DiffGR: Detecting differentially interacting genomic regions from Hi-C contact maps

Huiling Liu[1] and Wenxiu Ma[1,*]

[1]Department of Statistics, University of California, Riverside, 900 University Avenue, Riverside, California 92521, USA.
[*]Corresponding author: wenxiu.ma@ucr.edu

Abstract

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 large-scale chromatin organization such as topologically-associating domain (TAD) are still lacking. We proposed a new statistical method, DiffGR, for detecting differentially interacting genomic regions at the TAD level between Hi-C contact maps. We utilized the stratum-adjusted correlation coefficient (SCC) to measure similarity of local TAD regions. We then developed a nonparametric 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. DiffGR is publicly available at `https://github.com/wmalab/DiffGR`.

1 Introduction

Recent developments of chromatin conformation capture (3C)-based techniques—including 4C [1], 5C [2], Hi-C [3, 4, 5], ChIA-PET [6], and Hi-ChIP [7]—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 [8, 9, 10] are of particular interest because of their ability to map chromatin interactions at a genome-wide scale.

A Hi-C experiment yields a symmetric contact matrix in which each entry represents the chromatin contact frequency between the corresponding pair of genomic loci. 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 [11]. TADs are largely conserved across cell types and species. Moreover, CTCF and other chromatin binding proteins are enriched at the TAD boundaries, indicating that TAD boundary regions form chromatin loops and play an essential role in gene expression regulation [11, 12].

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 [13]. 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 diagonal bin to detect TADs and/or TAD boundaries. For instance, 11 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, 14 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 [15] identifies consistent TAD patterns across different resolutions by maximizing a quality scoring function of domain partition using dynamic programming. In addition, 16 proposed a TAD boundary detection method named HiCseg, which performs a 2D block-wise 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 [17, 18].

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 identification of differential chromatin interactions (DCIs), which represent different chromatin looping events between two Hi-C contact maps. In early studies, the most common strategy for DCI detection was to use the fold change values between two Hi-C contact maps. For instance, 19 used a simple fold-change strategy to detect the influence of estrogen treatment on chromatin interactions in MCF-7 Hi-C samples. Additionally, 20 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 [21]. In contrast, in more recent studies, count-based statistical methods, such as edgeR [22] and DESeq [23], have been adopted to identify pairwise chromatin interactions that show significant changes in contact frequencies. Among them, 24 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, 25 introduced a nonparametric test to detect statistically significant DCIs by permuting DCIs on each diagonal separately. While all these methods assumed independence among pairwise interactions, which holds true only in coarse-resolution Hi-C maps, 26 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, 27 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) [5, 10], to TADs (0.5–2 Mb on average) [11] and fine-scale chromatin interacting loops [8, 10]. 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 [11]. For example, 28 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, 13 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, 29 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 nonparametric 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.

2 Methods

The DiffGR method detects differentially interacting genomic regions in three steps, as described below (Sections 2.1-2.3). In addition, the simulation settings are outlined in Section 2.4 and real data preprocessing and analyses are described in Section 2.5.

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 [16] to the raw contact matrices and obtain the corresponding TAD boundaries. Note that one can exchange HiCseg with another TAD caller in this step. We choose HiCseg because it has been shown that HiCseg produces more robust TAD boundaries than other TAD-calling methods [17]. 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

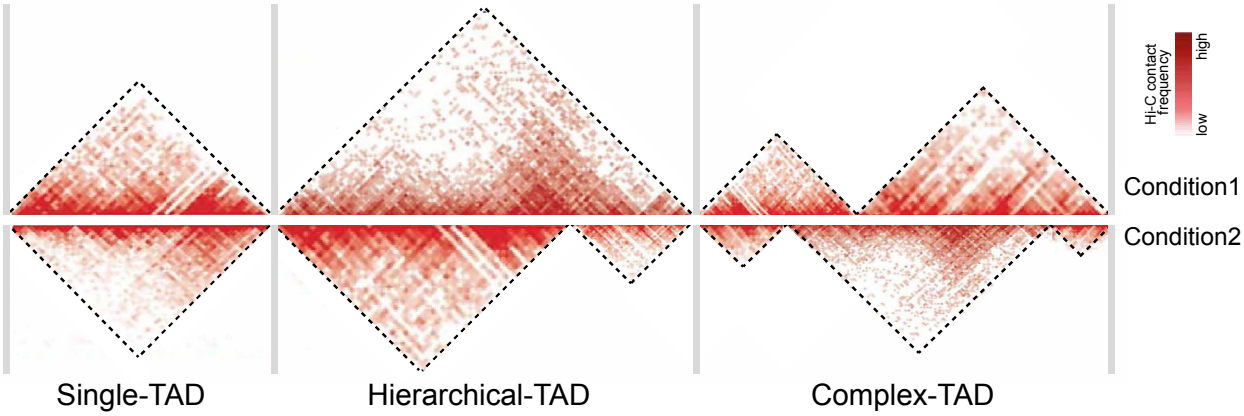

Figure 1: **Illustration of three candidate types of differential genomic regions.**

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 1.

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 k unique TAD boundaries in either one of the two Hi-C maps. In the single-TAD candidate region, $k = 0$; in the hierarchical-TAD or complex-TAD candidate regions, $k >= 1$. For each candidate region, we consider all $\binom{k+2}{2}$ 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) [30]) to measure the similarity of intra-TAD chromatin interactions between two Hi-C samples.

The SCC metric was introduced by 30 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 [31]. 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 preprocessing the Hi-C contact matrices, we first smooth each contact map by a 2D mean filter [30], 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 [32] on the smoothed matrices to remove potential biases.

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 nonparametric permutation test to estimate the p-values for these local SCC values (Section 2.3.1). Additionally, we propose a quantile regression strategy to speed up the permutation test (Section 2.3.2). 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.

2.3.1 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 nonparametric 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 [33].

2.3.2 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 nonparametric regression approach to estimate the quantiles of the SCC values. As shown in Supplementary Figure S1, 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 from the data. Therefore, the corresponding SCC values would be stabilized leading to relatively reliable and steady quantiles.

To speed up the permutation process, we adopt a nonparametric regression strategy to approximate the SCC quantiles (Supplementary Figure S1). 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 2.3.1.

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.

2.4 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 [10] as the original Hi-C data and simulated the altered Hi-C contact matrices as described below.

2.4.1 Single-TAD alternation

As TADs are conserved genomic patterns and TAD boundaries are relatively stable across cell types and even across species [11], 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 (Supplementary Figure S2). By default, we set the proportion of altered TADs to be 50%, which can vary from 20% to 70%.
- 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 outside that TAD region. 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 is computed as $\frac{\text{FP+FN}}{\text{N}}$, 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.

2.4.2 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 outside that TAD region. 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.

2.4.3 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 [10], 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 parameter $p = \{0.2, 0.4, 0.6, 0.8, 1.0\}$ represents the relative coverage level of the down-sampled contact matrix M'. In addition, 10% noise were added to the down-sampled matrices.

2.5 Real data preprocessing steps

In our real data analysis, we used the published Hi-C datasets by 10 (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 for subsequent 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 preprocessed by a 2D mean filter smoothing and then normalized by the KR method to eliminate potential biases.

In addition to Hi-C contact maps, CTCF and histone modification (including H3K4me1, H3K4me2, H3K27me3, and H3K36me3) ChIP-seq datasets from the same cell lines were obtained from ENCODE project [34] (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.

3 Results

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 2a, when the proportion of 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

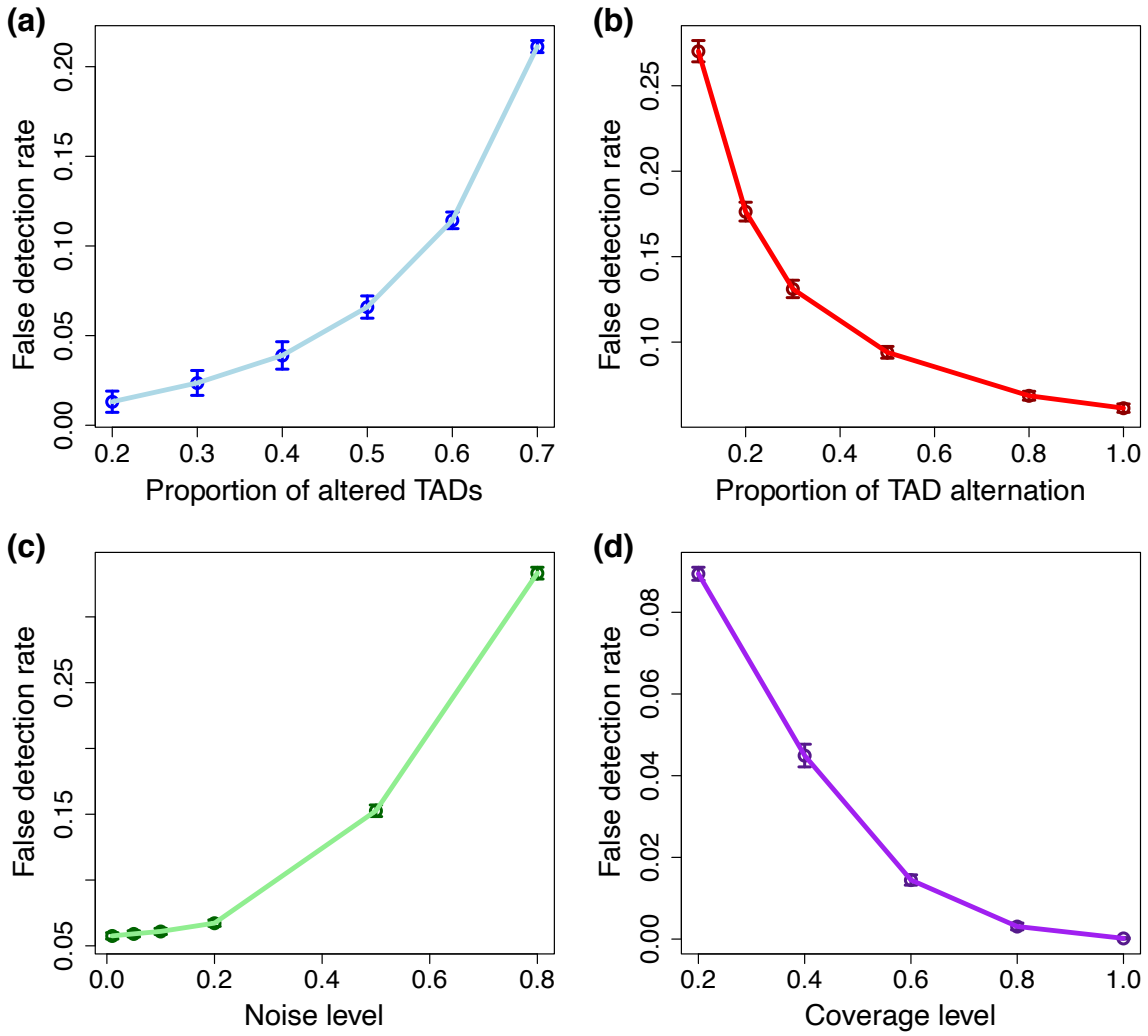

Figure 2: **Performance of single-TAD simulations.** The curves display the mean false detection rates at different levels of (a) proportion of altered TADs, (b) proportion of TAD alternation, (c) noise, and (d) sequencing coverage. Vertical bars represent 95% confidence intervals.

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 2b, the performance of DiffGR improved resulting in smaller false detection rates 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 false detection rate of our method was 0.27, suggesting that DiffGR can effectively detect subtle TAD differences.

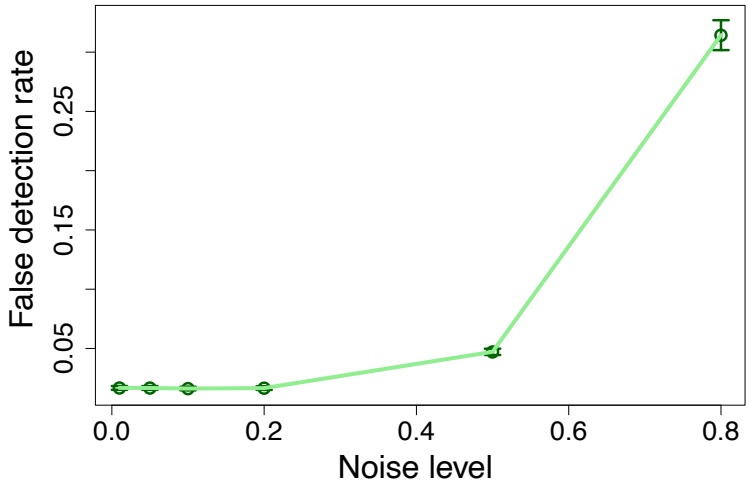

Figure 3: **Performance of hierarchical-TAD simulations.** The curve shows the mean false detection rates at various noise levels. Vertical bars represent 95% confidence intervals.

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 as the noise levels raised (Figure 2c). With moderate noise levels that were not greater than 20%, the false detection rates of DiffGR remained below 0.07, indicating that our method can accurately 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 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 2d 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.

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 3, the trend of the false detection rates across various noise levels under the hierarchical-TAD setting was similar to the pattern observed in the single-TAD case (Figure 2c). Furthermore, the false detection rates maintained low (less than 0.07) 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.

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 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 [30]. 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 4, 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.

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) [10] and identified

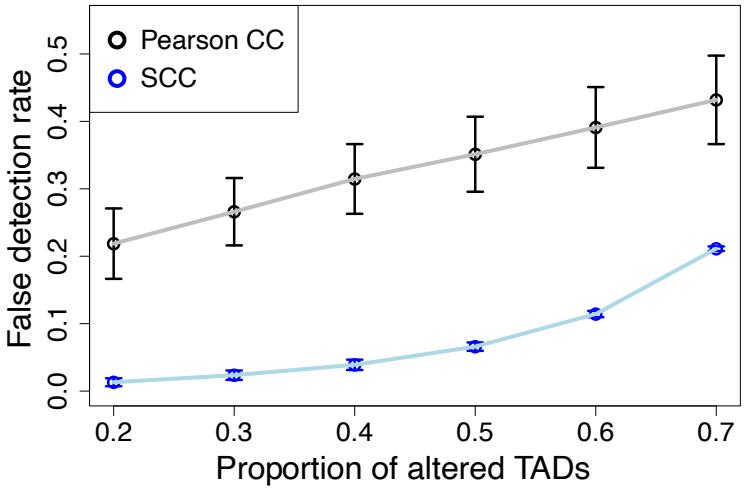

Figure 4: **Comparison between SCC and Pearson CC.** The curves represent the mean false detection rates at various proportions of altered TADs using either SCC (blue) or Pearson CC (black) as the local similarity metric. Vertical bars represent 95% confidence intervals.

statistically significant differential genomic regions between each pair of cell types with FDR cutoff 0.05. In total, we conducted ten pairwise comparisons among five cell types. In each pairwise comparison, we first applied HiCseg to identify TAD boundaries from the 50-kb contact matrix for each cell type and then partitioned the genome into three types of candidate regions: single-TAD candidate regions, hierarchical-TAD candidate regions, and complex-TAD candidate regions. As illustrated in Figure 5, of the candidate regions from all ten pairwise comparisons, 55.57% belonged to the single-TAD category (consistent with previous observations indicating that TAD boundaries are stable across cell types [11]), 31.88% to the hierarchical-TAD category, and 12.55% to the complex-TAD category. Our DiffGR analyses showed that only 24.79% of the single-TAD candidate regions showed statistically significant differences between two samples; 59.53% of the hierarchical-TAD candidate regions were determined to be differential; while the differential proportion of the complex-TAD category was as high as 90.20%. These observations indicated that candidate genomic regions with more distinct patterns of TAD boundaries are more likely to be detected as differential between 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 (Supplementary Figure S3).

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

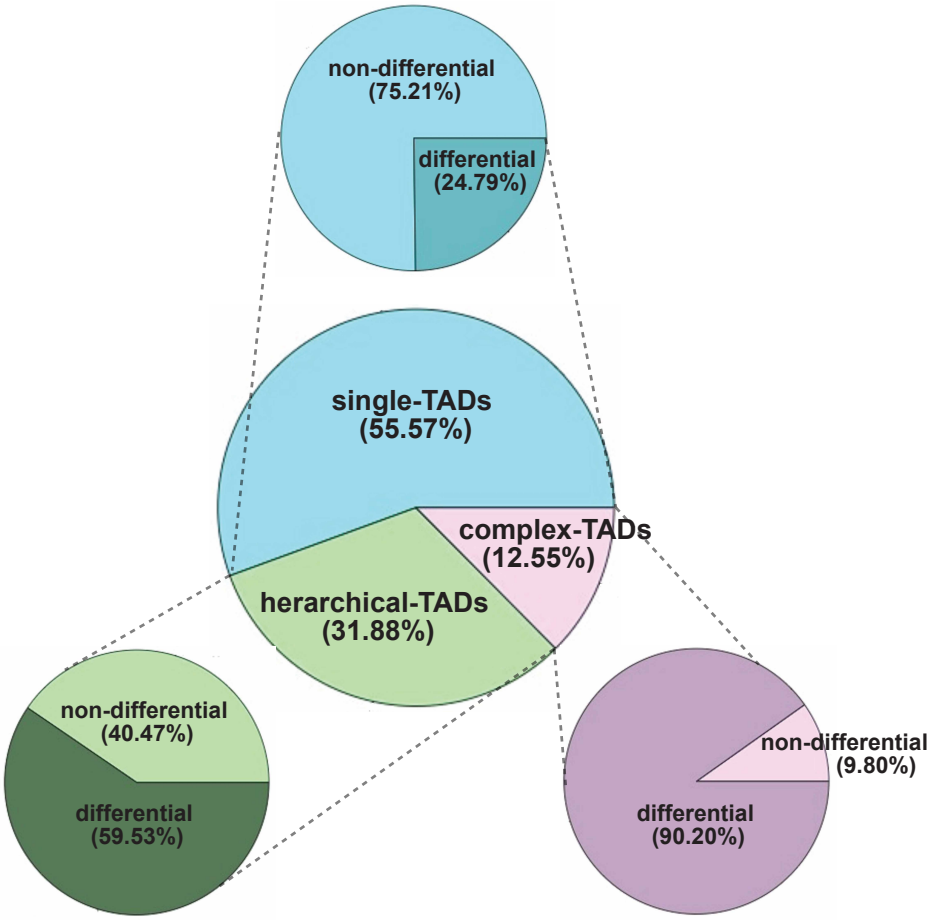

Figure 5: **Piecharts of DiffGR results obtained from human Hi-C datasets.** The center piechart presents the proportions of three types of candidate regions. The three outer piecharts display the proportions of detected differential genomic regions, one for each candidate category.

Table 1: **Agreements between ChIP-seq data and DiffGR-detected differential genomic regions.**

	100 kb	50 kb	25 kb
CTCF	78 (35.45%)	121 (55.00%)	138 (62.73%)
H3K4me1	60 (27.27%)	111 (50.45%)	144 (65.45%)
H3K4me2	57 (25.91%)	93 (42.27%)	118 (53.64%)
H3K27me3	57 (25.91%)	87 (39.55%)	119 (54.09%)
H3K36me3	38 (17.27%)	73 (33.18%)	121 (55.00%)

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 consistent. The counts and percentages of consistent results were reported for each ChIP-seq dataset at each resolution.

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. 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 [34]. 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 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 1, 78 out of 220 (35.45%) tests showed significantly higher absolute mean differences of CTCF values at the TAD boundaries in DiffGR-detected differential genomic regions than those in non-differential regions at 100-kb resolution. Whereas in the results at 25-kb resolution, 138 (62.73%) 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.

4 Discussion and Conclusions

With the fast accumulation 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 interactions, while few studies addressed the detection of differential chromatin organization at TAD scale. To tackle 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 nonparametric permutation test to assess the statistical significance of the local SCC values. In contrast to the parametric approaches that were used by most Hi-C data analysis methods, our nonparametric 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. Additionally, we utilized a nonparametric smoothing spline regression to speed up the permutation test. 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. Furthermore, DiffGR produced robust and stable detection results under various noise and coverage levels across different resolutions. To summarize, DiffGR provides 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.

We envision a few possible extensions and future directions based on this work. First, our method preforms pairwise comparison between 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 pool biological replicates together 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. Given the high sparsity and variability of single-cell Hi-C contact matrices, identifying differential chromatin interactions at single-cell level remains a significant challenge.

Funding

This work was supported by the National Science Foundation [DBI-1751317]; and the National Institute of Health [R35GM133678].

Conflict of interest: none declared.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

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

[16] 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.

[17] 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.

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

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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):550, 2014.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

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

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

[33] 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.

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