, the running time of DiffGR is noticeable, especially when the resolution of Hi-C maps increases. One possible solution is to further optimize the speed-up option by minimizing the number of points in smoothing spline

estimation or finding a proper parametric model to fit the curves. We could also parallelize the permutation processes with different TAD sizes to facilitate the detection. Second, our method performs pairwise comparison between Hi-C contact maps. A potential future direction is to conduct a more general statistical framework for differential analyses among three or more samples. Then we could further assign the differentially interacting genomic regions to cell type-specific or condition-specific changing areas. Third, we currently pool biological replicates together in our analyses. Extending DiffGR to incorporate multiple biological replicates to detect reproducible differences would enhance the reliability of the detection results. Lastly, our method is specifically designed for bulk Hi-C data. Given the high sparsity and variability of single-cell Hi-C contact matrices, identifying differential genomic regions at single-cell level remains a significant challenge.

Later, scHiCDiff performed comparative analyses on differentially interacting counts at single-cell level. Here, it provides a software tool identifying the differential chromatin interactions between two conditions by non-parametric tests (Kolmogorov–Smirnov test/ Cramér-von Mises test) and parametric likelihood ratio test with three regression models (Negative Binomial/ Zero-inflated Negative Binomial/ Negative Binomial Hurdle). Specifically, non-parametric tests are advantageous by allowing us detecting DCIs without any assumption on data distribution; negative binomial(NB) is the most common assumption for interaction counts in bulk Hi-C parametric approaches, while zero-inflated Negative Binomial(ZINB) and Negative Binomial Hurdle (NBH) regression models are specially designated for the interaction comparison at single-cell level by taking the excessive zeros feature into consideration.

As a result, ZINB and NBH models exceeded non-parametric tests and NB model, in terms of producing more accurate and consistent detection of differential chromatin interactions. In simulation part, it was demonstrated that parametric models outperformed non-parametric tests with respect to different fold changes, sample sizes and resolutions; the performance discrepancy between NB model and ZINB/NBH models were more pronounced at finer resolution in which excessive sparsity problem was more remarkable in scHi-C data. In real data studies, compared to non-parametric tests and NB model, the superiority and credibility of ZINB and NBH detection results were confirmed by the transcription factors and GO term enrichment analysis; highly consistency and stability between ZINB and NBH models were also been found.

However, there are a few challenges which require more efforts in the future:

(1) Although the functional roles of scHiCDiff detection results between H1Esc and GM12878 were revealed by the bulk ChIP-seq data and GO term enrichment analysis, we are looking for suitable single-cell ChIP-seq and RNA-seq to further evaluate. Moreover, due to the deficiency of more information and ChIP-seq/RNA-seq data for oocyte and zygote, the accuracy of detection results in Flaymer et al. dataset are required to explore.

(2) Since we assume mixture parameter for binary responses in NBH model to be a constant percentage, the ZINB and NBH models are equivalent in some extents, which conduct similar detection results between two models. Extending the mixture parameter to a more general density function $f_{zero}(0)$ in NBH model may enhance the robustness of detection results. Further, considering the potential influence of certain single cells on

interaction frequencies, we also tend to apply the mixed-effects models by adding cell-specific random effects to ZINB/NBH regression models.

(3) Our tool only performed differential comparison between two scHi-C populations. One potential future extension is to develop a more general framework for comparative analyses among three or more groups. Further, we can assign additional tests to distinguish the differential interactions to population-specific changes.

(4) In ZINB and NBH models, the changes in bin pairs might cause from the proportion of zero values or/and negative binomial parameters. Extending our tool to subdivide the types of differential contact interactions by patterns of differences of parameters would also be beneficial to better understand the relationship between chromatin interactions and gene expression regulation.

Bibliography

[1] Alan Agresti and Maria Kateri. *Categorical data analysis*. Springer, 2011.

[2] Theodore W Anderson. On the distribution of the two-sample cramer-von mises criterion. *The Annals of Mathematical Statistics*, pages 1148–1159, 1962.

[3] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

[4] Giacomo Cavalli and Tom Misteli. Functional implications of genome topology. *Nature structural and molecular biology*, 20(3):290, 2013.

[5] Fengling Chen, Guipeng Li, Michael Q Zhang, and Yang Chen. Hicdb: a sensitive and robust method for detecting contact domain boundaries. *Nucleic acids research*, 46(21):11239–11250, 2018.

[6] ENCODE Project Consortium et al. The encode (encyclopedia of dna elements) project. *Science*, 306(5696):636–640, 2004.

[7] Kate B Cook, Borislav H Hristov, Karine G Le Roch, Jean Philippe Vert, and William Stafford Noble. Measuring significant changes in chromatin conformation with accost. *Nucleic acids research*, 48(5):2303–2311, 2020.

[8] Olivia Corradin, Alina Saiakhova, Batool Akhtar-Zaidi, Lois Myeroff, Joseph Willis, Richard Cowper-Sal, Mathieu Lupien, Sanford Markowitz, Peter C Scacheri, et al. Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. *Genome research*, 24(1):1–13, 2014.

[9] Axel Cournac, Hervé Marie-Nelly, Martial Marbouty, Romain Koszul, and Julien Mozziconacci. Normalization of a chromosomal contact map. *BMC genomics*, 13(1):436, 2012.

[10] 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, 2015.

[11] Kellen G Cresswell and Mikhail G Dozmorov. Tadcompare: An r package for differential and temporal analysis of topologically associated domains. *Frontiers in Genetics*, 11:158, 2020.

[12] Darren A Cusanovich, Riza Daza, Andrew Adey, Hannah A Pliner, Lena Christiansen, Kevin L Gunderson, Frank J Steemers, Cole Trapnell, and Jay Shendure. Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science*, 348(6237):910–914, 2015.

[13] Mihails Delmans and Martin Hemberg. Discrete distributional differential expression (d 3 e)-a tool for gene expression analysis of single-cell rna-seq data. *BMC bioinformatics*, 17(1):1–13, 2016.

[14] Jesse R Dixon, David U Gorkin, and Bing Ren. Chromatin domains: the unit of chromosome organization. *Molecular cell*, 62(5):668–680, 2016.

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

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

[17] Mohamed Nadhir Djekidel, Yang Chen, and Michael Q Zhang. Find: differential chromatin interactions detection using a spatial poisson process. *Genome research*, 28(3):412–422, 2018.

[18] Josée Dostie, Todd A Richmond, Ramy A Arnaout, Rebecca R Selzer, William L Lee, Tracey A Honan, Eric D Rubio, Anton Krumm, Justin Lamb, Chad Nusbaum, et al. Chromosome conformation capture carbon copy (5c): a massively parallel solution for mapping interactions between genomic elements. *Genome research*, 16(10):1299–1309, 2006.

[19] Z. Duan, M. Andronescu, K. Schutz, S. McIlwain, Y. J. Kim, C. Lee, J. Shendure, S. Fields, C. A. Blau, and W. S. Noble. A three-dimensional model of the yeast genome. *Nature*, 465(7296):363–367, 2010.

[20] Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

[21] Darya Filippova, Rob Patro, Geet Duggal, and Carl Kingsford. Identification of alternative topological domains in chromatin. *Algorithms for Molecular Biology*, 9(1):14, 2014.

[22] Elizabeth H Finn, Gianluca Pegoraro, Hugo B Brandão, Anne-Laure Valton, Marlies E Oomen, Job Dekker, Leonid Mirny, and Tom Misteli. Extensive heterogeneity and intrinsic variation in spatial genome organization. *Cell*, 176(6):1502–1515, 2019.

[23] Ilya M Flyamer, Johanna Gassler, Maxim Imakaev, Hugo B Brandão, Sergey V Ulianov, Nezar Abdennur, Sergey V Razin, Leonid A Mirny, and Kikuë Tachibana-Konwalski. Single-nucleus hi-c reveals unique chromatin reorganization at oocyte-to-zygote transition. *Nature*, 544(7648):110–114, 2017.

[24] Mattia Forcato, Chiara Nicoletti, Koustav Pal, Carmen Maria Livi, Francesco Ferrari, and Silvio Bicciato. Comparison of computational methods for hi-c data analysis. *Nature methods*, 14(7):679, 2017.

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

[26] Geoffrey Fudenberg and Leonid A Mirny. Higher-order chromatin structure: bridging physics and biology. *Current opinion in genetics & development*, 22(2):115–124, 2012.

[27] Johan H Gibcus and Job Dekker. The hierarchy of the 3d genome. *Molecular cell*, 49(5):773–782, 2013.

[28] Zhijun Han and Gang Wei. Computational tools for hi-c data analysis. *Quantitative Biology*, 5(3):215–225, 2017.

[29] Sven Heinz, Christopher Benner, Nathanael Spann, Eric Bertolino, Yin C Lin, Peter Laslo, Jason X Cheng, Cornelis Murre, Harinder Singh, and Christopher K Glass. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and b cell identities. *Molecular cell*, 38(4):576–589, 2010.

[30] Ming Hu, Ke Deng, Siddarth Selvaraj, Zhaohui Qin, Bing Ren, and Jun S Liu. Hicnorm: removing biases in hi-c data via poisson regression. *Bioinformatics*, 28(23):3131–3133, 2012.

[31] 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. *Nature methods*, 9(10):999–1003, 2012.

[32] R. Kalhor, H. Tjong, N. Jayathilaka, F. Alber, and L. Chen. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nature biotechnology*, 30(1):90–98, 2012.

[33] Hyeon-Jin Kim, Galip Gürkan Yardımcı, Giancarlo Bonora, Vijay Ramani, Jie Liu, Ruolan Qiu, Choli Lee, Jennifer Hesson, Carol B Ware, Jay Shendure, et al. Capturing cell type-specific chromatin compartment patterns by applying topic modeling to single-cell hi-c data. *PLoS Computational Biology*, 16(9):e1008173, 2020.

[34] Philip A Knight and Daniel Ruiz. A fast algorithm for matrix balancing. *IMA Journal of Numerical Analysis*, 33(3):1029–1047, 2013.

[35] Donald E Knuth. *Art of computer programming, volume 2: Seminumerical algorithms*. Addison-Wesley Professional, 2014.

[36] Xiang Kong, Varun Gangal, and Eduard Hovy. Scde: Sentence cloze dataset with high quality distractors from examinations. *arXiv preprint arXiv:2004.12934*, 2020.

[37] Bryan R Lajoie, Job Dekker, and Noam Kaplan. The hitchhiker's guide to hi-c analysis: practical guidelines. *Methods*, 72:65–75, 2015.

[38] Celine Lévy-Leduc, Maud Delattre, Tristan Mary-Huard, and Stephane Robin. Two-dimensional segmentation for analyzing hi-c data. *Bioinformatics*, 30(17):i386–i392, 2014.

[39] Guoliang Li, Melissa J Fullwood, Han Xu, Fabianus Hendriyan Mulawadi, Stoyan Velkov, Vinsensius Vega, Pramila Nuwantha Ariyaratne, Yusoff Bin Mohamed, Hong-Sain Ooi, Chandana Tennakoon, et al. Chia-pet tool for comprehensive chromatin interaction analysis with paired-end tag sequencing. *Genome biology*, 11(2):R22, 2010.

[40] Xinjun Li, Fan Feng, Wai Yan Leung, and Jie Liu. schictools: a computational toolbox for analyzing single-cell hi-c data. *bioRxiv*, page 769513, 2020.

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

[42] Jie Liu, Dejun Lin, Galip Gürkan Yardımcı, and William Stafford Noble. Unsupervised embedding of single-cell hi-c data. *Bioinformatics*, 34(13):i96–i104, 2018.

[43] Tong Liu and Zheng Wang. schicnorm: a software package to eliminate systematic biases in single-cell hi-c data. *Bioinformatics*, 34(6):1046–1047, 2018.

[44] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15(12):1–21, 2014.

[45] Aaron TL Lun and Gordon K Smyth. diffhic: a bioconductor package to detect differential genomic interactions in hi-c data. *BMC bioinformatics*, 16(1):258, 2015.

[46] Wenxiu Ma, Ferhat Ay, Choli Lee, Gunhan Gulsoy, Xinxian Deng, Savannah Cook, Jennifer Hesson, Christopher Cavanaugh, Carol B Ware, Anton Krumm, et al. Fine-scale chromatin interaction maps reveal the cis-regulatory landscape of human lincrna genes. *Nature methods*, 12(1):71–78, 2015.

[47] Zhun Miao, Ke Deng, Xiaowo Wang, and Xuegong Zhang. Desingle for detecting three types of differential expression in single-cell rna-seq data. *Bioinformatics*, 34(18):3223–3224, 2018.

[48] John Mullahy. Specification and testing of some modified count data models. *Journal of econometrics*, 33(3):341–365, 1986.

[49] Maxwell R Mumbach, Adam J Rubin, Ryan A Flynn, Chao Dai, Paul A Khavari, William J Greenleaf, and Howard Y Chang. Hichip: efficient and sensitive analysis of protein-directed genome architecture. *Nature methods*, 13(11):919–922, 2016.

[50] Takashi Nagano, Yaniv Lubling, Tim J Stevens, Stefan Schoenfelder, Eitan Yaffe, Wendy Dean, Ernest D Laue, Amos Tanay, and Peter Fraser. Single-cell hi-c reveals cell-to-cell variability in chromosome structure. *Nature*, 502(7469):59–64, 2013.

[51] Takashi Nagano, Yaniv Lubling, Csilla Várnai, Carmel Dudley, Wing Leung, Yael Baran, Netta Mendelson Cohen, Steven Wingett, Peter Fraser, and Amos Tanay. Cell-cycle dynamics of chromosomal organization at single-cell resolution. *Nature*, 547(7661):61–67, 2017.

[52] Jennifer E Phillips and Victor G Corces. Ctcf: master weaver of the genome. *Cell*, 137(7):1194–1211, 2009.

[53] V. Ramani, D. A. Cusanovich, R. J. Hause, W. Ma, R. Qiu, X. Deng, C. A. Blau, C. M. Disteche, W. S. Noble, J. Shendure, and Z. Duan. Mapping 3D genome architecture through in situ DNase Hi-C. *Nature protocols*, 11(11):2104–2121, 2016.

[54] Vijay Ramani, Xinxian Deng, Ruolan Qiu, Kevin L Gunderson, Frank J Steemers, Christine M Disteche, William S Noble, Zhijun Duan, and Jay Shendure. Massively multiplex single-cell hi-c. *Nature methods*, 14(3):263–266, 2017.

[55] Vijay Ramani, Xinxian Deng, Ruolan Qiu, Choli Lee, Christine M Disteche, William S Noble, Jay Shendure, and Zhijun Duan. Sci-hi-c: a single-cell hi-c method for mapping 3d genome organization in large number of single cells. *Methods*, 170:61–68, 2020.

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

[57] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.

[58] Michael EG Sauria and James Taylor. Quasar: quality assessment of spatial arrangement reproducibility in hi-c data. *BioRxiv*, page 204438, 2017.

[59] 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, 2012.

[60] Brad T Sherman, Richard A Lempicki, et al. Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nature protocols*, 4(1):44, 2009.

[61] 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). *Nature genetics*, 38(11):1348–1354, 2006.

[62] Emily M Smith, Bryan R Lajoie, Gaurav Jain, and Job Dekker. Invariant tad boundaries constrain cell-type-specific looping interactions between promoters and distal elements around the cftr locus. *The American Journal of Human Genetics*, 98(1):185–201, 2016.

[63] John Stansfield and Mikhail G Dozmorov. Hiccompare: a method for joint normalization of hi-c datasets and differential chromatin interaction detection. *bioRxiv*, page 147850, 2017.

[64] John C Stansfield, Kellen G Cresswell, Vladimir I Vladimirov, and Mikhail G Dozmorov. Hiccompare: an r-package for joint normalization and comparison of hi-c datasets. *BMC bioinformatics*, 19(1):1–10, 2018.

[65] Phillippa C Taberlay, Joanna Achinger-Kawecka, Aaron TL Lun, Fabian A Buske, Kenneth Sabir, Cathryn M Gould, Elena Zotenko, Saul A Bert, Katherine A Giles, Denis C Bauer, et al. Three-dimensional disorganization of the cancer genome occurs coincident with long-range genetic and epigenetic alterations. *Genome research*, 26(6):719–731, 2016.

[66] Fuchou Tang, Catalin Barbacioru, Yangzhou Wang, Ellen Nordman, Clarence Lee, Nanlan Xu, Xiaohui Wang, John Bodeau, Brian B Tuch, Asim Siddiqui, et al. mrna-seq whole-transcriptome analysis of a single cell. *Nature methods*, 6(5):377–382, 2009.

[67] Oana Ursu, Nathan Boley, Maryna Taranova, YX 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, 2018.

[68] Junbai Wang, Xun Lan, Pei-Yin Hsu, Hang-Kai Hsu, Kun Huang, Jeffrey Parvin, Tim HM 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(1):70, 2013.

[69] Eitan Yaffe and Amos Tanay. Probabilistic modeling of hi-c contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nature genetics*, 43(11):1059, 2011.

[70] Koon-Kiu Yan, Galip Gürkan Yardımcı, Chengfei Yan, William S Noble, and Mark Gerstein. Hic-spector: a matrix library for spectral and reproducibility analysis of hi-c contact maps. *Bioinformatics*, 33(14):2199–2201, 2017.

[71] 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 research*, 27(11):1939–1949, 2017.

[72] Galip Gürkan Yardımcı, Hakan Ozadam, Michael EG Sauria, Oana Ursu, Koon-Kiu Yan, Tao Yang, Abhijit Chakraborty, Arya Kaul, Bryan R Lajoie, Fan Song, et al. Measuring the reproducibility and quality of hi-c data. *Genome biology*, 20(1):1–19, 2019.

[73] Rafal Zaborowski and Bartek Wilczynski. Difftad: Detecting differential contact frequency in topologically associating domains hi-c experiments between conditions. *bioRxiv*, page 093625, 2016.

[74] Jingtian Zhou, Jianzhu Ma, Yusi Chen, Chuankai Cheng, Bokan Bao, Jian Peng, Terrence J Sejnowski, Jesse R Dixon, and Joseph R Ecker. Robust single-cell hi-c clustering by convolution-and random-walk–based imputation. *Proceedings of the National Academy of Sciences*, 116(28):14011–14018, 2019.

Appendix A

DiffGR Source Code

DiffGR is a novel statistical method for detecting differential genomic regions at TAD level between two Hi-C contact maps. Briefly, DiffGR utilizes the stratum-adjusted correlation coefficient (SCC), which can effectively eliminate the genomic-distance effect in Hi-C data, to measure the similarity of local genomic regions between two contact matrices, and then applies a nonparametric permutation test on those SCC values to detect differential genomic regions.

A.1 Installation

The source code can be performed under R language version 4.0.2 with the installation of packages HiCcompare, HiCseg and R.utils, and it is available to download at https://github.com/wmalab/DiffGR

```
require(HiCcompare)
require(HiCseg)
require(R.utils)
```

A.2 Usage

The input arguments of the main function DiffGR are illustrated below:

Index	Description
dat1,dat2	numeric. N*N raw HiC contact maps, which would firstly be preprocessed with 2D mean filter smoothing and KR normalization in DiffGR function for the later use
tad1,tad2	numeric. A vector of TAD boundaries of contact maps. If the input is NA, the program will automatically detect the TADs by HiCseg
res	numeric. The resolution of HiC contact maps, eg:100kb will input 100,000
smooth.size	numeric. The size controlling the smoothing level (The size varies across different resolution and is guided by Hicrep paper). Here, we obtained the smoothing size with 11, 5 and 3 on real data analysis for the resolution of 25Kb, 50Kb and 100Kb respectively, and set the smoothing size with 0 in simulation.
N.perm	numeric. The number of iterations in permutation test
cutoff.default	logical. Whether set the SCC cutoff (meaningful SCC between the two Hi-C datasets that mustbe reached in order to call a differential TAD truly significant) with self-defined value(True) or with automatic computed value (False)
speedup.option	logical. Calculation with or without speed-up algorithm (True/FALSE)
alpha	numeric. Significant level of differential region testing

The function returns a list which contains a table for TAD result and one for genomic region result.

The TAD result table lists the following elements:

Index	Description
tad.start	the starting locus of TAD
tad.end	the end locus of TAD
scc	the SCC value of corresponding domain
pvalue	the pvalue of differential testing on corresponding domain
pvalue.adj	the adjusted pvalue of differential testing on corresponding domain (adjusted by Benjamin-Hochberg)

The genomic result table contains the following items:

Index	Description
genom.start	The starting locus of genomic region
genom.end	The end locus of genomic region
condition.type	The type if candidate genomic region belonging to 1:single-TAD, 2: Hierachical-TAD, 3: Alternating-TAD
detect.result	The differential testing result for corresponding genomic region. 1:Differential 0:Non-differential

A.3 Example

The raw HiC contact maps getting from chr10 of GM12878 and HMEC with resolution=50kb were untilized as sample data. An example of the usage of DiffGR with/without TAD inputs is shown below:

```
dat1 <- readRDS("path/dat.GM12878.chr10.rds")
dat2 <- readRDS("path/dat.K562.chr10.rds")
tad1 <- read.table("path/tad.GM12878.chr10.txt")
tad1 <- tad1$x
tad2 <- read.table("path/tad.K562.chr10.txt")
tad2 <- tad2$x
```

```
#with TAD inputs
result <- DiffGR(dat1=dat1,dat2=dat2,tad1=tad1,tad2=tad2,smooth.size=5,res=50000)

#without TAD inputs
result <- DiffGR(dat1=dat1,dat2=dat2,smooth.size=5,res=50000)
```

Appendix B

scHiCDiff Source Code

scHiCDiff is a novel statistical algorithm to detect differential chromatin interactions (DCIs) between two Hi-C experiments at single-cell level. Here, we introduced 5 ways to capture the DCIs: two non-parametric tests (Kolmogorov–Smirnov test/ Cramér-von Mises test) and parametric likelihood ratio test with three regression models (Negative Binomial/ Zero-inflated Negative Binomial/ Negative Binomial Hurdle). Non-parametric tests are advantageous by allowing us detecting DCIs without any assumption on data distribution; negative binomial(NB) is the most common assumption for interaction counts in bulk Hi-C parametric approaches, while zero-inflated Negative Binomial(ZINB) and Negative Binomial Hurdle (NBH) regression models are specially designated for the interaction comparison at single-cell level by taking the excessive zeros feature into consideration.

B.1 Installation

To accelerate data processing and use as less memory as possible, scHiCDiff requires the Matrix pacakges. For specific Hi-C data processing, we tend to use the HiTC, HiCcompare pacakges. For non-parametric tests, we utilize the R package twosamples to perform. In addition, to fit the regression models, we also need the R packages ggsci, VGAM etc.

Thus, with the installation of packages Matrix, mvtnorm, HiTC, HiCcompare, edgeR, ggsci, pscl, VGAM, maxLik, countreg and gamlss, the source code can be performed under R language version 4.0.2. The details about scHiCDiff source code is available at https://github.com/wmalab/scHiCDiff

```
require(Matrix)
require(mvtnorm)
require(HiTC)
require(HiCcompare)
require(edgeR)
require(ggsci)
require(pscl)
require(VGAM)
require(maxLik)
require(countreg)
require(gamlss)
require(twosamples)
```

B.2 Usage

The functions in scHiCDiff can be classified as two types: The first type is the simulation function (scHiCDiff.sim) and the other type is the detection function (scHiCDiff.KS, scHiCDiff.CVM, scHiCDiff.NB, scHiCDiff.ZINB and scHiCDiff.NBH).

Simulation Function

The inputs of the simulation function scHiCDiff.sim are illustrated below:

Index	Description
file.path	The pathway of single cell files. All scHi-C data used in simulation should be stored in this pathway. Each scHi-C file is performed as three-column format containing the first interacting region of the bin pair, the second interacting region of the bin pair and the interaction frequency of the bin pair.
fold.change	The amount of fold change.
resolution	The resolution of singel-cell HiC data, eg:200kb will input 200,000
sample.num	The number of single cells tending to generate in each condition.(¡= the number of inputted singel cells)
pDiff	The probability that an interaction will be differential.

The function returns a list that contains the simulated replicates and the matrix of the true DCI regions. The list contains the following elements:

Index	Description
Hic1.sim	A list containing the simulated scHi-C matrices of the first condition.
Hic2.sim	A list containing the simulated scHi-C matrices of the second condition.
diff.sim	A sparseMatrix containing the position of the differential interactions.

Simulation Example: The simulation test data is a dataset with 8 single-cells getting from chr1 of Diploid ESC cultured with 2i in Nagano et al. with resolution=200kb.

```
data.file <- "path/sampledata/sim.test.data"
simRes <- scHiCDiff.sim(data.file,fold.change=5,resolution=200000,sample.num=8,
          pDiff=0.01)
```

Detection Functions

The inputs for all detection functions are illustrated below:

Index	Description
count.table	A non-negative matrix of scHi-C normalized read counts. The rows of the matrix are bin pair and columns are samples/cells.
group	A vector of factor which mentions the two condition to be compared,corresponding to the columns in the count table.

The detection function will return a data frame containing the differential chromatin interaction (DCI) analysis results, rows are bin pairs and columns lists the related statistics.

The outputs for the three parametric models are listed below:

Index	Description
bin_1,bin_2	The interacting region of the bin pair.
mu_1,mu_2, theta_1,theta_2 (pi_1,pi_2)	MLE of the parameters of NB/ZINB/NBH of group 1 and group 2, where mu and theta represent the mean and dispersion estimate of negative binomial, pi denotes the estimate of zero percentage
norm_total_mean_1, norm_total_mean_2	Mean of normalized read counts of group 1 and group 2.
norm_foldChange	norm_total_mean_1/norm_total_mean_2.
chi2LR1	Chi-square statistic for hypothesis testing of H0.
pvalue	P value of hypothesis testing of H0 (underlying whether a bin pair is a DCI).
pvalue.adj.FDR	Adjusted P value of H0's pvalue using FDR method.
Remark	Record of abnormal program information.

The outputs for the non-parametric tests are shown below:

Index	Description
bin_1,bin_2	The interacting region of the bin pair.
test_statistic	The statistic given by KS/CVM test.
pvalue	P value of hypothesis testing of H0 (underlying whether a bin pair is a DCI).
pvalue.adj.FDR	Adjusted P value of H0's pvalue using FDR method.

Example: The data getting from chr11 of oocyte and zygote cells with resolution=200kb (Flyamer et.al.) were untilized as sample data. In the sample data file, it lists all bin pairs with at least one non-zero counts in one of cell types. The first two columns represent the interacting region of each listed bin pair, then followed 86 columns denote the normalized read counts for oocyte cells and the last 34 columns denote the normalized read counts for zygote cells.

```
count.table <- read.table(paste("path/sampledata/oocyte.zygote.filtered.chr11.txt")
count.table <- as.matrix(count.table)
group <- factor(c(rep(1,86), rep(2,34)))
result.ks <- scHiCDiff.KS(count.table,group)
result.cvm <- scHiCDiff.CVM(count.table,group)
result.nb <- scHiCDiff.NB(count.table,group)
result.zinb <- scHiCDiff.ZINB(count.table,group)
result.nbh <- scHiCDiff.NBH(count.table,group)
```