FIND: differential chromatin INteractions Detection using a spatial Poisson process

Mohamed Nadhir Djekidel,[1] Yang Chen,[1] and Michael Q. Zhang[1,2]

[1]*MOE Key Laboratory of Bioinformatics and Bioinformatics Division, Center for Synthetic and System Biology, TNLIST/Department of Automation, Tsinghua University, Beijing 100084, China;* [2]*Department of Molecular and Cell Biology, Center for Systems Biology, The University of Texas, Dallas, Richardson, Texas 75080-3021, USA*

Polymer-based simulations and experimental studies indicate the existence of a spatial dependency between the adjacent DNA fibers involved in the formation of chromatin loops. However, the existing strategies for detecting differential chromatin interactions assume that the interacting segments are spatially independent from the other segments nearby. To resolve this issue, we developed a new computational method, FIND, which considers the local spatial dependency between interacting loci. FIND uses a spatial Poisson process to detect differential chromatin interactions that show a significant difference in their interaction frequency and the interaction frequency of their neighbors. Simulation and biological data analysis show that FIND outperforms the widely used count-based methods and has a better signal-to-noise ratio.

[Supplemental material is available for this article.]

Chromatin folding constitutes one of the key mechanisms by which cells control their transcriptional program and cellular identity (Schmitt et al. 2016). Hi-C is one of the most widely adopted biochemical techniques to probe the genome-wide spatial organization of chromatin (Rao et al. 2014; Schmitt et al. 2016; Dekker et al. 2017). Computationally, the analysis of Hi-C data is challenging, largely due to the various sources of biases introduced by the various experimental steps (Imakaev et al. 2012). Therefore, much of the existing computational effort is focused on developing more reliable data preprocessing techniques, such as filtering and normalization, to extract much of the signal from the data (Hu et al. 2012; Sauria et al. 2015; Servant et al. 2015; Forcato et al. 2017). However, with the increasing accumulation of Hi-C data (Rao et al. 2014; Du et al. 2017), there is interest in performing more comparative analyses to study the structural variability between the different tissues and cellular conditions.

Surveying the literature, we noticed that there has not been a globally adopted conventional method for the detection of differential chromatin interactions (DCIs). One of the simplest strategies is to use fold change as a norm for the detection of DCIs. This strategy was generally adopted in the early Hi-C analysis papers. For example, Wang et al. (2013) used a simple fold-change strategy to detect DCIs between MCF-7 Hi-C samples before and after estrogen treatment. In a more elaborate model, Dixon et al. (2015) used large fold-change chromatin interactions to train a random-forest model to detect the epigenetic signals that are more predictive of the chromatin structural changes.

Other strategies use the binomial model to compare two normalized Hi-C contact maps and detect the pairwise interactions that show a significant change in their frequency. This type of test is adopted by the HOMER software (Heinz et al. 2010). However, in many of the recently published studies (Paulsen et al. 2014; Lun and Smyth 2015; Taberlay et al. 2016; Ulianov et al. 2016), we noticed an increasing adoption of count-based methods such as edgeR (Robinson et al. 2010) to detect DCIs. By comparing edgeR to binomial-based methods, Lun and Smyth (2015) showed that edgeR could outperform HOMER's results.

Since most of the methods for detecting DCIs were developed to analyze relatively low-resolution Hi-C contact maps (40 kb or more), the authors assumed independence among the different pairwise interactions. At low resolutions, this assumption is logical, as we only capture the spatial proximity between distant chromatin fibers. However, in the case of high-resolution contact maps (Rao et al. 2014), this assumption may break down. Due to the polymeric nature of the chromatin fiber, the establishment of a chromatin loop that brings two interacting loci (i, j) into spatial proximity will also influence the spatial distance between their adjacent loci ($i-1$, $i+1$ and $j-1$, $j+1$). In the Hi-C contact map, the distance change between the anchor points i and j should also be reflected in the interaction frequencies within the window centered around the pairwise interaction (i, j) (Fig. 1A). Clearly, in high-resolution Hi-C data, the detection of DCIs under the independence assumption can have a high error rate.

To resolve this issue, we developed a new computational method, FIND, that considers the local spatial dependency between interacting loci. FIND exploits a spatial Poisson process model to detect differential chromatin interactions that show both a significant change in their interaction frequency and the interaction frequency of their adjacent bins.

Results

High resolution Hi-C captures the spatial dependency between adjacent chromatin fibers

Improvement in the biochemistry of the Hi-C experiment, from the dilution Hi-C (Lieberman-Aiden et al. 2009) to the in situ Hi-C (Rao et al. 2014), helped reveal a more detailed snapshot of the chromatin folding and the spatial dependency among adjacent interacting loci. In the Hi-C contact map, the spatial dependency

Corresponding authors: yc@tsinghua.edu.cn, michael.zhang@utdallas.edu

Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.212241.116. Freely available online through the *Genome Research* Open Access option.

28:412–422 Published by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/18; www.genome.org

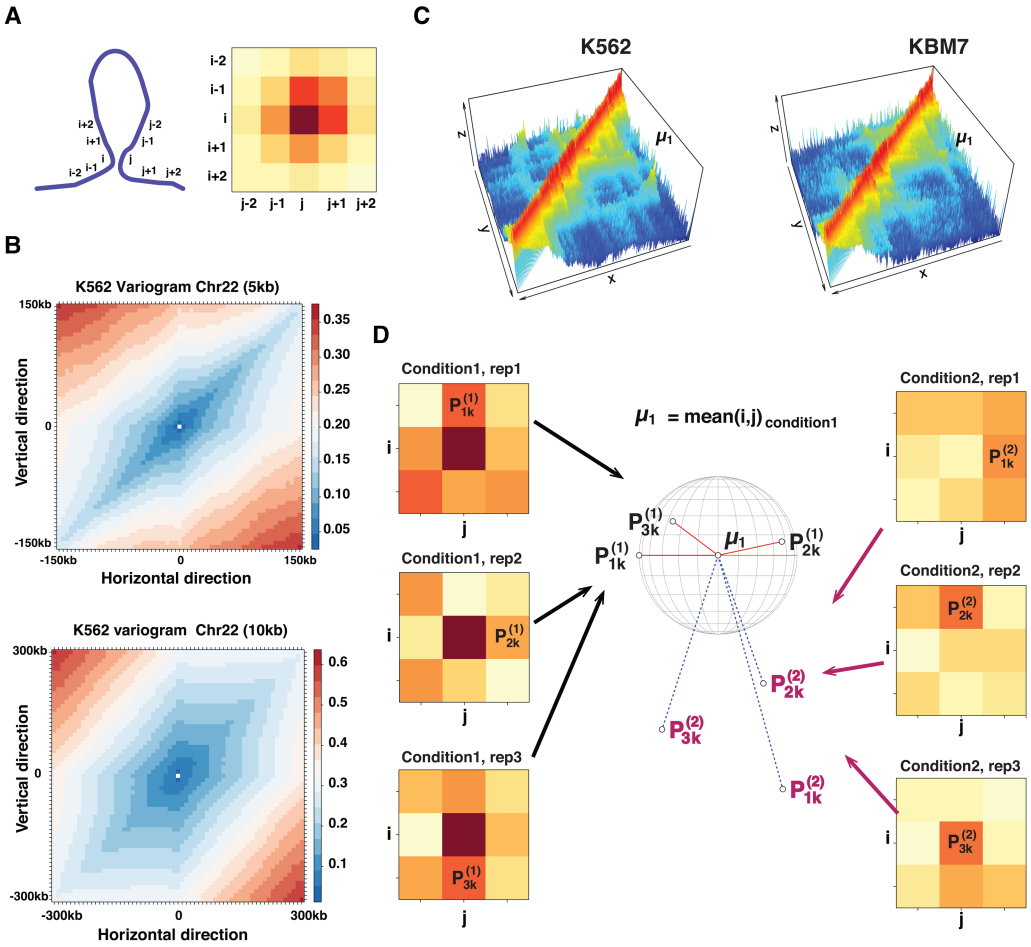

Figure 1. Existence of the spatial dependency and the idea behind our model. (*A*) Illustration of the spatial dependency along neighboring loci in the Hi-C contact map. (*B*) Semivariogram showing the directional variability between interaction bins separated by a certain horizontal and vertical distance in the Hi-C contact map. (*C*) A differential interaction can be considered as a change of the intensity around the 3D coordinate (i, j, f_{ij}) of a reference point μ_1. (*D*) Principle of the *k*-nearest neighbor (KNN) intensity estimation in a 3×3 window around a pairwise interaction (i, j). Given an interaction (i, j) with a mean frequency $\mu_1 = (i, j, \overline{f_{ij}})$ in the first condition (represented by the mountain tip in *C*), if there is no structural change, we expect the interaction frequencies from the second condition to have a similar density around the point μ_1 in the 3D space. Thus, for each condition and each replicate, we calculate the 3D distance between each bin in the surrounding window and μ_1 and order them according to their distance. We note $P^{(c)}_{nk}$ as the *k*th-nearest neighbor to the point μ_1 from the *n*th replicate of condition $c \in \{1, 2\}$. Then, we estimate the density of the KNNs around $\mu_1$in the first condition ($\lambda^{(1)}_k$) and in the second condition ($\lambda^{(2)}_k$). The density of the KNNs around μ_1 is expected to be stable between the two conditions. To decide if the change of the KNN density around μ_1 ($\lambda^{(2)}_k/\lambda^{(1)}_k$) is significantly large or small, we use a Fisher distribution. The same principle applies if we use μ_2 (the mean in condition 2) as our reference point.

between nearby chromatin fibers can be illustrated by the schematic example in Figure 1A.

Given the polymeric nature of chromatin, a loop formation is characterized by the bending of the polymer chain and the establishment of a chromatin interaction between two physically proximal loci, i and j. When the spatial distance between i and j changes, the distance between their adjacent loci (such as $i-1$, $i+1$, $j-1$ and $j+1$) will also be influenced. In the Hi-C contact map, this phenomenon should be characterized by a change in the interaction frequencies of the bins within the 2D window centered at the pairwise interaction (i, j) (Fig. 1A). In this work, we call the pairwise interactions located in this window the "neighborhood interactions" of (i, j). To quantitatively measure the extent of the spatial dependency between adjacent loci in the Hi-C contact map, we calculated the directional variogram under different Hi-C resolutions (Fig. 1B; Supplemental Figs. S1, S2). The variogram is a measure widely used in geostatistics to describe the degree of spatial correlation of a given spatial process. It is based on calculating the mean variance between the values of all the pairs of points $Z(x)$ and $Z(x+h)$ separated by a given distance h in a given map (Pebesma 2004). If we consider the Hi-C contact map as a 2D surface and the interaction frequencies as values sampled from a spatial process $Z(i, j)$, the variogram should give us an idea of the spatial dependency between adjacent bins. The distance-dependent contact probability plot generally used to describe Hi-C data informs us about the expected contact frequency along each diagonal within the Hi-C contact map. However, it does not inform us about the relationships between the interaction frequencies along the other directions in the 2D Hi-C contact map. These relationships can be captured through the variogram as it measures the mean interaction frequency variability between all the bins (i, j) and the bins separated by a distance h in the horizontal, vertical, and diagonal directions.

In Figure 1B, we show the directional variogram up to a separation distance of 30 bins calculated from 5- and 10-kb resolution-normalized Hi-C contact maps of Chromosome 22 in the cell line K562 (Rao et al. 2014). We notice that there is a small amount of variability among adjacently located pairwise interactions that are separated by short distances in all the directions of the 2D Hi-C contact map. We also notice that the interaction frequencies vary much more slowly along the diagonal than in the other directions. This indicates that the "zone of influence" of an interaction (i, j) is limited to its neighborhood and is independent from the distal interactions. The radius of the influence zone of (i, j) decreases with increasing resolution (Supplemental Figs. S1, S2), which confirms that in high-resolution analysis, the dependency between neighboring bins cannot be ignored.

Based on these observations, we built our differential chromatin interaction detection model to take into account the spatial dependences among neighboring loci.

Proposition of a spatial Poisson process to model differential chromatin interactions

If we consider the interaction frequencies of the Hi-C contact map in a three-dimensional parameter space in which the x-axis and y-axis represent the genomic coordinates and the z-axis represents the interaction frequencies, then the highly interacting regions would form mountain-like structures (Fig. 1C). In the case of a differential interaction, we would expect to see a significantly correlated change in the "mountain" shape (Fig. 1C). Conversely, if the difference is due to technical noise, we would expect a more random shape change in which the frequency change of a pairwise interaction will have no effect on its adjacent interactions.

Intuitively, to estimate this shape change between two conditions, we can take the three-dimensional location of the tip of the "mountain" in the first condition as a point of reference. Then, we calculate the change in the density of points around it between the first and second conditions (Fig. 1D).

Given an experimental design in which the Hi-C contact maps are generated in two biological conditions, each with n_c replicates ($c \in \{1, 2\}$). Let (i, j) be a pairwise interaction of interest for which we want to check the differential interaction state, and let W be the width of the window centered around it. For example, in the case of a window of width 3, the window will include all nine pairwise interactions with coordinates in the Cartesian product between the loci $\{i-1, i, i+1\}$ and $\{j-1, j, j+1\}$. We also define μ_1 as the mean interaction frequency of the pairwise interaction (i, j) in experimental condition 1 and μ_2 as the mean interaction frequency in experimental condition 2.

In the case of no differential interaction, the interaction frequencies within the window centered at (i, j) are expected to be similarly distributed in the two conditions around μ_1. Hence, we expect the probability of observing the k-nearest neighbor (KNN) at a distance $x \in \mathbb{R}^3$ from the reference point (i, j, μ_1) to be similar between the two biological conditions. Under the assumption that the neighboring interaction frequencies are sampled from a homogeneous spatial Poisson process, the probability of observing the KNN at a distance x depends only on the density of the KNNs around the reference point (i, j, μ_1) (Methods). Thus, in a window of width W, we consider the interaction (i, j) to be a DCI, if the majority of the KNNs show a significant change in their intensity.

More specifically, using the triplet (i, j, μ_1) as our reference and for each replicate in each condition, we can rank the W^2 interactions in the neighborhood of (i, j) according to their distance from the point (i, j, μ_1). Let $P_{nk}^{(c)}$ indicate the kth-nearest neighbor of (i, j, μ_1) in the nth replicate of condition c (Fig. 1D). For a fixed k (for example, the first nearest neighbor), we can use the point intensity estimator developed by Burguet (Burguet et al. 2009, 2011; Burguet and Andrey 2014) to estimate the intensity $\lambda_k^{(1)}(\mu_1)$ of the KNNs $P_{nk}^{(1)}$ from the first condition and the intensity $\lambda_k^{(2)}(\mu_1)$ of the KNNs $P_{nk}^{(2)}$ from the second condition around μ_1 (Methods). We consider the KNNs to show a change in their intensity at μ_1 if the ratio $R_k(\mu_1) = \lambda_k^2(\mu_1)/\lambda_k^1(\mu_1)$ is significantly different from unity. Under the null hypothesis that $\lambda_k^1(\mu) = \lambda_k^2(\mu)$, the ratio $R_k(\mu_1)$ can be shown to follow a Fisher distribution with $2n_1$ k and $2n_2$ k degrees of freedom, respectively (Burguet et al. 2009; Methods).

Using the same window and the points (i, j, μ_1) and (i, j, μ_2) as references, we can calculate $2W^2$ P-values for the pairwise interaction (i, j). These P-values will then be combined using the rth ordered P-value statistic (Song and Tseng 2014). The final false discovery rates will be estimated using the Benjamini-Hochberg method (Benjamini and Hochberg 1995; Methods).

Neighboring pairwise interactions have different sensitivity to the Hi-C contact frequency changes

To understand the behavior of our model and to check the contribution of each k-nearest neighbor to the final DCI decision, we simulated two Hi-C conditions, each with two replicates (Methods). Then, we used a major voting strategy in which a pairwise interaction (i, j) is considered a DCI if at least half of the KNNs in the window centered at (i, j) show a significant change in their density around μ_1 or μ_2.

To assess the sensitivity of each KNN to the interaction frequency change, we used a window of width 3 and plotted the pairwise interactions that show a significant density change in their kth-nearest neighbor ($k \in [1, 9]$, P-value <0.001) between the two conditions (Fig. 2A). We observe that the furthest neighbors are more sensitive to the variability in interaction frequency, because they tend to report more significant interaction changes, whereas the nearest neighbors are less sensitive. The analysis of the distribution of the P-values obtained from each KNN in Figure 2A indicated a similar conclusion (Fig. 2B), with distant neighbors having many of their P-values located at the extremities outside of the null hypothesis acceptance region

$$\left[R_k^{\frac{\alpha}{2}},\ R_k^{1-\frac{\alpha}{2}}\right],$$

and the nearest neighbors have more of their P-values more inside the acceptance region (Fig. 2B). In fact, this sensitivity is due to the speed of convergence of the Fisher-distribution cumulative density function (CDF), as illustrated in Figure 2C. We noticed that the smaller the degree of freedoms is, the slower is the convergence of the CDF to 1. Thus, for smaller k values, in order for the ratio $R_k(\mu)$ to be significant, it needs to be either very small ($\hat{\lambda}_k^1(\mu) \ll \hat{\lambda}_k^2(\mu)$) or very large ($\hat{\lambda}_k^1(\mu) \gg \hat{\lambda}_k^2(\mu)$). In contrast, the distant neighbors need to have only a small amount of variability to be in the null hypothesis rejection zone. This indicates that we need a decision scheme that accounts for the neighbors' sensitivity.

Different strategies can be adopted to combine the $2W^2$ P-values calculated for each pairwise interaction (i, j). Traditional methods such as Fisher's combined probability test (Fisher 1925) and Stouffer's Z-test (Riley et al. 1949) or their weighted variants can be used; however, they are designed to report a significant P-value if at least one of the $2W^2$ tests is non-null, which makes them very

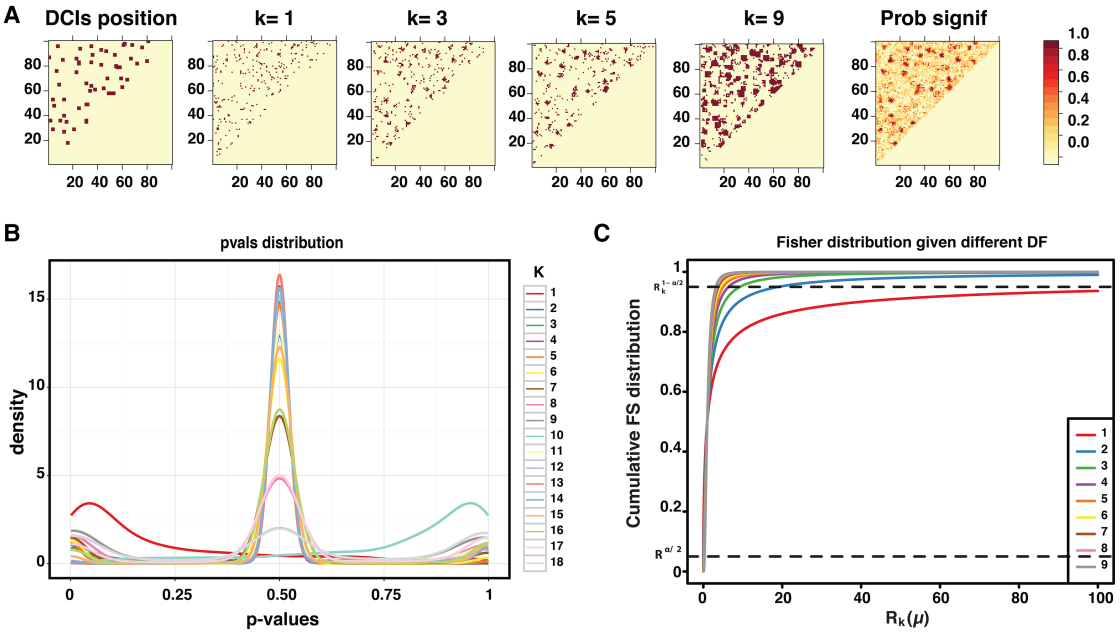

Figure 2. Sensitivity of the different KNN to variability. (*A*) Heatmaps showing the sensitivity of the different KNNs to variability. The *left* heatmap indicates the position of the simulated differential interactions. The heatmaps labeled from $k = 1$ to $k = 9$ show the positions of the pairwise interactions showing a significant difference from their *k*th-nearest neighbor (*P*-value <0.001). The *right* heatmap shows the majority vote heatmap of the KNNs. We noticed that the furthest neighbors give noisier predictions. (*B*) The distribution of the *P*-values of the KNN heatmaps calculated in *A*. Plots 1–9 are the *P*-values obtained using μ_1 as reference, and plots 10–18 are the *P*-values obtained using μ_2 as reference. (*C*) Plot showing the convergence of the Fisher distribution using *k* as the degrees of freedom ($k \in [1, 9]$). We noticed that the larger the *k* is, the faster the Fisher distribution converges out of the acceptance zone.

sensitive to changes in the farthest KNNs. Other methods, such as the maxP approach (Wilkinson 1951) are too stringent, as this method considers only the largest ordered *P*-value among the $2W^2$ tests and potentially misses many differential interactions. Ideally, a weighted *P*-value alternative should be adopted; however, they tend to be computationally demanding. A good tradeoff is the use of the *r*th ordered *P*-value (rOP) method (Song and Tseng 2014). The rOP method considers a collection of tests to be significant if at least *r* of them are significant. Under the null hypothesis, the rOP tests on the *r*th order statistic among the sorted $2W^2$ *P*-values using beta distribution with degrees of freedom $\alpha = r$ and $\beta = 2W^2 - r + 1$. In our model, the value of *r* is equal to the largest integer value smaller than $p \times W^2$, such that $p \in (0, 1]$ is the percentage of the significantly variable KNNs between the two biological conditions required for a pairwise interaction to be considered a DCI.

The spatial dependency model shows more accurate detection of DCIs compared to the spatial dependency-free models

To assess the accuracy of our model, we compared its performance to the edgeR method as a representative of models that assume the independence between the Hi-C pairwise interactions. We did a simulation analysis in which we generated a Hi-C experiment with two conditions, each with two replicates. We used the counts of the K562 Hi-C at 5-kb resolution as a reference to generate the simulated interaction counts (Methods). Then, given the known positions of the differential and nondifferential interactions, we calculated the area under the curve (AUC) values to assess the performance of FIND versus edgeR given different window sizes (*W*), the percentage of significantly variable KNNs (*P*), and fold-change (FD) values. For each fixed setting, the simulation was repeated 10 times. By summarizing these results in Figure 3A, we observed that for a window less than or equal to 5 bins (25 kb), FIND outperforms edgeR. However, depending on the selected proportion of significantly differential KNNs, the performance of FIND varies. When we required all the neighboring pairwise interactions in the sliding window to be significantly variable ($P = 1$), which is also equivalent to the maxP statistic, we observed that FIND still maintains a good DCI detection performance but drops significantly compared to other percentage values. This strict setting could be useful to detect reliable DCIs, but it would also miss many other important DCIs. On the other hand, we noticed a lower performance if we only required a small proportion of the neighboring pairwise interactions ($P < 0.5$) to be differential, the result is comparable to those of Fisher's combined probability test and Stouffer's *Z*-test.

When the window size is very large, >35 kb, FIND's performance degrades and shows generally worse performance than edgeR. This behavior can be explained by the loss of the local dependency between the interacting chromatin fibers at larger resolutions. This relationship can be seen in the 5-kb variogram in Figure 1B, in which we observed high variance between bins with increasing distance. These results indicate that FIND is suitable for high-resolution Hi-C data, whereas in the case of a low-resolution contact map, a count-based method such as edgeR can be sufficient.

As edgeR is independent from the window size and the confidence level α, it has similar performance for a fixed fold-change value. By allowing differential interaction to have increasing fold-change values, we observed that edgeR performs similarly to FIND in high fold-change regions. For example, in the case in which we allowed the DCI bins to have a fold change of 10 or more, edgeR and FIND performed similarly. This indicates that edgeR is more suitable to detect very significant changes that

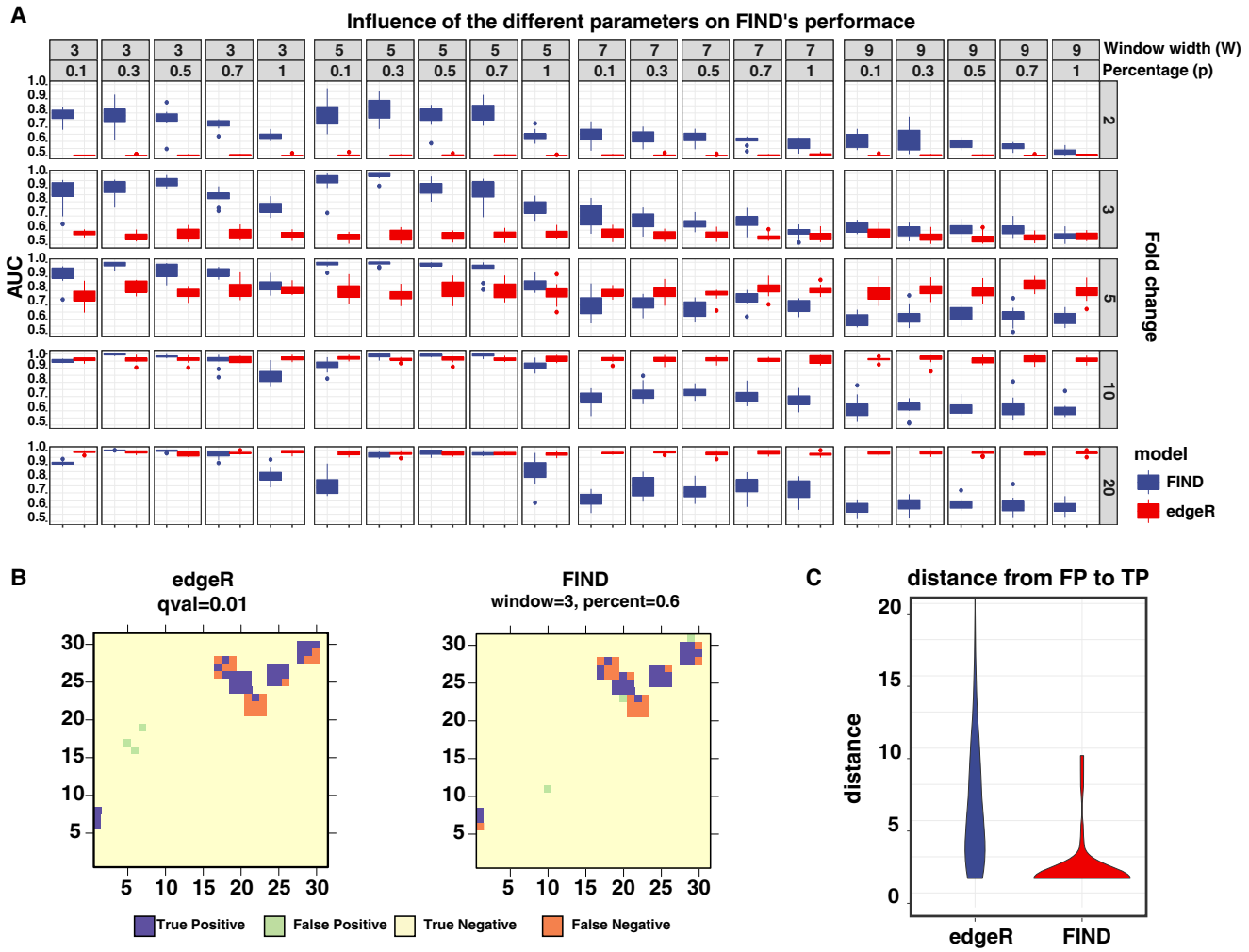

Figure 3. Performance comparison between FIND and edgeR on simulated data. (*A*) Tile plot comparing the performance between FIND and edgeR given different window sizes (*W*), percentages of significantly variable KNNs (*P*) and fold-change values. For each fixed configuration, we ran 10 simulations and calculated the box plots shown in each tile. (*B*) An example simulation showing the positioning of the reported DCI by edgeR and FIND (blue and green) relative to the real signal (blue and orange) in a case in which FIND and edgeR have similar performance. We observed that regions reported by edgeR tend to be located far from the real signal. (*C*) Distribution of the distances of edgeR and FIND's false-positive signals relative to true DI regions. We observed that edgeR DIs tended to be more scattered in the heatmap, whereas FIND's results tended to be near the real signal.

have a visible impact on the neighborhood. However, edgeR still misses many of the DCIs that show lower fold change but have a significant structural change in their neighborhoods.

We then looked at the case in which edgeR and FIND have comparable behavior (high fold-change values) to check which method was more reliable. In other words, how far are the false-positive values from the real signal? In Figure 3B, we show an example of the location of the wrongly predicted DCI (shown in green) to the accurately simulated DCIs (blue and orange). We clearly see that FIND's false positives are located near the real signal, whereas edgeR false positives tend to be located far away. Most of FIND's false positives are located one bin away from the real signal, while for edgeR, the false-positive predictions tend to be located far away (Fig. 3C). This observation indicates that the DCIs reported by FIND tend to be in the neighborhood of regions of high structural variability, and edgeR has more potential to detect technical noise in the Hi-C contact map, which can lead to more erroneous conclusions.

To sum up, these simulation results reveal the advantage of our method over the count-based methods for high-resolution Hi-C contact map as it uses the nearest neighbors' information to decide about the significance of a structural change. Compared to count-based methods, this strategy avoids the detection of noisy DCIs, because it borrows information for adjacent interactions. We also showed that in high resolution Hi-C data, count-based methods may be used, but they require more cautious manipulation as detailed in Lun and Smyth (2015).

FIND shows more reliable behavior than count-based methods on real data

To assess the reliability of our model on real data, we compared the 5-kb resolution Hi-C contact maps of K562 and GM12878 cells (Rao et al. 2014). For each cell line, we used two replicates. The data were normalized using the Knight-Ruiz matrix-balancing algorithm (Rao et al. 2014). Additionally, to remove the between-

replicates variability, we normalized the replicates of each condition using the MA-plot strategy (Methods).

We found that FIND detects approximately 1.6 times more DCIs than edgeR (Fig. 4A). A reasonable explanation is that edgeR missed some of the differential interactions with relatively small fold change, whereas FIND can consider these interactions as true differential interactions because it is backed up by the information from the neighboring interactions.

Several Hi-C-based comparative studies indicated that the majority of the chromatin structural changes tend to occur within topologically associated domains (TADs) (Rao et al. 2014; Dixon et al. 2015; Smith et al. 2016). Hence, we checked the span of the reported DCIs by both methods, and the results indicated that the majority of the DCIs detected by FIND have an interaction span <1 Mb (mean span 58,229.4 kb); for edgeR, approximately 20% of the interactions span >1 Mb, with a mean span of 107,555 kb (Fig. 4B; Supplemental Fig. S4). We also calculated the proportion of DCIs located within TADs (Fig. 4C; Supplemental Fig. S5). The results are consistent with Figure 4B, showing that ~70% of the DCIs detected by FIND are located within TADs; for edgeR, only ~20% of the DCIs are within TADs. Additionally, because CTCF is a master controller of the chromatin architecture (Rao et al. 2014; Tang et al. 2015), we expected the DCIs to be located in the neighborhoods of the differential CTCF peaks. Thus, we calculated the distance of edgeR- and FIND-detected DCIs to the differential CTCF peaks (Fig. 4D). Compared with edgeR, FIND has a larger proportion of peaks located <100 bp from the CTCF peaks. All of these results indicated that FIND tends to detect larger numbers of reliable DCIs than edgeR.

Topological changes have a large effect on the cross-talk between enhancers and promoters that can alter gene expression (Rao et al. 2014; Dixon et al. 2015). We classified genes located in the proximity of DCIs according to their expression fold change (FC) into two categories: FC ≤ 2, genes that did not show a significant difference in expression between the two cell lines, and FC > 2, genes that have a noticeable difference in expression. For FIND, 71.46% of the DCI-related genes show a significant expression change (FC ≥ 2); for edgeR, approximately 50.63% of the DCI-related genes show an expression change between the cell lines (Fig. 5A). Consistent with transcription results, DCIs detected by FIND are closer to active gene signals such as H3K4me3, Pol II binding sites, and EP300 (Fig. 5B–D). All of these results indicated that FIND tends to detect larger numbers of functional DCIs than edgeR.

Interestingly, the functional analysis of the genes located in the proximity of DCIs (≤5 kb) detected by FIND, shows a high enrichment for GO terms related to the immune system (Table 1). Meanwhile, for edgeR, only the term "immune system process" showed a significant enrichment. These results further support the reliability of FIND. These findings are consistent with previous reports that indicate that many of the H3K4me1 peaks overlap with known autoimmune disorder SNPs in the B lymphoblast cell line GM12878 (Corradin et al. 2014).

Role of chromatin structure in K562 differentiation

The role of chromatin loops is well characterized at the classical alpha-globin locus, which is known for its exclusive expression in erythroid cells (K562) and its silencing in lymphoblastoid cells (GM12878) (Fig. 6; Vernimmen et al. 2007; Baù et al. 2011). Genes at this locus are regulated by a cluster of remote DNase I hypersensitive sites (HSs) located approximately 30–60 kb upstream of the alpha-globin genes (Vernimmen et al. 2007), and the silencing of these genes is due to the absence of the enhancer-promoter interaction.

In Figure 6, we marked the different results reported by FIND using the *P*-value cutoff of 1×10^{-4}. We notice that all the reported

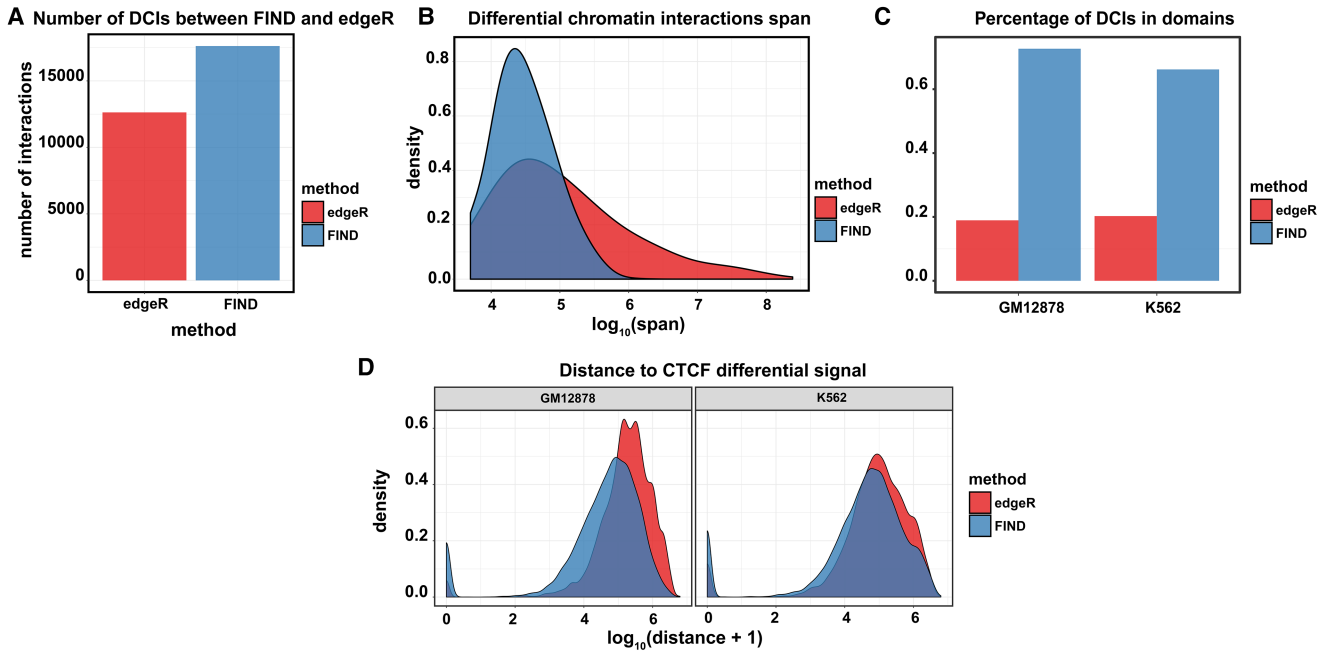

Figure 4. Performance comparison between FIND and edgeR with genomic characteristics. (*A*) The numbers of DCIs detected by FIND and edgeR. (*B*) The distribution of the span of the DCIs reported by edgeR and FIND. We observed that the majority of FIND's results have a span <1 Mb. (*C*) The proportion of the genome-wide DCIs located inside TADs for both edgeR and FIND. (*D*) The distances of edgeR- and FIND-reported DCIs to the CTCF differential peaks.

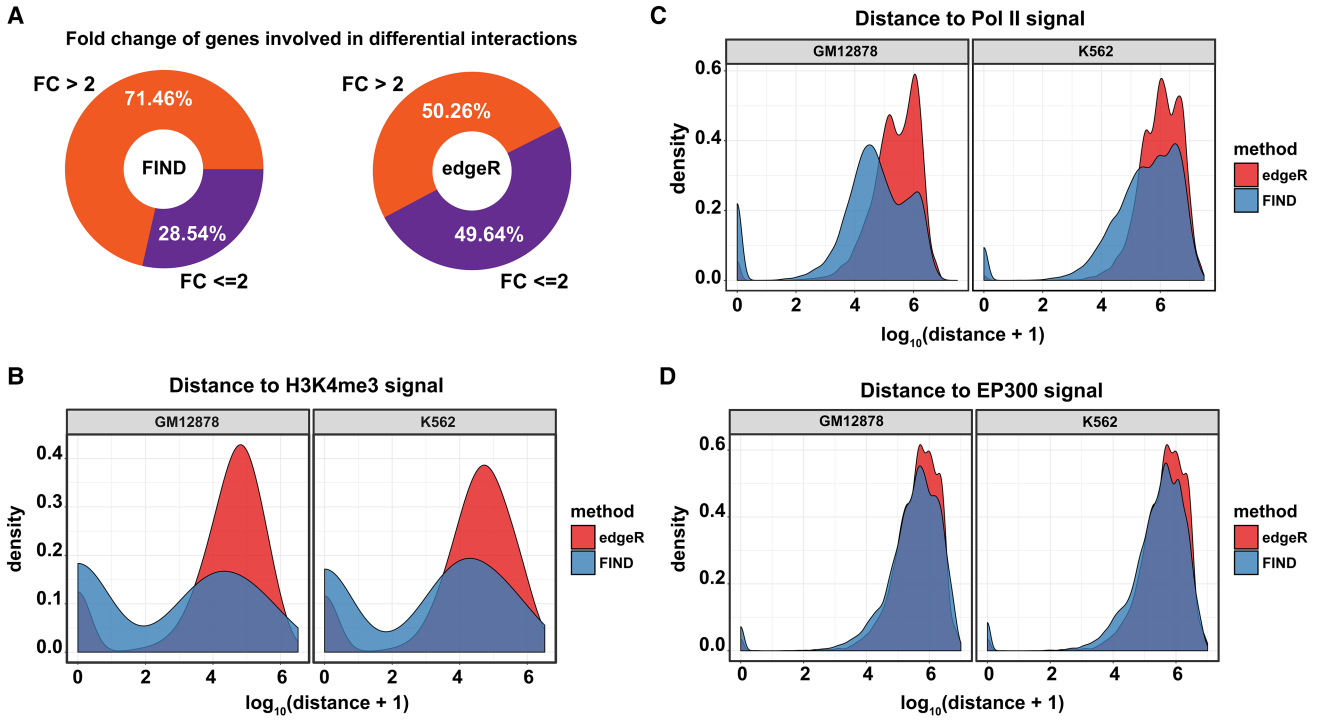

Figure 5. Comparison of performance between FIND and edgeR with transcriptional characteristics. (*A*) The proportion of genes with different fold changes located near the DCIs. (*B*) The distribution of the distances between DCIs and the H3K4me3 peaks. (*C*) The distribution of the distances between DCIs and the Pol II peaks. (*D*) The distribution of the distances between DCIs and the EP300 peaks.

differently interacting regions are located around interactions that show a significant differential binding of the CTCF protein. ChromHMM tracks at this region also indicate a major change in chromatin state (Ernst and Kellis 2012), from a strong enhancer state (shown in orange in the ChromHMM track) to a heterochromatin state (shown in gray in the ChromHMM track). In addition to the alpha-globin locus, the localization of FIND's result around differential CTCF binding sites can be observed in many regions (Supplemental Fig. S6).

Job Dekker and his group used the low-throughput 5C experiment to investigate the chromatin cross-talk between selected regions in different cell lines (Sanyal et al. 2012). The comparison of our results in the proximity of the differential 5C interactions indicates that FIND'S DCIs are closer to the differential 5C peaks than edgeR ones (Supplemental Fig. S7).

Discussion

Here, we present a novel computational method that detects differential chromatin interactions between two Hi-C experiments. We argue that in high-resolution Hi-C maps, the spatial dependency between neighboring interactions should be considered. First, we used the directional semivariogram metric to verify the existence of the neighborhood dependency in Hi-C data. Then, taking this relationship into account, we developed a computational method that detects differential chromatin interactions that show a correlated change with the pairwise interactions within the surrounding window.

In our model, we consider the Hi-C matrix in the 3D space, which enabled us to consider an interaction change as a change in intensity (height). Then, we exploited a spatial Poisson process to estimate the changes in the intensities of the *k*-nearest neighbors around the pairwise interactions between two Hi-C conditions. We showed that this change can be estimated using a Fisher distribution. Given the Fisher cumulative distribution function (CDF), we showed that more-distant neighbors show more sensitivity to change due to the fast convergence of their associated CDF. Therefore, we used the *r*th ordered *P*-value (rOP) method to minimize the effect of this sensitivity.

To assess the performance of our method, we performed some simulation analyses in which we compared the performance of our method to the widely adopted count-based method edgeR. We showed that, in general, our method outperformed edgeR. For high fold-change interactions, we showed that there is essentially no large performance gap between our model and edgeR; however, our false positives tend to localize near the real differential interactions, whereas edgeR false positives tend to be scattered along the heatmap.

Table 1. Functional enrichment of genes near the DCI sites predicted by FIND

GO term	*Q*-value	Type
GO:0002376 immune system process	3.83×10^{-18}	Up
GO:0001775 cell activation	3.48×10^{-7}	Up
GO:0002520 immune system development	4.86×10^{-5}	Up
GO:0002521 leukocyte differentiation	1.27×10^{-4}	Up
GO:0002682 regulation of immune system process	4.08×10^{-4}	Up
GO:0001816 cytokine production	1.141×10^{-3}	Up
GO:0006325 establishment or maintenance of chromatin architecture	0.02	Down

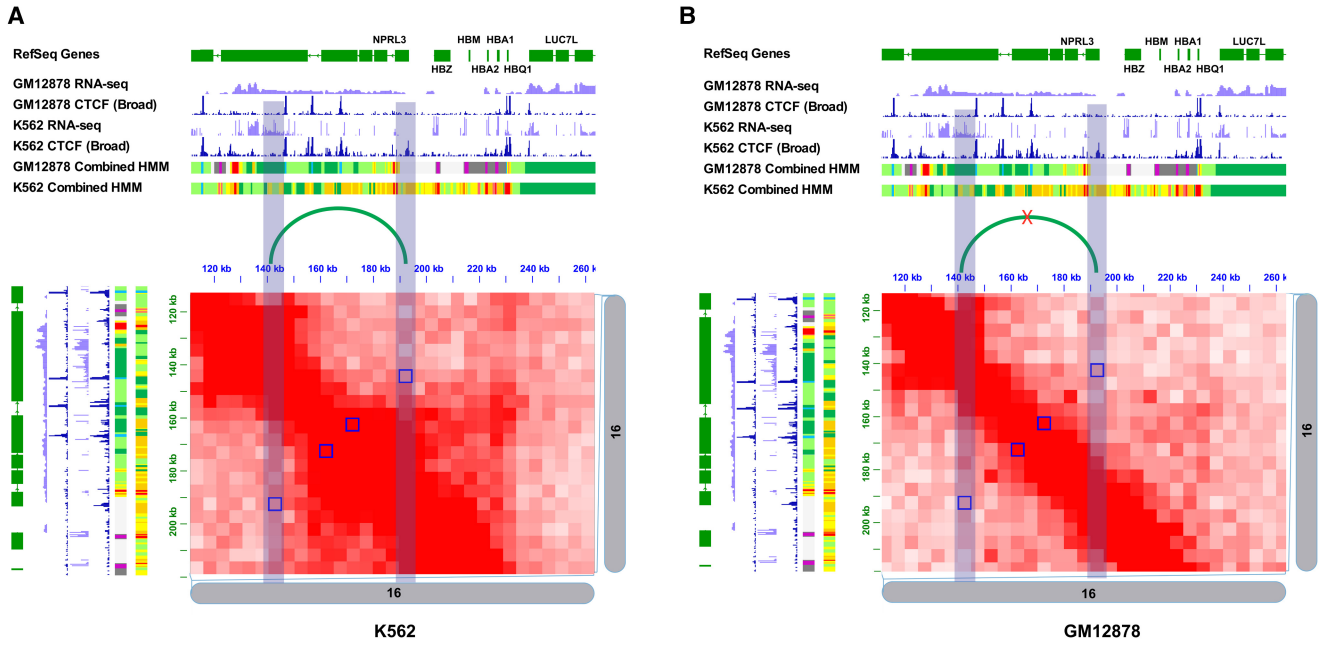

Figure 6. DCIs in the alpha-globin region as detected by FIND between K562 and GM12878 cells. (*A*) Hi-C contact map of the K562 at the alpha-globin locus. Differential interactions (Q-value $<1 \times 10^{-4}$) are shown in blue squares; the CTCF signal change region is highlighted by the blue bars. (*B*) The corresponding region shown in the GM12878 with the same regions highlighted.

In addition to the simulation experiments, we tested our method on Hi-C interaction data comparing K562 and GM12878 cells. We showed that the DCIs detected by FIND tend to be more located inside TADs, whereas a large proportion of the edgeR results are outside TADs. Additionally, the DCIs reported by our method are located near differential CTCF binding cites and are associated with differentially expressed genes.

The increasing availability of high-resolution chromatin conformation data opens the door to understanding the principles that govern the spatial organization of the chromatin between different species and cell types. However, with the lack of differential chromatin interaction detection tools, it is hard to make significant conclusions. Using existing DCI detection methods without considering the spatial dependence between neighboring interactions may be prone to serious errors when analyzing high-resolution contact maps. Our tool, FIND, has resolved this issue, hence presenting a valuable tool for many investigators.

However, more room remains for improvement; for example, the methods can be extended to allow comparisons of more than two conditions. It would also be beneficial to be able to remove the rOP statistics step and replace it with a unified statistic.

Methods

Hi-C data analysis

We used Hi-C data published by Lieberman-Aiden's group (Rao et al. 2014). For the GM12878 cell line, we used the samples GSM1551574 and GSM1551575; for K562, we used the samples GSM1551620 and GSM1551623. Hi-C matrices were normalized using the VC-squared method available in the Juicebox tool (Durand et al. 2016). For each condition, inter-sample normalization was performed using the MA-plot approach (described below).

For data visualization, we used the R Environment for Statistical Computing (R Core Team 2016) for the generation of Figures 1 and 2 and Supplemental Figures S1–S3; for Figure 6 and Supplemental Figure S6, we used the Juicebox tool (Durand et al. 2016).

Gene expression data

The gene expression data were obtained from the ENCODE project (The ENCODE Project Consortium 2012) with accession numbers GSE78553 for GM12878 cells and GSE78625 for K562 cells.

Gene set enrichment

The gene set enrichment was done using the GAGE method (Luo et al. 2009).

Simulation analysis

To simulate the contact frequencies of the different replicates, we used the K562 Hi-C heatmap as a reference. For each pairwise interaction (i, j), we used a negative-binomial distribution with a dispersion of 1×10^4 using the R function rnbinom. The nondifferential interactions are sampled from a negative binomial with a mean equal to the value of the corresponding pairwise interaction in the K562 matrix, whereas the differential interactions are sampled from a negative binomial with a mean equal to the fold change of their corresponding pairwise interaction in the K562 Hi-C contact map. We tried to make the simulated DCIs as sparsely distributed as possible by selecting a small number of interactions to be DCIs (approximately 1%). Among these DCIs, 40% showed an increase in their interaction count with a given fold-change value. The obtained new mean value (for the DCIs) is used to sample the frequency counts of the DCI region of the replicates in the second condition. Then, we applied a Gaussian smoother around the DCI bins to simulate the effect of changes in the neighbors. We also allowed the non-DCI sample to be an outlier with a probability of 10%.

Semivariogram calculation

The semivariogram calculation was done using the gstat (Pebesma 2004) package in R. The variogram (or semivariogram) measure is widely used in geostatistics to describe the degree of spatial correlation of a given spatial process. If $\gamma(x)$ is a spatial process defined in a two-dimensional space, the variability between all pairs of points separated by a distance h can be calculated as follows:

$$\gamma(h) = \frac{1}{2\,|N(h)|}\sum_{x}\left[\gamma(x) - \gamma(x+h)\right]^2. \quad (1)$$

For example, x and $h \in \mathbb{R}^2$ and $N(h)$ is the set of all points separated by h, and $|N(h)|$ indicates the size of this set.

Testing the significance of the *k*th-nearest neighbor intensity change

The k-nearest neighbor intensity change test is based on the k-nearest neighbor density estimator developed by Burguet et al. (Burguet et al. 2009, 2011; Burguet and Andrey 2014). In this part, we summarize their main model. If we consider the Hi-C interactions in the 3D space, each interaction can be represented by the Cartesian coordinates (i, j, f_{ij}), such that i and j are the genomic coordinates and f_{ij} is the interaction frequency. Then, from Burguet's work, we know that for a given pairwise interaction (i, j) with an interaction frequency μ, the probability of observing the kth-nearest neighbor at the distance x_{ik} from (i, j, μ) in the nth Hi-C replicate is

$$f(x_{n,k}) = \frac{(4\lambda\pi)^k}{3^{k-1}\,(k-1)!}\,x_{n,k}^{k-1}\exp\left(-\lambda\frac{4\pi}{3}x_{n,k}^3\right). \quad (2)$$

In Equation 2, we can see that the only parameter that needs to be estimated is the density $\lambda(\mu)$, written as λ for clarity. Given the Hi-C matrix of the experimental condition c with n_c replicates, the density of the kth-nearest neighbor around the point μ can be estimated from the maximum likelihood of Equation 2, which gives

$$\hat{\lambda}_k^{(c)}(\mu) = \frac{n_c k - 1}{\frac{4\pi}{3}\sum_{n=1}^{n_c} x_{n,k}^3}. \quad (3)$$

By doing some variable change and algebraic manipulation (detailed in Supplemental Methods), we can show that under the null hypothesis the ratio between the kth-nearest neighbor density at μ in the first and second conditions follows a Fisher distribution with $2n_1\ k$ and $2n_2\ k$ degrees of freedom:

$$\hat{R}_k(\mu) = \frac{\hat{\lambda}_k^{(2)}(\mu)}{\hat{\lambda}_k^{(1)}(\mu)} \sim FS(2n_1k, 2n_2k). \quad (4)$$

Thus, given Equation 4 and given a confidence level α, the two-sided P-value that $\hat{R}_k(\mu)$ is significant will be equal to

$$P\text{-value}(k) = 2 \times \min\left(\Pr\left(R_k(\mu) < \hat{R}_k(\mu)\right),\ \Pr\left(R_k(\mu) > \hat{R}_k(\mu)\right)\right). \quad (5)$$

FIND's algorithm

Consider two Hi-C experiments performed under two conditions c_1 and c_2 each with $n_1$and n_2 replicates. For a given pairwise interaction (i, j) between two genomic bins i and j, let μ_1 be the mean interaction frequency of (i, j) in c_1, and let μ_2 be the mean interaction frequency of (i, j) in c_2. Let W be the size of the window around (i, j). The window around (i, j) will then be of size W^2 and will include all the pairwise interactions of their coordinates in the Cartesian product

$$\left\{i - \frac{W}{2}, \ldots, i, \ldots, i + \frac{W}{2}\right\} \times \left\{j - \frac{W}{2}, \ldots, j, \ldots, j + \frac{W}{2}\right\}, \quad (6)$$

where $W/2$ indicates the largest integer value smaller than $W/2$.

Using the mean interaction frequency of (i, j) in the first biological condition μ_1 as a point of reference, we can associate with each pairwise interaction in our defined window a P-value that indicates if it significantly changes around μ_1 between the two conditions. A total of W^2 P-values will be obtained. Then, using the mean interaction frequency of (i, j) in the second biological condition μ_2 as a point of reference, we can also calculate W^2 P-values that indicate the change of the interactions in the defined window around μ_2. In total, $2W^2$ P-values will be obtained.

The rth ordered P-value (rOP) statistic will be used to estimate the probability that r out of the $2W^2$ tests are significant. We define r as $r = p \times 2W^2$, where $p \in (0, 1]$ is the percentage of the significantly variable KNNs between the two biological conditions required for a pairwise interaction to be considered as a DCI. Briefly, given $2W^2$ P-values estimated through $2W^2$ tests, the rOP statistic defines the following hypothesis setting HS_r:

$$HS_r : \left\{ H_0 : \bigcap_{k=1}^{2W^2}\{\theta_k = 0\} \text{ versus } H_1 : \sum_{k=1}^{2W^2} I\{\theta_k \neq 0\} \geq r \right\}, \quad (7)$$

where θ_k is the effect size of the test k. If S_r is the rth order statistic of P-values, then

$$S_r \sim \text{Beta}(r,\ 2W^2 - r + 1)\,. \quad (8)$$

ROC calculation

We used the ROCR package to estimate the prediction performances of each of edgeR and FIND. The true-positive signals are the regions simulated to be DCI and reported to be DCI by the algorithm. The false-positive signals are the regions that are not DCI in the simulation but reported as DCI by the algorithm. The false-negative signals are the regions simulated as DCI but reported as not DCI by the algorithm. The true negatives are the regions not DCI in the simulation that are reported as not DCI by the algorithm.

MA normalization

To ensure consistency between the replicates of the same condition, we performed a between-samples normalization. The procedure is similar to the MA-plot normalization for gene expression. Briefly, for each interaction point (i, j) with interaction frequencies $f_{ij}^{(1)}$ and $f_{ij}^{(2)}$ in replicate one and replicate two, respectively, we calculated the log intensity (A) and the log ratio (M) as follows:

$$\begin{cases} A = \dfrac{\log2\left(f_{ij}^{(1)} f_{ij}^{(2)}\right)}{2} \\ M = \log2\left(\dfrac{f_{ij}^{(1)}}{f_{ij}^{(2)}}\right) \end{cases} \quad (9)$$

The expected bias, M_{bias}, is then estimated by fitting a loess curve in the MA-plot. The corrected M-value, $M_{correct} = M - M_{bias}$, is calculated, and the rescaled values of $f_{ij}^{(1)}$ and $f_{ij}^{(2)}$ are calculated as follows:

$$\begin{cases} \widehat{f_{ij}^{(1)}} = 2^{(A + 0.5 \times M_{correct})} \\ M = 2^{(A - 0.5 \times M_{correct})} \end{cases} \quad (10)$$

5C data processing

The five interaction matrices were downloaded from the GEO under the accession number GSE39510. Both the first and second sets of primers, ENm and ENr, were used. We considered an interaction to be differential in 5C if the interaction between two primers is absent in K562 or GM12878.

Selection of the *Q*-value threshold in the real data case

Two possible ways can be used to select the cutoff *Q*-value. One is using a hard cutoff. In this approach, it is advisable to use a *Q*-value that corresponds to the rOP statistic. For example, if we require that the *r*th *P*-value in each window should not be larger than 1×10^{-3}, we can use a *Q*-value cutoff of Beta($1\times10^{-3}|\alpha=r,\ \beta=2W^2-r+1$).

Using a hard cutoff can sometimes be very stringent and generally will remove some long-range DCIs due to the relatively weak signal of the large interactions. Moreover, some of the long-range DCIs will show a significantly different *Q*-value than their counterparts that have the same interaction span. Thus, we used quantile regression to model the relationship between interaction-span and *Q*-value (Supplemental Fig. S8). In our analysis, all the *Q*-values above the 99th percentile were considered, and none were larger than Beta($1\times10^{-3}|\alpha=13,\ \beta=18-13+1$).

Software availability

The software is published under the GNU GPL v3.0 license. The source code of FIND is available in the Supplemental Material and at https://bitbucket.org/nadhir/find.

Acknowledgments

This work is supported by the National Basic Research Program of China (2017YFA0505503), the National Natural Science Foundation of China (91729301, 31671384, 31301044, and 81630103), and Tsinghua National Laboratory for Information Science and Technology Cross-Discipline Foundation. The authors thank Professor Minping Qian (Beijing University) for her insightful discussions and Zhengyu Liang (Tsinghua University) for his discussion and criticism.

Author contributions: M.N.D., Y.C., and M.Q.Z. conceived the project. M.N.D. implemented FIND and wrote the analysis scripts. Y.C. helped with the data analysis. M.N.D. and Y.C. wrote the manuscript with contributions from all authors.

References

Baù D, Sanyal A, Lajoie BR, Capriotti E, Byron M, Lawrence JB, Dekker J, Marti-Renom MA. 2011. The three-dimensional folding of the α-globin gene domain reveals formation of chromatin globules. *Nat Struct Mol Biol* **18:** 107–114.

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

Burguet J, Andrey P. 2014. Statistical comparison of spatial point patterns in biological imaging. *PLoS One* **9:** e87759.

Burguet J, Andrey P, Rampin O, Maurin Y. 2009. Three-dimensional statistical modeling of neuronal populations: illustration with spatial localization of supernumerary neurons in the locus coeruleus of quaking mutant mice. *J Comp Neurol* **513:** 483–495.

Burguet J, Maurin Y, Andrey P. 2011. A method for modeling and visualizing the three-dimensional organization of neuron populations from replicated data: properties, implementation and illustration. *Pattern Recognit Lett* **32:** 1894–1901.

Corradin O, Saiakhova A, Akhtar-Zaidi B, Myeroff L, Willis J, Cowper-Sal·lari R, Lupien M, Markowitz S, Scacheri PC. 2014. Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. *Genome Res* **24:** 1–13.

Dekker J, Belmont AS, Guttman M, Leshyk VO, Lis JT, Lomvardas S, Mirny LA, O'Shea CC, Park PJ, Ren B, et al. 2017. The 4D nucleome project. *Nature* **549:** 219–226.

Dixon JR, Jung I, Selvaraj S, Shen Y, Antosiewicz-Bourget JE, Lee AY, Ye Z, Kim A, Rajagopal N, Xie W, et al. 2015. Chromatin architecture reorganization during stem cell differentiation. *Nature* **518:** 331–336.

Du Z, Zheng H, Huang B, Ma R, Wu J, Zhang X, He J, Xiang Y, Wang Q, Li Y, et al. 2017. Allelic reprogramming of 3D chromatin architecture during early mammalian development. *Nature* **547:** 232–235.

Durand NC, Robinson JT, Shamim MS, Machol I, Mesirov JP, Lander ES, Aiden EL. 2016. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Syst* **3:** 99–101.

The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489:** 57–74.

Ernst J, Kellis M. 2012. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods* **9:** 215–216.

Fisher RA. 1925. *Statistical methods for research workers*. Oliver and Boyd, Edinburgh.

Forcato M, Nicoletti C, Pal K, Livi CM, Ferrari F, Bicciato S. 2017. Comparison of computational methods for Hi-C data analysis. *Nat Methods* **14:** 679–685.

Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK. 2010. Simple combinations of lineage-determining transcription factors prime *cis*-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38:** 576–589.

Hu M, Deng K, Selvaraj S, Qin Z, Ren B, Liu JS. 2012. HiCNorm: removing biases in Hi-C data via Poisson regression. *Bioinformatics* **28:** 3131–3133.

Imakaev M, Fudenberg G, McCord RP, Naumova N, Goloborodko A, Lajoie BR, Dekker J, Mirny LA. 2012. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat Methods* **9:** 999–1003.

Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, et al. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326:** 289–293.

Lun ATL, Smyth GK. 2015. diffHic: a Bioconductor package to detect differential genomic interactions in Hi-C data. *BMC Bioinformatics* **16:** 258.

Luo W, Friedman MS, Shedden K, Hankenson KD, Woolf PJ. 2009. GAGE: generally applicable gene set enrichment for pathway analysis. *BMC Bioinformatics* **10:** 161.

Paulsen J, Sandve GK, Gundersen S, Lien TG, Trengereid K, Hovig E. 2014. HiBrowse: multi-purpose statistical analysis of genome-wide chromatin 3D organization. *Bioinformatics* **30:** 1620–1622.

Pebesma EJ. 2004. Multivariable geostatistics in S: the gstat package. *Comput Geosci* **30:** 683–691.

R Core Team. 2016. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Rao SSPSP, Huntley MHH, Durand NCC, Stamenova EKK, Bochkov IDD, Robinson JTT, Sanborn ALL, Machol I, Omer ADD, Lander ESS, et al. 2014. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159:** 1665–1680.

Riley JW, Stouffer SA, Suchman EA, Devinney LC, Star SA, Williams RM. 1949. The American soldier: adjustment during army life. *Am Sociol Rev* **14:** 557.

Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26:** 139–140.

Sanyal A, Lajoie BR, Jain G, Dekker J. 2012. The long-range interaction landscape of gene promoters. *Nature* **489:** 109–113.

Sauria MEG, Phillips-Cremins JE, Corces VG, Taylor J. 2015. HiFive: a tool suite for easy and efficient HiC and 5C data analysis. *Genome Biol* **16:** 237.

Schmitt AD, Hu M, Jung I, Xu Z, Qiu Y, Tan CL, Li Y, Lin S, Lin Y, Barr CL, et al. 2016. A compendium of chromatin contact maps reveals spatially active regions in the human genome. *Cell Rep* **17:** 2042–2059.

Servant N, Varoquaux N, Lajoie BR, Viara E, Chen CJ, Vert JP, Heard E, Dekker J, Barillot E. 2015. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol* **16:** 259.

Smith EM, Lajoie BR, Jain G, Dekker J. 2016. Invariant TAD boundaries constrain cell-type-specific looping interactions between promoters and distal elements around the *CFTR* locus. *Am J Hum Genet* **98:** 185–201.

Song C, Tseng GC. 2014. Hypothesis setting and order statistic for robust genomic meta-analysis. *Ann Appl Stat* **8:** 777–800.

Taberlay PC, Achinger-Kawecka J, Lun ATL, Buske FA, Sabir K, Gould CM, Zotenko E, Bert SA, Giles KA, Bauer DC, et al. 2016. Three-dimensional

disorganization of the cancer genome occurs coincident with long-range genetic and epigenetic alterations. *Genome Res* **26:** 719–731.

Tang Z, Luo OJ, Li X, Zheng M, Zhu JJ, Szalaj P, Trzaskoma P, Magalska A, Wlodarczyk J, Ruszczycki B, et al. 2015. CTCF-mediated human 3D genome architecture reveals chromatin topology for transcription. *Cell* **163:** 1611–1627.

Ulianov SV, Khrameeva EE, Gavrilov AA, Flyamer IM, Kos P, Mikhaleva EA, Penin AA, Logacheva MD, Imakaev MV, Chertovich A, et al. 2016. Active chromatin and transcription play a key role in chromosome partitioning into topologically associating domains. *Genome Res* **26:** 70–84.

Vernimmen D, De Gobbi M, Sloane-Stanley JA, Wood WG, Higgs DR. 2007. Long-range chromosomal interactions regulate the timing of the transition between poised and active gene expression. *EMBO J* **26:** 2041–2051.

Wang J, Lan X, Hsu PY, Hsu HK, Huang K, Parvin J, Huang TH, Jin VX. 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.

Wilkinson B. 1951. A statistical consideration in psychological research. *Psychol Bull* **48:** 156–158.

Received July 4, 2016; accepted in revised form January 8, 2018.