Bioinformatics, 35, 2019, i145–i153
doi: 10.1093/bioinformatics/btz362
ISMB/ECCB 2019

Selfish: discovery of differential chromatin interactions via a self-similarity measure

Abbas Roayaei Ardakany[1,*], Ferhat Ay[2,3] and Stefano Lonardi [1,*]

[1]Department of Computer Science and Engineering, University of California, Riverside, Riverside, CA 92521, USA, [2]Division of Vaccine Discovery, La Jolla Institute for Allergy and Immunology and [3]School of Medicine, Department of Pediatrics, UC San Diego, La Jolla, CA 92037, USA

*To whom correspondence should be addressed.

Abstract

Motivation: High-throughput conformation capture experiments, such as Hi-C provide genome-wide maps of chromatin interactions, enabling life scientists to investigate the role of the three-dimensional structure of genomes in gene regulation and other essential cellular functions. A fundamental problem in the analysis of Hi-C data is how to compare two contact maps derived from Hi-C experiments. Detecting similarities and differences between contact maps are critical in evaluating the reproducibility of replicate experiments and for identifying differential genomic regions with biological significance. Due to the complexity of chromatin conformations and the presence of technology-driven and sequence-specific biases, the comparative analysis of Hi-C data is analytically and computationally challenging.

Results: We present a novel method called Selfish for the comparative analysis of Hi-C data that takes advantage of the structural self-similarity in contact maps. We define a novel self-similarity measure to design algorithms for (i) measuring reproducibility for Hi-C replicate experiments and (ii) finding differential chromatin interactions between two contact maps. Extensive experimental results on simulated and real data show that Selfish is more accurate and robust than state-of-the-art methods.

Availability and implementation: https://github.com/ucrbioinfo/Selfish

Contact: aroay001@ucr.edu or stelo@cs.ucr.edu

1 Introduction

Recent studies have revealed that genomic DNA in eukaryotes is not arbitrarily packed into the nucleus. The chromatin has a well-organized and regulated structure in accordance to the stage of the cell cycle and environmental conditions (Ma *et al.*, 2015; Pederson, 1972). The chromatin structure in the nucleus plays a critical role in many essential cellular processes, including regulation of gene expression and DNA replication (Dixon *et al.*, 2012, 2015; Rao *et al.*, 2014; Sexton *et al.*, 2012).

Technological and scientific advancements in genome-wide DNA proximity ligation (Hi-C) have enabled life scientists to study how chromatin folding regulates cellular functions (Cavalli and Misteli, 2013; Chen *et al.*, 2015; Gorkin *et al.*, 2014; Lieberman-Aiden *et al.*, 2009). The analysis of Hi-C led to the discovery of new structural features of chromosomes such as topologically associating domains (TADs) (Dixon *et al.*, 2012; Zufferey *et al.*, 2018) and chromatin loops (Cao *et al.*, 2018; Rao *et al.*, 2014).

With the decreasing cost of Hi-C experiments and the higher availability of Hi-C data for different cell types in diverse conditions, there is a growing need for reliable and robust measures to systematically compare contact maps to discover similarities and differences. However, the comparative analysis of Hi-C data presents computational and analytical challenges due to presence of technology-driven and sequence-specific biases. Technology-driven biases include sequencing depth, cross-linking conditions, circularization length, and restriction enzyme sites length (Cournac *et al.*, 2012; O'Sullivan *et al.*, 2013; Stansfield and Dozmorov, 2017). Sequence-specific biases include GC content of trimmed ligation junctions, sequence uniqueness and nucleotide composition (Yaffe and Tanay, 2011). For instance, it is well-known that contact maps from replicate experiments can contain significant differences solely due to these biases, which could be falsely interpreted as biological differences if these biases were not accounted for (Yardimci *et al.*, 2018) . Several normalization methods have been developed to compensate for these biases and improve the reproducibility of Hi-C experiments (e.g. Imakaev *et al.*, 2012; Knight and Ruiz, 2013; Lieberman-Aiden *et al.*, 2009; Yaffe and Tanay, 2011). While several computational methods have been proposed to extract statistically significant differences in normalized contact maps (Ay *et al.*, 2014; Cairns *et al.*, 2016; Rao *et al.*, 2014; Ron *et al.*, 2017), their performance is still not entirely satisfactory due to the inherent

complexity, inter-dependency and unaccounted biases in chromatin interaction data.

There are two major domains of application for the comparative analysis of Hi-C contact maps. The first application domain is focused on quantifying the reproducibility of Hi-C biological/technical replicate experiments (Yardimci *et al.*, 2018). For instance, Yang *et al.* (2017) defined a reproducibility measure based on the *stratum-adjusted correlation coefficient statistic* defined on the unique spatial features of Hi-C data. Their method HiCRep (i) reduces the effect of noise and biases by applying a 2D averaging filter on the data, (ii) addresses the distance dependence of Hi-C data by stratifying the data with respect to the genomic distance, (iii) calculates a Pearson correlation coefficient for each stratum and (iv) aggregates the computed stratum-specific correlation coefficients using a weighted average.

In the same application domain aimed at quantifying reproducibility or concordance of contact maps, Ursu *et al.* (2018) presented a method called GenomeDISCO to measure the differences between smoothed contact maps. GenomeDISCO represents contact map as a graph, where each node represents a genomic locus and each edge represents an interaction between two loci. Edges are weighted by the normalized frequency of the corresponding pairs of loci. GenomeDisco executes iteratively the two following steps: (i) traverses the graph using random walks, which has the effect of denoising (smoothing) the data, (ii) computes the normalized difference between smoothed contact maps using the L_1 distance between two contact maps.

The second application domain is aimed at finding statistically significant differences between contact maps for cells in different states (tissues, developmental states, healthy/diseased, time-points, etc.). It is well-known that chromatin interactions that are mediated by specific protein can have distinct frequencies in different cell types or in different cell conditions (Gong *et al.*, 2011; Patel *et al.*, 2012). Differences in chromatin interactions can be associated with cell-type-specific gene expression or mis-regulation of oncogenes or anti-oncogenes (Hnisz *et al.*, 2016; Liu *et al.*, 2008; Schmitt *et al.*, 2016).

Wang *et al.* (2013) proposed the first method to discover differences in Hi-C contact maps. The authors used a simple fold change of the normalized local interactions to discover that estrogen stimulation significantly impacts chromatin interactions in MCF7 cells. Building on this idea, Dixon *et al.* (2015) proposed a method that (i) quantile normalizes contact maps to compensate for the bias induced by different sequencing depth, and (ii) determines the significance of normalized differences between two contact maps (augmented by feature vectors representing epigenetic signals) using a Random Forest model. Their method can (i) determine whether the epigenetic signal is predictive of changes in interaction frequency and (ii) discover which epigenetic signals are most predictive of changes in higher-order chromatin structure.

Stansfield and Dozmorov (2017) developed a non-parametric method to account for between-datasets biases. They used locally weighted polynomial regression to fit a simple model trained on the difference between the two datasets. Based on the assumption that the majority of the interactions should be relatively unchanged among similar Hi-C datasets and by centering the average difference to zero, loci which are far from the average are considered potentially significant differential interaction.

Unlike other methods which assume independence among pairwise interactions (which holds true only for low resolution Hi-C) Djekidel *et al.* (2018) presented a method that takes into account the dependency of adjacent loci in higher resolutions. Based on the fact that interacting neighboring loci are known to be inter-dependent, structural differences can be detected by observing the differences in a neighborhood of the corresponding loci pair. In contrast, random noises tend to affect singular pairwise interactions only. By considering a three-dimensional space in which the x and y are the coordinates of the genomic loci and z is their pairwise interaction frequency, the authors define a chromatin interaction between two conditions to be *differential* when the intensity of the majority of k-nearest neighbors of (x, y) exhibit a significant change.

In this work, we address three major weaknesses of these existing methods for the comparative analysis of contact maps, namely, (i) ignoring the inter-dependency of chromatin interactions, (ii) requiring a pre-processing (normalization) step based on a flawed assumption that biases between two contact maps can be accurately modeled and (iii) being extremely computationally demanding for the analysis of high-resolution Hi-C data. We present new comparative methods for the analysis of Hi-C data based on the notion of self-similarity (Shechtman and Irani, 2007). We show that our self-similarity measure is robust to biases and does not need complex and computationally intensive normalization steps, such as Minus versus Average (MA) (Dudoit *et al.*, 2002) or Minus versus Distance (MD) (Stansfield and Dozmorov, 2017). In the first part of the paper, we show that our self-similarity measure can be used as a tool to quantify the reproducibility of Hi-C biological/technical replicate experiments. In the second part, we show that our measure can also be employed for finding statistically significant differences between Hi-C contact maps.

2 Materials and methods

Although existing methods for comparing contact maps vary widely, they all share the following assumption. Given two contact maps that are expected to be similar (e.g. technical replicates of the same biological experiment), it is possible to devise (or train) a common underlying model can faithfully represent both. We believe that this approach is fundamentally flawed, because the inherent biases present in the Hi-C data are very hard to model and completely eliminate. Here we propose to use the intrinsic self-similarity structure in contact maps to avoid dealing with the modeling problem.

In the application domain of object detection in complex visual data, the notion of self-similarity was first introduced by Shechtman and Irani (2007). The idea of self-similarity on images can be explained as follows. Shechtman and Irani (2007) showed that given two images of a certain object, the most relevant correlations between them are not necessarily the raw values of pixels (or an underlying model describing those pixel values) but the internal organization of self-similarities of local regions at similar relative geometric positions. Given two images of the same object, the relation between these local self-similarities tend to be more preserved than the similarities between the images.

2.1 Self-similarity and reproducibility

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

When we compare two contact maps that are expected to be similar, e.g. for two technical replicates of the same biological experiment, we expect to have similar internal layout of interactions. More precisely, given two contact maps A and B for two replicates,

if we observe more chromatin interactions in block α than in block β in contact map A, we expect to have more chromatin interactions in block α than block β in contact map B as well, for several local choices of α, β. In other words, to measure similarity we do not need to depend on the absolute number of interactions in each contact map, rather we can rely on pairwise comparison between many local interactions. Here we claim that the Boolean vectors representing binary comparison between local interactions encode enough information to define a similarity measure that can be used to quantify reproducibility for contact maps.

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

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

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

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

2.2 Self-similarity and differential chromatin interactions

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

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

To incorporate locality information in our self-similarity representation, each interaction $A(i, j)$ is represented by a linear combination of its neighboring interactions. To penalize interactions which are progressively farther from (i, j), we weight these local interactions via a Gaussian filter centered at (i, j) (see Fig. 1 for an example). By gradually increasing the size of the Gaussian filter, we capture impact regions with larger and larger radii. We denote with $G^A_{r_k}$ the matrix resulting from the convolution between a Gaussian filter with radius r_k and the contact map A. We first compute $G^A_{r_k}$ for a set of n radii $\{r_1, r_2, \ldots, r_n\}$, and collect them in vector Γ_A as follows

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

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

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

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

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

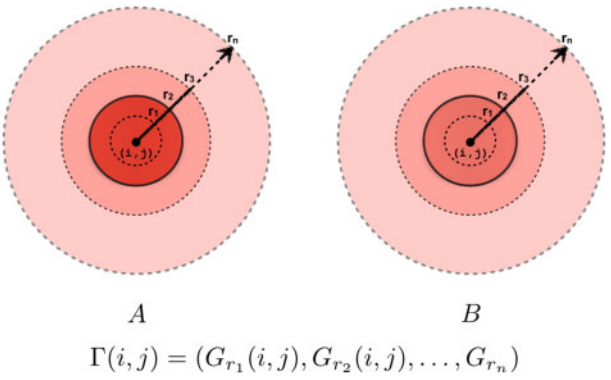

Fig. 1. Our self-similarity metric for representing chromatin interactions is obtained by first convolving a contact map with a set of Gaussian filters with radii $\{r_1, r_2, \ldots, r_n\}$. The shade represents the intensity of the convolution for different radii. In this example, a sharp frequency change can be observed between radius r_2 and r_3 in contact map A but not in contact map B. This difference can indicate a potential DCI

of self-similarity, i.e. by using local comparison of local regions in contact maps.

According to Mikolajczyk (2002), the Gaussian filter scale r must be distributed exponentially between the inner (r_1) and outer (r_n) scale limits (impact radii) $r_n = r_0 s^n$, in order to maintain a uniform change of information between successive levels of Gaussian filtering. For 5-kb resolution data, we set $r_0 = 7$, $s = 2$ and $n = 10$.

Inspired by the work of Lowe (2004), instead of using Γ to define the behavior of interaction frequency across a set of impact regions, we use the first-order derivative of Γ with respect to impact radius r, which can be estimated by the difference $G_{r_{k+1}} - G_{r_k}$.

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

By computing the first-order derivative for various choices of the impact radii, we carry out a comparison of local contact map regions ($G_{r_{k+1}} - G_{r_k}$). Figure 2 shows the first-order derivatives of $\Gamma_{\hat{A}}$ and $\Gamma_{\hat{B}}$ and the difference between them for the DCI reported later in Figure 8. Observe the sharp change between the derivatives at radius r_2 which corresponds to a DCI with that radius.

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

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

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

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

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

These P-values $P_{A,B}(i,j)$ are finally fed into the Benjamini–Hochberg algorithm to calculate the final probabilities (Benjamini and Hochberg, 1995).

3 Results

3.1 Reproducibility

We evaluated our reproducibility measure on a Hi-C dataset obtained from Schmitt *et al.* (2016) that has a variable total number of interactions and resolution. The dataset consists of five different cell types: hESC (H1), Mesendoderm (MES), Mesenchymal Stem Cell (MSC), Neural Progenitor Cell (NPC) and Trophoblast-like Cell (TRO). Each cell type has two biological replicates. All experiments were carried out on a single chromosome (chromosome 1 for this work) with 40 kb resolution. According to our experience, parameter c is dependent on the particular Hi-C protocol used. Parameter k must be chosen such that the resulting blocks enclose the primary structures of contact maps which are likely to be preserved between cell types, e.g. TADs. Parameter c has to be set to the largest integer value such that the computed reproducibility for biological replicates is at least 0.9. For this dataset, we set $k = 100$ and $c = 5$. We compared our method Selfish against two state-of-the-art reproducibility methods, namely HiCRep (Yang *et al.*, 2017) and GenomeDISCO (Ursu *et al.*, 2018).

First, we assessed the effect of the total number of intra-chromosomal interactions captured by Hi-C experiment on different reproducibility measure. Given two biological replicates, we generated pseudo-replicates by first summing the two Hi-C matrices and then down-sampling the resulting matrix. Any pair of contact maps, which are either replicates or pseudo-replicates are called *non-replicates*. Next, each individual replicate was down-sampled to a wide range of total interactions ($10^5, 5\times10^5, 10^6, 2\times10^6, 5\times10^6, 10^7$). For each of these choices, we computed the pair-wise reproducibility score.

Figure 3a–c illustrates the effect of the total number of interactions (which depends on the Hi-C sequencing depth) on the performance of reproducibility measures. A desirable feature for a reproducibility measure is to produce similarity scores that are invariant from the total number of interactions. Observe in Figure 3a–c that Selfish is much more invariant to the total number of interactions than HiCRep and GenomeDISCO. Both these latter methods failed to report stable reproducibility scores which are independent from the sequencing depth.

In the next experiment, we evaluated the effect of binning resolution of Hi-C data on the reproducibility methods. For this experiment, we used deeply sequenced Hi-C data of cell type GM12878

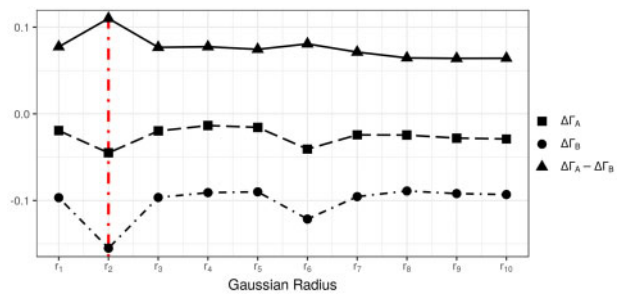

Fig. 2. The first-order derivatives $\Delta\Gamma_A$ and $\Delta\Gamma_B$ and the difference between them for the DCI reported later in Figure 8. A large difference between $\Delta\Gamma_A$ and $\Delta\Gamma_B$ at radius r_2 indicates the presence of a potential DCI

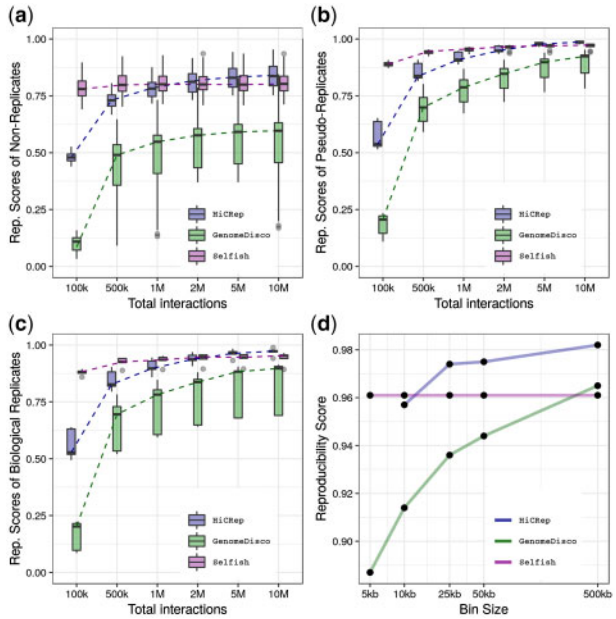

Fig. 3 Illustrating the effect of the total number of interactions on reproducibility score of (**a**) non-replicates, (**b**) pseudo-replicates and (**c**) biological replicates. Panel (**d**) illustrates the effect of data resolution (bin size) on reproducibility score of two replicates of cell type GM12878 from Rao *et al.* (2014)

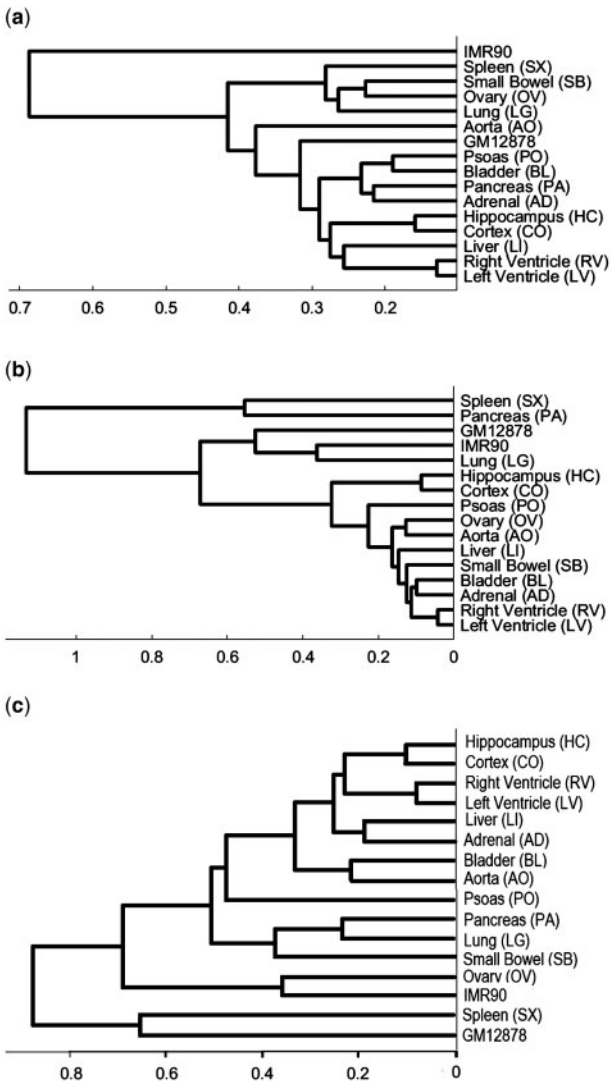

Fig. 4. Clustering of 14 human primary tissues and two cell lines obtained from Schmitt *et al.* (2016). The dendrograms are computed based on the pairwise similarity calculated using (**a**) GenomeDisco, (**b**) Selfish and (**c**) HiCRep

Table 1. Average running time of HiCRep, GenomeDisco and Selfish for different choices of the data resolution

Method	500 kb	50 kb	25 kb	10 kb	5 kb
HiCRep	6 s	636 s	1045 s	34 479 s	–[a]
GenomeDisco	7 s	2989 s	4933 s	30 238 s	61 004 s
Selfish	0.75 s	85 s	184 s	345 s	474 s

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

from Rao *et al.* (2014). Again, a desirable property of a reproducibility score is to be robust to changes in resolution. Figure 3d shows that reproducibility scores for 5 kb, 10 kb, 25 kb, 50 kb and 500 kb resolutions are very stable for Selfish, whereas HiCRep and GenomeDISCO scores are resolution dependent, in particular for GenomeDISCO.

We also tested our reproducibility measure to cluster different cell and tissue types. Contact maps of 14 different tissues and 2 cell types were obtained from Schmitt *et al.* (2016). A visual inspection

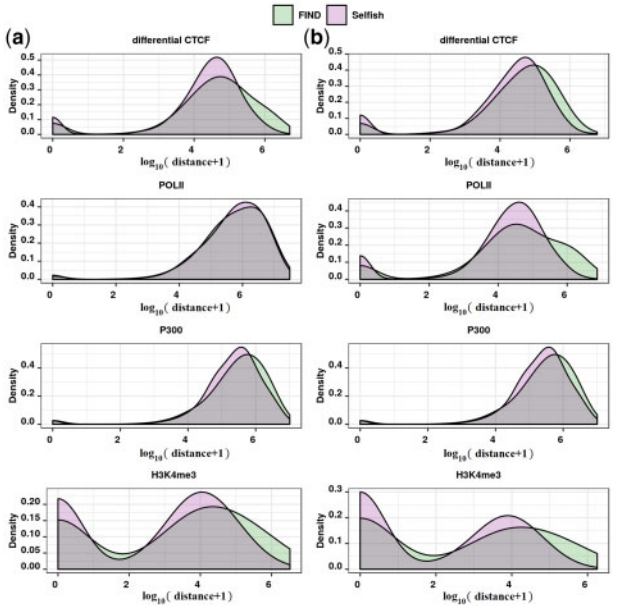

Fig. 5. Enrichment of differential transcription factor binding and epigenetic marks (CTCF, POLII, P300 and H3K4me3) around reported DCIs for (**a**) cell type GM12878 and (**b**) cell type K562

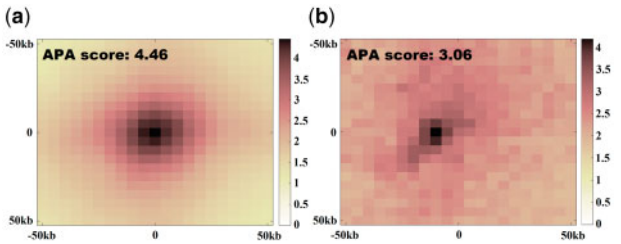

Fig. 6. A modified APA plots for reported DCIs between two cell types GM12878 and K562 by (**a**) Selfish and (**b**) FIND

of Figure 4 shows that our reproducibility measure can cluster similar cell and tissue types. For instance, observe that Selfish correctly clusters the right and left ventricles, IMR90 and lung as well as hippocampus and cortex together.

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

We also compared the average running time of the three methods on two replicates of cell type GM12878 from Rao *et al.* (2014). Table 1 shows that Selfish is by far more efficient than HiCRep and GenomeDISCO.

3.2 Differential chromatin interaction

We compared Selfish to the current state-of-the-art method for detecting differential chromatin interaction called FIND (Djekidel *et al.*, 2018). To the best of our knowledge, FIND is the only DCI detection method which works on high-resolution Hi-C data by taking into account the chromatin interactions inter-dependency. Extensive experimental results in Djekidel *et al.* (2018) show that FIND performs better than previously published methods for DCI detection.

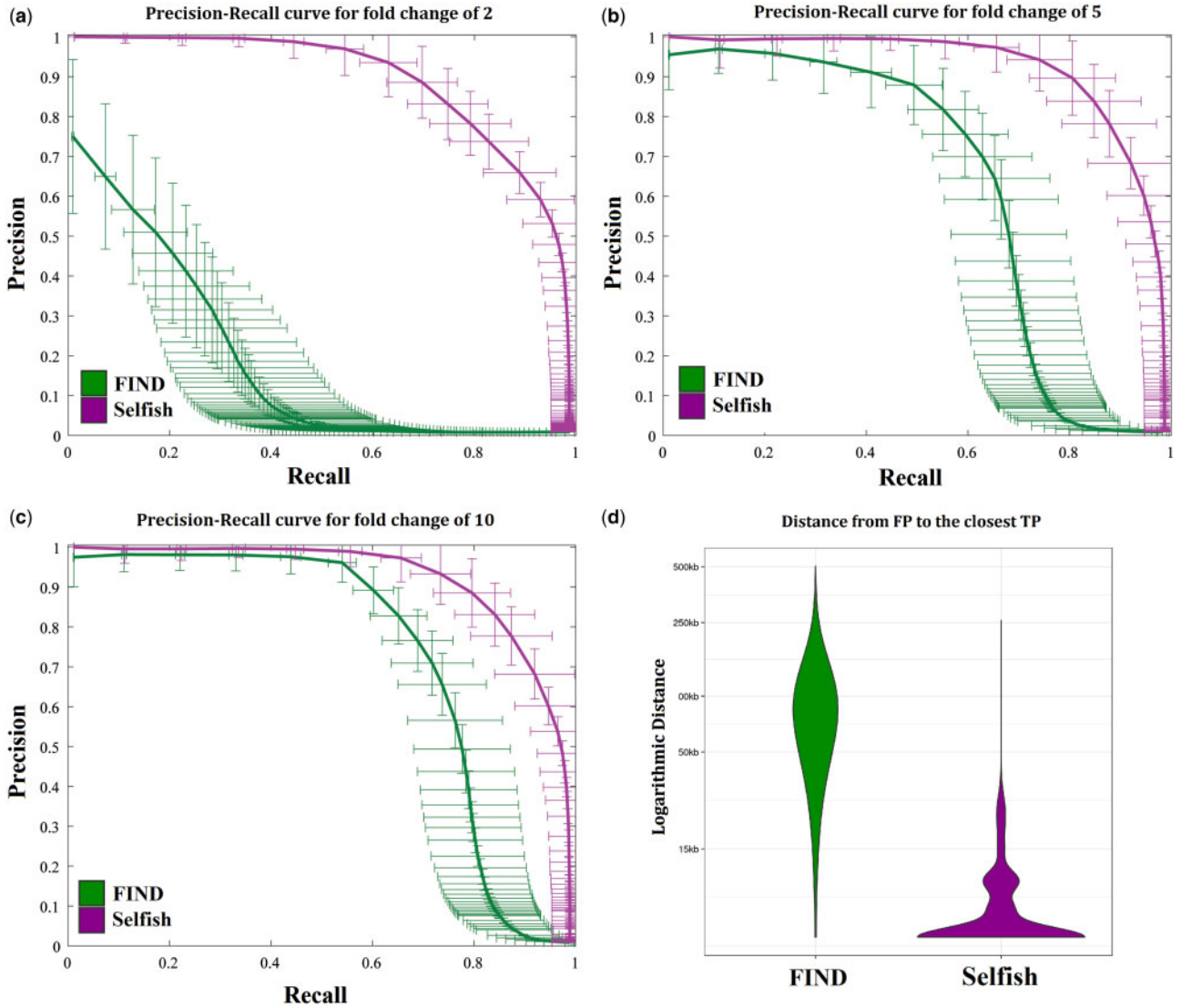

Fig. 7. Precision–recall curves for Selfish (magenta) and FIND (green) for (**a**) 2-fold, (**b**) 5-fold and (**c**) 10-fold DCIs. The vertical and horizontal bars represent the 95% confidence interval for precision and recall at that threshold respectively. (**d**) The distribution of distances of FPs to closest TPs

We ran Selfish and FIND on Hi-C contact maps for cell types GM12878 and K562 obtained from Rao *et al.* (2014). We replicated some of the experiments proposed in Djekidel *et al.* (2018) in order to make a fair comparison with FIND. First, we analyzed the enrichment of epigenetic signals in the neighborhood of detected DCIs as well as the percentage of nearby genes having significant expression fold changes. We evaluated the enrichment of four important epigenetic markers, namely the binding of CTCF, POLII and P300, as well as the presence of histone modification H3K4me3. CTCF is widely recognized as a main driver of chromatin structure (Rao *et al.*, 2014; Tang *et al.*, 2015). We computed the enrichment of CTCF differential peaks (i.e. peaks that are different between two cell types) around detected DCIs. For this part of the analysis, we obtained the FIND's detected DCIs from Djekidel *et al.* (2018).

To compute the enrichment of each marker in the neighborhood of the detected DCIs, we calculated the distance of marker peaks to their closest anchor of DCIs. Figure 5 shows the enrichment of epigenetic markers near DCIs. Observe that CTCF and H3K4me3 are more enriched around the Selfish's reported DCIs than those detected by FIND, even though the number of reported DCIs for Selfish is twice as large (30 456 versus 14 131).

We also calculated the expression fold change of nearby genes for the two cell types. We first determined the set of genes which have overlap with any of the detected DCIs' anchors. Then we computed the percentage of those genes having an expression fold change of two or greater. For the set of genes overlapping FIND's DCIs, 71.46% of them were over-expressed. For Selfish, 78.78% were over-expressed. This analysis confirmed that the differences in chromatin structure are strongly associated with the changes in gene regulation. However, the DCIs detected by Selfish have stronger associations to differences in gene regulation than FIND. For the gene expression analysis, we used the dataset from Djekidel *et al.* (2018). The expression data were obtained from ENCODE, accession numbers GSE78553 and GSE78625, for cell types GM12878 and K562, respectively.

To quantify how well the Hi-C data supported the detected DCIs between two cell types GM12878 and K562, we generated a modified aggregate peak analysis (APA) plots (Phanstiel *et al.*, 2015; Rao *et al.*, 2014). The interaction frequencies in contact maps were

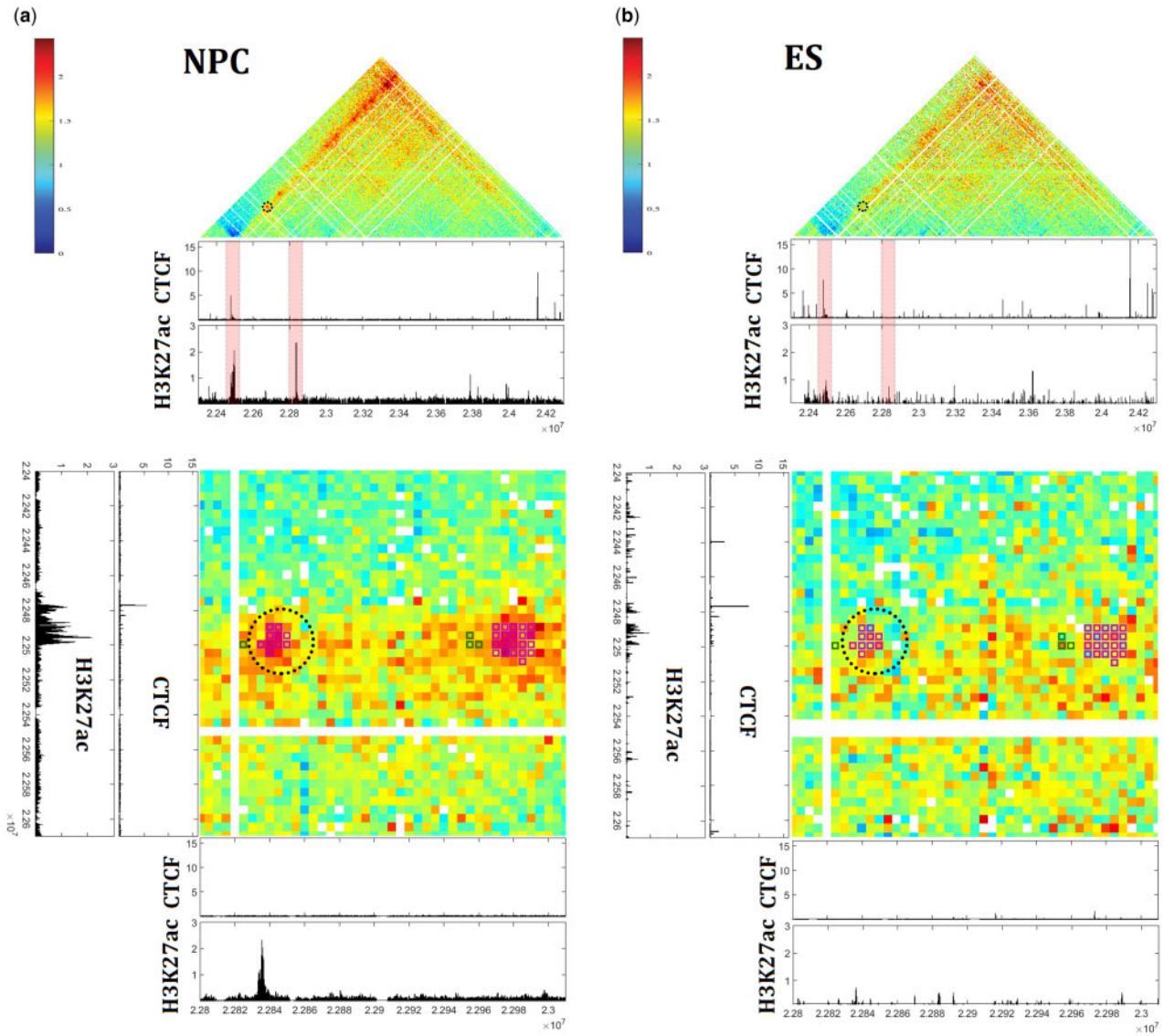

Fig. 8. A 2-Mb region shown around Brn2 promoter of chromosome 4 of mouse neural cells: (**a**) ES and (**b**) NPC. Dashed circles show the contact between the Brn2 promoter and an NPC-specific enhancer. Insets show the magnified view of this contact

first Z-normalized along the diagonals as explained in Section 2. Then, for each detected DCI, we calculated the interaction differences between the contact maps in a ±-50Kb neighborhood. By averaging over all DCIs, we computed the APA plot for differences. The differential APA score, i.e. the value of the central index in the plot compared to neighboring regions shows how different the interactions are at reported DCIs with respect to their expected interaction frequency with that genomic distance.

Figure 6 shows that Selfish produces the expected Gaussian-shaped plot around its reported DCIs while the DCIs from FIND failed to generate a similar pattern. Correspondingly, the computed APA score is higher for Selfish (4.46) compared to FIND (3.06) suggesting a stronger detection of DCIs. Finally the peak pixel of the APA plot for FIND (score of 4.17) is not centered on the called DCI pairs suggesting that Selfish performs better at pinpointing anchor points of chromatin interactions at high resolution.

We also compared the runtime of FIND and Selfish on the 5-kb-resolution dataset described above. FIND failed to run on the whole genome or even on large segments of contact maps. To compare the efficiency of Selfish and FIND, we computed the average runtime required to process random contact map segments of 6 Mb. Selfish took an average of 101.6 s. FIND required an average of 12 839.6 s, about 120× slower than Selfish.

To further investigate the accuracy of detected DCIs, we used simulated Hi-C data generated by the method proposed in Zhou *et al.* (2014). We generated 100 pairs of simulated contact maps, each of which had known location for DCIs. After running Selfish and FIND on these simulated datasets, we obtained a *P*-value for each DCI location for all 100 simulated pairs of contact maps. Given the *P*-values and the true locations of DCIs (true positives), we computed a precision–recall curve for each simulated pair of contact maps. We used the threshold averaging method proposed by Fawcett (2006) to combine the 100 precision–recall curves to get the overall performance curve. We thresholded over the ratio of all indices in the contact map used for computing the precision and recall for each simulated pair. To combine the curves, we averaged all 100 calculated precision and recall values for each threshold. Figure 7a–c shows the performance of both methods for 2-fold, 5-fold and

10-fold DCIs. The vertical and horizontal bars represent the 95% confidence interval for precision and recall at that threshold respectively. Selfish performed better than FIND on all fold change settings, confirming the stronger performance that we observed on real Hi-C data. The performance difference is most striking for small fold change values, which are more relevant for comparisons of real Hi-C datasets and yet have a very large effect on gene regulation (Greenwald *et al.*, 2018). It is also important to note that Selfish's performance is quite consistent across different samples as indicated by small confidence intervals.

Figure 7d shows the distribution of distances between false positive DCIs produced by Selfish and FIND to true DCIs (true positives). To generate this figure, we set the number of returned DCIs equal to the number of true DCIs. Then, for each falsely detected DCI, we calculated its distance to the closest true DCI. These results clearly show that most of Selfish's false positives are located in close proximity of true DCIs confirming their relevance to true differences and their non-random distribution. FIND's false positive are instead much farther from true DCIs and are more scattered in the contact map.

In our final experiment to assess the performance of two methods, we tested Selfish and FIND on a real test case from Bonev *et al.* (2017). Figure 8 shows a 2-Mb region around the *Brn2* promoter (also known as *Pou3f2*) for mouse embryonic stem cells (ES) and neuronal progenitor cells (NPC). Dashed circles show the contact between the *Brn2* promoter and an NPC specific enhancer. Insets show the magnified view of this contact. Observe that the contact between the promoter and enhancer is strongly present in the NPC cell (Fig. 8a) in contrast with the ES cell in which this interaction is weak (Fig. 8b). The mentioned contrast shows itself as a subtle but important difference of interactions between two cell types. The highlighted regions of the epigenetic signals show the difference in the specified regions between two cell types. Detected DCIs by Selfish and FIND for q-value $< 10^{-4}$ are shown in magenta and green squares, respectively. Observe that Selfish can identify this contact region as a DCI between two cell types, but FIND fails to detect it.

4 Conclusion

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

We showed that a simple binary comparison operation between blocks in the contact maps can be used to encode the local and global features in a manner that is robust to the data resolution and sequencing depth. This encoded information is used to build a feature vector for each contact map, which in turn allows to define a simple but effective similarity metric using the distance between their feature vectors. Experimental results showed that our self-similarity-based measure outperformed two state-of-the-art methods (HiCRep and GenomeDISCO) for measuring reproducibility of replicated Hi-C experiments.

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

Funding

This work was supported, in part, by the US National Science Foundation [IOS-1543963, IIS-1526742, IIS-1814359].

Conflict of Interest: none declared.

References

Ay,F. *et al.* (2014) Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. *Genome Res.*, **24**, 999–1011.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B*, **57**, 289–300.

Bonev,B. *et al.* (2017) Multiscale 3D genome rewiring during mouse neural development. *Cell*, **171**, 557–572.e24.

Cairns,J. *et al.* (2016) CHiCAGO: robust detection of DNA looping interactions in capture Hi-C data. *Genome Biol.*, **17**, 127.

Cao,Y. *et al.* (2018) Accurate loop calling for 3d genomic data with cLoops. bioRxiv, doi:10.1101/465849.

Cavalli,G. and Misteli,T. (2013) Functional implications of genome topology. *Nat. Struct. Mol. Biol.*, **20**, 290–299.

Chen,H. *et al.* (2015) Functional organization of the human 4D nucleome. *Proc. Natl. Acad. Sci. USA*, **112**, 8002–8007.

Cournac,A. *et al.* (2012) Normalization of a chromosomal contact map. *BMC Genomics*, **13**, 436.

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

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

Djekidel,M.N. *et al.* (2018) FIND: difFerential chromatin INteractions Detection using a spatial Poisson process. *Genome Res.*, [Epub ahead of print, doi:10.1101/gr.212241.116, February 12, 2018].

Dudoit,S. *et al.* (2002) Statistical methods for identifying differentially expressed genes in replicated CDNA microarray experiments. *Stat. Sin.*, **12**, 111–139.

Fawcett,T. (2006) An introduction to ROC analysis. *Pattern Recognit. Lett.*, **27**, 861–874.

Gong,F. *et al.* (2011) The BCL2 gene is regulated by a special AT-rich sequence binding protein 1-mediated long range chromosomal interaction between the promoter and the distal element located within the 3'-UTR. *Nucleic Acids Res.*, **39**, 4640–4652.

Gorkin,D.U. *et al.* (2014) The 3D genome in transcriptional regulation and pluripotency. *Cell Stem Cell*, **14**, 762–775.

Greenwald,W.W. *et al.* (2018) Integration of phased Hi-C and molecular phenotype data to study genetic and epigenetic effects on chromatin looping. bioRxiv, doi:10.1101/352682.

Hnisz,D. *et al.* (2016) Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science*, **351**, 1454–1458.

Imakaev,M. *et al.* (2012) Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat. Methods*, **9**, 999–1003.

Knight,P.A. and Ruiz,D. (2013) A fast algorithm for matrix balancing. *IMA J. Numer. Anal.*, **33**, 1029–1047.

Lieberman-Aiden,E. *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289–293.

Liu,X. *et al.* (2008) TiGER: a database for tissue-specific gene expression and regulation. *BMC Bioinformatics*, **9**, 271.

Lowe,D.G. (2004) Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.*, **60**, 91–110.

Ma,Y. *et al.* (2015) How the cell cycle impacts chromatin architecture and influences cell fate. *Front. Genet.*, **6**, 19.

Mikolajczyk,K. (2002) Detection of local features invariant to affine transformations: application to matching and recognition. PhD Thesis, Grenoble INPG, Grenoble, France.

O'Sullivan,J.M. *et al.* (2013) The statistical-mechanics of chromosome conformation capture. *Nucleus*, **4**, 390–398.

Patel,B. *et al.* (2012) CTCF mediated enhancer and promoter interaction regulates differential expression of TAL1 oncogene in normal and malignant hematopoiesis. *Blood*, **120**, 281–281.

Pederson,T. (1972) Chromatin structure and the cell cycle. *Proc. Natl. Acad. Sci. USA*, **69**, 2224–2228.

Phanstiel,D.H. *et al*. (2015) Mango: a bias-correcting ChIA-PET analysis pipeline. *Bioinformatics*, **31**, 3092–3098.

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

Ron,G. *et al*. (2017) Promoter-enhancer interactions identified from Hi-C data using probabilistic models and hierarchical topological domains. *Nat. Commun*., **8**, 2237.

Schmitt,A.D. *et al*. (2016) A compendium of chromatin contact maps reveals spatially active regions in the human genome. *Cell Rep*., **17**, 2042–2059.

Sexton,T. *et al*. (2012) Three-dimensional folding and functional organization principles of the drosophila genome. *Cell*, **148**, 458–472.

Shechtman,E. and Irani,M. (2007) Matching local self-similarities across images and videos. In: *2007 IEEE Conference on Computer Vision and Pattern Recognition*. pp. 1–8, 17-22 June 2007, Minneapolis, MN, USA.

Stansfield,J. and Dozmorov,M.G. (2017) HiCcompare: an R-package for joint normalization and comparison of HI-C datasets. https://doi.org/10.1186/s12859-018-2288-x.

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

Ursu,O. *et al*. (2018) GenomeDISCO: a concordance score for chromosome conformation capture experiments using random walks on contact map graphs. *Bioinformatics*, **34**, 2701–2707.

Wang,J. *et al*. (2013) Genome-wide analysis uncovers high frequency, strong differential chromosomal interactions and their associated epigenetic patterns in e2-mediated gene regulation. *BMC Genomics*, **14**, 70.

Yaffe,E. and Tanay,A. (2011) Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat. Genet*., **43**, 1059–1065.

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

Yardimci, GG. *et al*. (2018) Measuring the reproducibility and quality of Hi-C data. bioRxiv, doi:10.1101/188755.

Zhou,X. *et al*. (2014) Robustly detecting differential expression in RNA sequencing data using observation weights. *Nucleic Acids Res*., **42**, e91.

Zufferey,M. *et al*. (2018) Comparison of computational methods for the identification of topologically associating domains. *Genome Biol*., **19**, 217.