Liang Niu and Shili Lin*

A Bayesian mixture model for chromatin interaction data

Abstract: Chromatin interactions mediated by a particular protein are of interest for studying gene regulation, especially the regulation of genes that are associated with, or known to be causative of, a disease. A recent molecular technique, Chromatin interaction analysis by paired-end tag sequencing (ChIA-PET), that uses chromatin immunoprecipitation (ChIP) and high throughput paired-end sequencing, is able to detect such chromatin interactions genomewide. However, ChIA-PET may generate noise (i.e., pairings of DNA fragments by random chance) in addition to true signal (i.e., pairings of DNA fragments by interactions). In this paper, we propose MC_DIST based on a mixture modeling framework to identify true chromatin interactions from ChIA-PET count data (counts of DNA fragment pairs). The model is cast into a Bayesian framework to take into account the dependency among the data and the available information on protein binding sites and gene promoters to reduce false positives. A simulation study showed that MC_DIST outperforms the previously proposed hypergeometric model in terms of both power and type I error rate. A real data study showed that MC_DIST may identify potential chromatin interactions between protein binding sites and gene promoters that may be missed by the hypergeometric model. An R package implementing the MC_DIST model is available at <http://www.stat.osu.edu/~statgen/SOFTWARE/MDM>.

Keywords: Bayesian mixture model; ChIA-PET; R package.

DOI 10.1515/sagmb-2014-0029

1 Introduction

Long-range chromatin interactions mediated by a particular protein are of great interest for the study of gene regulation. This is particularly relevant in cancer research, as scientists have discovered that a number of cancer gene regulating transcription factors have binding sites that are located far away from the target gene promoter regions in certain cancers. Examples of such transcription factor binding sites include estrogen receptor (ER) binding sites in breast cancer (Carroll et al., 2006) and androgen receptor (AR) binding sites in prostate cancer (Jia et al., 2008; Wang et al., 2009; Yu et al., 2010). Studies of such long-range chromatin interactions have been aided by the molecular technique chromosome conformation capture (3C) (Dekker et al., 2002) and many other developments derived from 3C. As a principle, 3C has been demonstrated for its ability of detecting long-range chromatin interactions (Tolhuis et al., 2002; Murrell et al., 2004; Spilianakis and Flavell, 2004; Vernimmen et al., 2007). However, 3C is a low throughput technique since it only studies chromatin interaction between two pre-selected DNA loci. As such, many further developments to improve the capability of such a technique have flourished. They include circularized 3C (4C) (Simonis et al., 2006; Zhao et al., 2006), carbon-copy 3C (5C) (Dostie et al., 2006), ChIP-loop (Horike et al., 2005), and Hi-C (Lieberman-Aiden et al., 2009). These new techniques address the limitations of 3C, with Hi-C being the technique that is capable of capturing genome-wide chromatin interactions.

*Corresponding author: Shili Lin, Department of Statistics, Ohio State University, Columbus, OH 43210, USA, e-mail: shili@stat.osu.edu

Liang Niu: Department of Environmental Health, School of Medicine, University of Cincinnati, Cincinnati, OH 45267, USA

Chromatin interaction analysis by paired-end tag sequencing (ChIA-PET) is another molecular technique for detecting genome-wide chromatin interactions (both long-range ones and short-range ones), but with restriction to those that are mediated by a particular protein. Here is a schematic description of ChIA-PET, full details can be found in Fullwood et al. (2009). First, the chromatin is cross-linked with formaldehyde and sonication is used to break up the chromatin; second, specific antibody of choice is used to precipitate the chromatin fragments bound by the protein of interest (chromatin immunoprecipitation, or ChIP) and the biotinylated oligonucleotide half-linkers containing flanking MmeI (a restriction enzyme) sites are used to ligate the tethered DNA fragments; third, the linkers are ligated and MmeI is introduced to digest the ligated fragments; finally, the ligation products are purified using streptavidin-coated beads and the pair end tags are sequenced through high throughput paired-end sequencing. Note that the ChIP step is for the purpose of focusing on only interactions involving a particular protein of interest, which sets it apart from the Hi-C protocol. A full comparison between ChIA-PET and Hi-C can be found in de Wit and de Laat (2012).

A ChIA-PET experiment generates millions of paired-end sequencing reads that can be mapped uniquely to the reference genome (Fullwood et al., 2009). Pairs of DNA regions, which we will refer to as fragment pairs, can then be determined based on the overlapping of the read alignments. For each such fragment pair, the count of alignment pairs that fall into the fragment pair can be obtained. Such counts, together with their associated fragment pairs, are referred to as the *ChIA-PET count data* in this paper. However, not all fragment pairs in the count data are due to true biological interactions: some of them are present because of “random collision” of two sonicated DNA fragments during ligation due to their close proximity. Such pairings are termed *false pairs* to distinguish them from the *true pairs*, those indeed the results from chromatin interactions. In general, the larger the count is, the more likely the fragment pair is a true pair. However, false pairs may also have large counts as revealed by simulation. Therefore, modeling the observed frequencies of the fragment pairs from such experiments to distinguish true pairs from false pairs is an interesting and important problem. Few methods are available for performing this task to date. The hypergeometric test employed by Fullwood et al. (2009) is an example, however the underlying assumption for such a test to be valid appears to be oversimplified. Therefore, the hypergeometric test can sometimes perform poorly, as we can see from the simulation study.

In this paper, we propose MC_DIST based on a mixture modeling framework for ChIA-PET count data to distinguish true pairs from false ones. We compared its performance with two other models, including a model (denoted as *NIP*) proposed in the paper for comparison purpose and the hypergeometric model (denoted as *HG* in this paper), through an extensive simulation study. We also applied MC_DIST on a real ChIA-PET data set from Fullwood et al. (2009) and identified potential chromatin interactions between ER binding sites and ER regulated genes that are missed by the hypergeometric model. An R package MDM (Marginal count Distance Model) implementing the MC_DIST model is available at <http://www.stat.osu.edu/~statgen/SOFTWARE/MDM>.

2 Model

Let $\mathbf{X}=\{X_i, i=1, 2, \dots, n\}$ be a set of random variables where $X_i \geq k$ represents the count of fragment pair i , i.e., the number of read pairs uniquely mapped to the i -th pair of fragments in a ChIA-PET experiment, n is the total number of fragment pairs, and k is the cut-off value that one may use to decide whether a pair should be kept in the analysis. Usually, $k=1$, that is, one keeps all the pairs. In Fullwood et al. (2009), k was set to be 2, as a pair that has been observed once was assumed by the authors to be a false pair and should be excluded from the analysis. This threshold can be set by investigators based on their knowledge of the experiments.

We assume that X_i follows a two-component mixture distribution:

$$X_i \sim \sum_{j=0}^1 w_{ji} f(\cdot | \lambda_{ji}), \text{ independently for } i=1, 2, \dots, n,$$

where $w_{oi} + w_{ii} = 1$, and $\lambda_{oi} < \lambda_{ii}$ for $1 \leq i \leq n$ is also assumed for identifiability. The component probability mass function $f(\cdot|\lambda)$ is the $(k-1)$ -truncated Poisson distribution:

$$f(X=x|\lambda) = \frac{\lambda^x}{x!(e^\lambda - (1 + \lambda + \frac{\lambda^2}{2!} + \dots + \frac{\lambda^{k-1}}{(k-1)!))},$$

$$\text{for } x = k, k+1, \dots$$

The use of a $(k-1)$ -truncated Poisson distribution is motivated by the nature that X_i is a count and that $X_i \geq k$. We further note that Poisson has been used to model next-generation sequencing (NGS) count data in the literature (Marioni et al., 2008; Jiang and Wong, 2009; Srivastava and Chen, 2010), although negative binomial is also a popular one (Anders and Huber, 2010; Robinson et al., 2010).

For pair i ($1 \leq i \leq n$), the first component $f(\cdot|\lambda_{oi})$ models the count conditional on it being a false pair, whereas the second component $f(\cdot|\lambda_{ii})$ models the count conditional on it being a true pair. The requirement that $\lambda_{oi} < \lambda_{ii}$ comes from a biologically supported assumption (Rousseau et al., 2011) that we expect the interaction intensity to be higher for a true pair.

It is easy to see that the counts of two pairs with a common fragment are correlated with each other and therefore the data are dependent. We also noted that the extent to which the locations of a fragment pair relative to a pair of known DNA loci of interest is useful for our classification purpose. In light of these two observations, we cast the above model into a Bayesian framework and incorporate the data dependency and proximity of genomic positions of pairs to known DNA loci into the modeling of the prior distributions of w_{oi} and w_{ii} . The details of the Bayesian framework are as follows.

We assume that

$$\lambda_{oi} \sim \Gamma(b, \frac{b}{r_0}), \text{ and}$$

$$\lambda_{ii} | \lambda_{oi} \sim N(\lambda_{oi}, \sigma^2) \cdot I(\lambda_{ii} > \lambda_{oi}),$$

independently for $i=1, 2, \dots, n$, where b is the shape parameter of the gamma model and $\frac{b}{r_0}$ is the rate parameter [thus the mean of $\Gamma(b, \frac{b}{r_0})$ is $\frac{b}{b/r_0} = r_0$]. For the hyperparameters b , r_0 and σ , we assume that, independently,

$$b \sim U(0, A),$$

$$r_0 \sim U(0, B),$$

$$\text{and } \sigma \sim U(0, C),$$

where A , B and C are large constants, leading to uninformative distributions. In the simulation study and real data study, we chose $A=B=C=1000$.

To take the data dependency and the genomic locations of the pairs relative to known DNA loci into account, we define the following two terms for each pair and use them in the specification of the priors of the w 's. For pair i ($1 \leq i \leq n$), mc_i , or *marginal count*, is defined as the sum of two *marginal counts* of the two fragments of the pair, where the marginal count of a fragment is defined as the sum of the counts of all pairs containing that fragment. On the other hand, $dist_i$, or *distance*, is defined as the minimum of the following two values: the sum of the distance between the midpoint of $frag_{ii}$ and the nearest midpoint of a transcription factor binding site (TFBS) of the protein P and the distance between the midpoint of $frag_{zi}$ and the nearest gene transcription start site (TSS); and the sum of the distance between the midpoint of $frag_{ii}$ and the nearest gene TSS and the distance between the midpoint of $frag_{zi}$ and the nearest midpoint of TFBS of the protein P . Here $frag_{ii}$ and $frag_{zi}$ are the two fragments of the pair i , and P is the protein of interest in the ChIA-PET experiment. Here, the marginal counts of the pairs reflect the dependency among the data and the distance of each pair measures the extent to which the pair is related to a pair of DNA loci of interest.

With the above two quantities, we model w_{ii} by a beta distribution as follows:

$$w_{ii} \sim \beta(\text{rmcd}_i \cdot d, \text{rmcd} \cdot d) \text{ independently for } i=1, 2, \dots, n,$$

where $\text{rmcd}_i = \frac{mc_i}{\text{dist}_i}$, $\text{rmcd} = \frac{1}{n} \sum_{i=1}^n \text{rmcd}_i$ and d is a hyperparameter. The rmcd_i stands for the “ratio of marginal count to distance of pair i ”. The motivations for specifying such a prior distribution are that the larger mc_i is, the more active the two fragments of pair i are, thus pair i is more likely to be a true pair; and that the smaller dist_i is, the more likely the pair is observed because of the interaction between the nearby TFBS and the nearby gene TSS. Since w_{ii} represents the probability of pair i being a true pair, it is naturally to include the rmcd_i 's into the prior distribution of w_{ii} as described above. It is also worth to re-emphasize that, although the ChIA-PET protocol can identify the TFBSs of the protein of interest and the chromatin interactions with their regulated genes, the existence of random collisions may include pairs that do not have the correct genomic features. The introduction of the additional information (external to ChIA-PET) is precisely to address this issue and can lead to more accurate identification of random collisions without sacrificing the power of detecting true interactions. We also assume that d follows a uniform distribution, i.e., $d \sim U(0, D)$, where D is a large constant and should be large enough to balance the difference between mc and dist , since they are in different units. In the simulation study and real data study, we chose $D=10,000$. Note that we do not need to specify the prior distribution of w_{oi} since $w_{oi}=1-w_{ii}$.

To determine whether pair i truly represents interacting chromatins, we define a latent variable Z_i for $i=1, \dots, n$, such that

$$Z_i \sim \text{Bernoulli}(w_{ii})$$

and

$$X_i | Z_i = j \sim f(\cdot | \lambda_{ji}), \text{ for } j=0, 1,$$

independently for $i=1, 2, \dots, n$. The Z_i is an indicator variable in which $Z_i=1(0)$ implies that the pair i is a true (false) pair. We conclude that pair i is a true pair whenever $P(Z_i=1|\mathbf{X})$ is bigger than a cutoff, say 0.5. The posterior probabilities are calculated using samples obtained by a Markov Chain Monte Carlo (MCMC) method. Specifically, in each MCMC iteration Z_i , w_{ii} and σ^2 are sampled from their full conditional distributions (Bernoulli distribution for Z_i , beta distribution for w_{ii} , and inverse gamma distribution for σ^2). On the other hand, since the full conditional distributions of the other parameters are not of a known form, we utilize the Metropolis-Hastings algorithm using either a log-normal distribution (for λ_{oi} , λ_{ii} , r_o or b) or a uniform distribution (for d) as the proposal distribution.

We refer to the above model as *Marginal Count–Distance Model* and denote it by *MC_DIST*. A graphical representation of MC_DIST is shown in Figure 1. We also proposed another model, *non-informative prior model (NIP)*, which does not take locations and dependency into account. It may be considered as a special case of MC_DIST. Specifically, in NIP we assume that w_o and w_i (weights of the mixture components) are common for all pairs and

$$w_i \sim U(0, 1).$$

The inclusion of the NIP is to show, through a simulation study, the importance of including such information in the MC_DIST model.

3 Inference

3.1 Simulation study

To evaluate the performance of MC_DIST and compare it with NIP and HG, we carried out a simulation study to evaluate the type I error rate, power and the effect of interaction intensity.

for $i=1, \dots, n$, where n is the number of fragment pairs.

Figure 1 A graphical representation of MC_DIST. The circles represent model parameters and the rectangles represent data or prior information.

3.1.1 Data generation

To generate the count data, we imitated a ChIA-PET experiment aiming to study genome-wide chromatin interactions bound by estrogen receptor alpha ($ER\alpha$) in breast cancer cell line MCF7.

First, we determined 6348 pairs of DNA loci and designated each pair either as a pair with interaction or a pair without interaction. The DNA loci used to form those 6348 pairs included $23 \times 2 = 46$ randomly selected $ER\alpha$ binding sites in MCF7 cell line (two on each human chromosome), $23 \times 2 = 46$ randomly selected gene transcription start sites (two on each human chromosome), and $23 \times 4 = 92$ randomly selected non-specific loci (not a $ER\alpha$ binding site or a gene transcription start site, four on each human chromosome). The $ER\alpha$ binding site information is obtained from database ChIPBase (Yang et al., 2013) and the gene transcription start site information is from the RefSeq gene model for human genome GRCh37/hg19 (Pruitt et al., 2014). After we determined the DNA loci, we paired each selected $ER\alpha$ binding site with each selected transcription start site to get a set of $46 \times 46 = 2116$ pairs. We also paired each selected $ER\alpha$ binding site with each selected non-specific loci to get another set of $46 \times 92 = 4232$ pairs. The former set of pairs were designated as pairs with interaction (“true pairs”) and were assigned higher sampling probabilities in the later sampling process, and the latter set of pairs were designated as pairs without interaction (“false pairs”) and were assigned lower sampling probabilities in the later sampling process. To facilitate the following discussion, we denoted the 46 $ER\alpha$ binding sites by $TFBS_{ij}$ and the 46 transcription start sites by TSS_{ij} , where $1 \leq i \leq 23$ is the chromosome index and $1 \leq j \leq 2$ is the locus index.

Next, we sampled 100,000 pairs of DNA loci from the above $2116 + 4232 = 6348$ pairs (with replacement). To study the effect of interaction intensity on the performance of MC_DIST, we divided the set of 2116 pairs of DNA loci with interaction into two groups (group 1 and group 2) of equal size and assigned the sampling probabilities in such a way that the sampling probability of a pair in group 1 is twice of that of a pair in group 2. The group 1 consisted of those 1058 pairs whose two DNA loci shared the same locus index, i.e., a $TFBS_{ij}$ and a TSS_{kj} for some $1 \leq i, k \leq 23$ (i and k can be equal) and some $1 \leq j \leq 2$. The group 2 consisted of the remaining 1058 pairs. For each pair in the set of the 4232 pairs that we designated as pairs without interaction (referred to as group 3), we assigned the sampling probability to be one third of the sampling probability of a pair in

group 2. Thus for each pair in the above three groups, the sampling probabilities are 0.000436, 0.000218 and 0.000072, respectively.

For each sampled pair, we cut along the related chromosome(s) according to a Poisson process so that the expected fragment length be 250 bp (imitating a sonication process) and selected the two DNA fragments containing the two DNA loci of the sampled pair; we then randomly selected two ends from the total four ends of the two fragments and ligated the two ends with probability 0.8 (imitating a ligation process). We sequenced each ligated product to get a pair of 50 bp reads centered at, and with the read directions toward, the ligation point (imitating a restriction enzyme digestion process followed by sequencing). Then we aligned such read pairs to the reference human genome GRCh37/hg19 and discarded those read pairs with at least one read that could not be aligned uniquely to the genome, also we discarded those read pairs of which the two reads were aligned too close to each other (the criterion of “too close” are explained below in more details), as those read pairs were likely to be from self-ligated products (formed by the ligation between the two ends of the same sonicated DNA fragment). Next we extended each alignment of the remaining alignment pairs to 250 bp along the direction opposite to the read direction [imitating the read extension step in ChIA-PET analysis (Li et al., 2010)]. We then define disjoint chromatin regions by overlapping extended alignments [imitating the step of identifying the ChIP-enriched interaction anchor regions in ChIA-PET analysis (Li et al., 2010)]. Finally, for each pair of the obtained chromatin regions (fragment pair), we counted the number of alignment pairs that fell into it, i.e., those alignment pairs with two alignments overlapped with the two chromatin regions respectively, to obtain a data set of counts of the fragment pairs. Note that in the obtained ChIA-PET count data set we excluded those fragment pairs with count 0.

To determine whether the two reads of a read pair are too close to each other so that to be discarded in the above simulation, we calculated the logarithm of the genomic distance between the two aligned reads for each aligned read pair, and then performed a k -means clustering on all those values to partition the read pairs into two groups. Those read pairs in the group with smaller genomic distances were then assumed to be those with the two reads too close to each other. The above clustering method is also used in the ChIA-PET protocol (Li et al., 2010).

3.1.2 Data analysis and results

The simulated data is summarized in Figure 3, which shows count versus the logarithm of the ratio of marginal count to distance [$\log(rmcd)$ ’s] for the fragment pairs. The dark green and red points represent true pairs, while the dark red and gold points represent false pairs. We see from the figure that, in general, true pairs have both large counts and large $rmcd$ s, while false pairs have both small counts and small $rmcd$ ’s. However, there exist pairs that have large $rmcd$ s but small counts (red points), or large count but small $rmcd$ ’s (dark red points to the right of the vertical line). These are pairs that are difficult to classify correctly, and it is therefore of particular interest to see how the algorithms perform on these points.

We applied three models (MC_DIST, NIP and HG) to the simulated data, which contains 6110 fragment pairs whose counts range from 1 to 46. For each MCMC algorithm used for MC_DIST or NIP, we set the total number of iterations to be 2,000,000 and set the burn-in period to be 400,000 to ensure its convergence, which was also confirmed by the trace plots of posterior samples. Furthermore, we used Raftery and Lewis diagnostic (as in Raftery and Lewis, 1995) to ensure that the total number of iterations and the burn-in were large enough; we also used the Gelman and Rubin’s convergence diagnostic (as in Gelman and Rubin, 1992) to further check for convergence of the MCMC algorithms. We used a machine with a Intel Xeon processor (E5-2643 at 3.30 GHz) and the computing time is 4.5 h for NIP model and 9 h for the MC_DIST model. For more details on the MCMC algorithms, such as the tuning parameters used in the proposal distribution of the Metropolis Hastings algorithms, the achieved acceptance rates, convergence, please see the Supplementary materials (Sections S1 and S2). Provided there is also an analysis of the posterior distribution (Section S3).

The results for the three models (MC_DIST, NIP and HG) in the simulation study are summarized in Table 1. Note that in this table, the “type I error” refers to the false positive rate, i.e., the proportion of fragment pairs

Table 1 Results of three analysis methods on the simulated data.

						Actual
	–			+		
	–	+	Type I error	–	+	Power
HG ^a	3618	217	0.057	1074+250^b	30+921	0.418
NIP	3618	217	0.057	152+139	952+1032	0.872
MC_DIST	3834	1	0.0003	144+138	960+1033	0.876

^aThe HG classification is based on a significance level (for each pair) such that the type I error rate of HG is the same as that of NIP. ^bBold numbers represent pairs with weak interactions and italic numbers represent pairs with strong interactions.

that are classified as true pairs by a particular method among all false pairs; and the “power” refers to the true positive rate, i.e., the proportion of fragment pairs that are classified as true pairs by a particular method among all true pairs. In this table, the result of HG model was obtained by setting the cut-off of p -values in such a way that the classification result had the same type I error rate (0.057) as the NIP classification result. Note that setting the cut-off for the HG model in such a way is to facilitate direct comparison of power between HG and NIP as they now have the same empirical type I error rate. The HG model performed the worst: it had the smallest power (0.418). The NIP model had a much higher power (0.872) than the HG model. On the other hand, the MC_DIST model performed the best. It had the smallest type I error rate (0.0003, only one type I error) and the highest power (0.876). By looking at those bold (true fragment pairs of which the two DNA loci were assigned lower sampling probabilities) and italic (true fragment pairs of which the two DNA loci were assigned higher sampling probabilities) entries in Table 1, we found that pairs with strong interaction are easier to be classified correctly than those with weak interaction for the three models (especially for the HG model), as one would expect. By varying the cut-off value for the posterior probabilities of NIP, MC_DIST and that for the p -values obtained by HG, we obtained three ROC curves shown in Figure 2. Again we see that the order of the models, from the best to the worst, is MC_DIST, NIP and HG.

We dissect the advantage of MC_DIST over NIP as an illustration of the importance of taking data dependency and the genomic locations of the pairs relative to known DNA loci into account. In particular, we compared the results of MC_DIST with NIP by plotting the count versus the logarithm of the ratio of marginal count to distance for the fragment pairs (Figure 3). The coloring scheme is determined by results from MC_DIST: dark green points represent the *true positives*; gold points represent the *false positives*; dark red points represent the *true negatives* and red points represent the *false negatives*. The blue vertical line at 7.5 depicts the classification scheme of NIP, that is, pairs with count larger than seven are all classified as true pairs. We see that many false positives by NIP were corrected by MC_DIST (dark red points to the right of the blue line). The reason is that these pairs have small ratios of marginal count to distance, although their count are relatively large (from 8 to 14). This result demonstrates the flexibility of MC_DIST: pairs classified as “truth pairs” do not necessarily have larger counts than those classified as “false pairs”; the additional information from marginal counts and genomic features contribute to the classification rule as well.

To investigate how reproducible the results from MC_DIST are across multiple simulations, we replicated the same simulation 100 times to obtain 100 simulated data sets. MC_DIST was then applied to analyze each of them. The type I error rate across these 100 data sets ranges from 0 to 0.0079, while the power is between 0.852 and 0.938. These results show that MC_DIST performs consistently. See Section S4 and Figure S6 in the Supplementary materials for further details.

3.2 Real data analysis

We also evaluated MC_DIST on a real ChIA-PET data set. The data set is the IHM001F data set from Fullwood et al. (2009), which was created from a ChIA-PET experiment to study the chromatin interactions mediated

Figure 2 Comparison of ROC curves. Different colors represent different models. The black dashed line on the diagonal represents random classification.

Figure 3 Comparison of MC_DIST with NIP to demonstrate the need for incorporating data dependency and the relevance of the pairs to known DNA loci into the modeling. The x-axis is the count and the y-axis is the logarithm of the ratio of marginal count to distance. The color of each point indicates the classification given by the MC_DIST model: dark green points are true positives; gold points are false positives; dark red points are true negatives; red points are false negatives. The blue vertical line at 7.5 gives the classification rule of NIP, i.e., all points to the right of it are positives and all points to the left of it are negatives.

by estrogen receptor alpha ($ER\alpha$) in MCF-7 cell line. It contains 2289 pairs of DNA fragments whose counts range from 2 to 83. In order to calculate the distance for each fragment pair, we used the TSS information in the UCSC human gene annotation (hg18) (Karolchik et al., 2014) and the $ER\alpha$ binding site information for the MCF7 cell line from Fullwood et al. (2009).

We applied MC_DIST to the real data set and compared the classification result with that of the HG model. The classification of the HG model was based on the q -values [p -values after false discovery rate (FDR) adjustment] and the FDR is controlled at 0.05, i.e., all pairs with q -values < 0.05 was classified as true pairs, otherwise

were classified as false pairs. The result is illustrated in Figure 4, which shows the logarithm of the count versus the logarithm of the ratio of marginal count to distance for the fragment pairs. The coloring scheme is determined by the classification results from MC_DIST and HG: dark green points represent the pairs that were classified as true pairs by both models (+/+); gold points represent pairs that were classified as true pairs by MC_DIST but as false pairs by HG (+/-); dark red points represent pairs that were classified as false pairs by both models (-/-); red points represent pairs that were classified as false pairs by MC_DIST but as true pairs by HG (-/+). We see that all but nine pairs that were classified as true pairs by MC_DIST were also classified as true pairs by HG. However, many pairs that were classified as false pairs by MC_DIST were classified as true pairs by HG (1641 out of 1857). The above observation is consistent with the property of MC_DIST model: that is, making using of the data dependency and the proximity of known DNA loci may reduce the false positive rate.

Interestingly, for the nine fragment pairs that were classified as true pairs by MC_DIST but were classified as false pairs by HG, they have high q -values (all 1 except a 0.43) and also high posterior probabilities (0.6–0.8). They are summarized in Table 2. From the literature, we see that, among them, there are two genes (*CYP24A1*, *PARAD6B*) that are $ER\alpha$ regulated (Labhart et al., 2005; Parisi et al., 2009) and one (*SNX24*) that is directly regulated by estrogen (Wright et al., 2009). It turns out that all three interactions involve the chromosomal region chr17:57151973-57159980, which in fact harbors three $ER\alpha$ binding sites. These results showed that MC_DIST may be able to identify chromatin interactions that are missed by the hypergeometric model. However, we do realize that the “ground truth” is unknown, and therefore, the results need to be substantiated by other external data. Nevertheless, pairs identified by MC_DIST can be helpful in aiding the designs of further investigation and/or experimental validations.

4 Discussion

In this paper we proposed the MC_DIST model for distinguishing true pairs of DNA fragments brought to close proximity by chromatin interactions from false pairs (due to random collisions) based on ChIA-PET count

Figure 4 Comparison of MC_DIST and HG on real data. The x-axis is the logarithm of the count and the y-axis is the logarithm of the ratio of marginal count to distance. The color of each point indicates the classifications given by the MC_DIST model and the HG model: dark green points represent pairs that were classified as true pairs by both models (+/+); gold points represent pairs that were classified as true pairs by MC_DIST but as false pairs by HG (+/-); dark red points represent pairs that were classified as false pairs by both models (-/-); red points represent pairs that were classified as false pairs by MC_DIST but as true pairs by HG (-/+). The false discovery rate (FDR) of HG classification was controlled at 0.05.

Table 2 Nine chromatin interactions identified by MC_DIST but missed by the hypergeometric model.

Fragment 1	Fragment 2	Count	Proximity ^a	Nearby gene	HG ^b	MC_DIST ^c
chr17:57151973-57159980	chr20:52170453-52175517	5	BP	<i>CYP24A1</i>	1.00	0.59
chr17:57151973-57159980	chr20:48771221-48781451	3	BP	<i>PARD6B</i>	1.00	0.60
chr17:57151973-57159980	chr5:122210789-122214348	3	BP	<i>SNX24</i>	1.00	0.67
chr1:144436649-144443486	chr11:129066696-129073013	2	BP	<i>AX746800</i>	1.00	0.71
chr12:96483832-96488523	chr16:82536611-82542231	2	BP	<i>OSGIN1</i>	1.00	0.60
chr14:102661740-102666179	chr20:46913393-46917450	2	PB	<i>TNFAIP2</i>	1.00	0.60
chr17:56787225-56790580	chr20:48842790-48846927	2	BP	<i>BCAS4</i>	1.00	0.68
chr17:57151973-57159980	chr3:195833301-195835602	2	BP	<i>TMEM44</i>	1.00	0.73
chr17:46374734-46378746	chr9:122677689-122680279	2	BP	<i>PHF19</i>	0.43	0.81

^a“PB” or “BP.” “PB” indicates that fragment 1 is close to a promoter and fragment 2 is close to a ER α binding site, otherwise it is “BP.”

^bq-Value from the hypergeometric model.

^cPosterior probability of being a true pair from MC_DIST.

data. We utilized a mixture modeling framework and cast the problem into a Bayesian setting to incorporate data dependency and the relative proximity of fragment pairs to DNA loci of special interest in the ChIA-PET study. With this feature, MC_DIST outperforms NIP and the hypergeometric model in Fullwood et al. (2009), as shown in the simulation study. We also evaluated MC_DIST on a real ChIA-PET data and showed that MC_DIST can identify potential chromatin interactions that can be missed by the hypergeometric model. An R package that implemented the MC_DIST method is available from the website provided.

To confirm that the superiority of MC_DIST over NIP is due to MC_DIST incorporating useful information (marginal counts and distances) rather than simply due to MC_DIST being more flexible in its modeling than NIP (i.e., allowing for pair specific weights w_{ii} 's), we applied a variant of NIP to the simulated data and compared its performance with NIP. Specifically, in the NIP variant we make the mixture weight to be pair specific (i.e., w_{ii} for pair i) and assume that $w_{ii} \sim \text{beta}(\alpha, \beta)$, where $\alpha \sim U(0, 1000)$ and $\beta \sim U(0, 1000)$. The classification result of the NIP variant turned out to be identical to that of NIP. Moreover, for each pair, the posterior probability of it being a true pair under the NIP variant was almost identical to its counterpart under the NIP model (the difference was between 0 and 0.013 for all pairs). This study showed that the MC_DIST outperforms NIP because the extra information plays a supplementary, but important, role in the classification.

In addition to the simulation study presented in Subsection 3.1, we also performed several other simulation studies to investigate the sensitivity and stability of MC_DIST. First, we explored an alternative prior distribution of λ_{ii} given λ_{oi} to study the sensitivity of the prior specification. The alternative prior distribution is $N(\lambda_{oi} + m, \sigma^2) \cdot I(\lambda_{ii} > \lambda_{oi})$ with $m > 0$ pre-selected, which may be considered as more sensible because of the constraint that $\lambda_{ii} > \lambda_{oi}$. Our exploration shows that such MC_DIST variant does not outperform MC_DIST. Specifically, we ran MC_DIST on the simulated data set and calculated the 10th, 20th, 30th and 40th percentiles of the posterior samples of $\{\lambda_{ii} - \lambda_{oi} | 1 \leq i \leq n\}$. We then set m to be each of those percentiles and ran the MC_DIST variant with the alternative prior. The type I error rates of those MC_DIST variants are between 0 to 0.0003, and the powers are between 0.871 to 0.882 (see the Supplementary materials Section S5 and Table S7 for details). We have also considered m as a hyperparameter that follows a uniform distribution and modified the above MC_DIST variant correspondingly, but the resulted model had smaller type I error rate (0) but also smaller power (0.847) than MC_DIST. These results suggest that MC_DIST is quite robust.

We also considered the effect of our definition of marginal counts (mc) on the MC_DIST classification as the information contained therein seems to be extremely important for separating true pairs from the false ones. Our definition of mc as in Section 2 double counts the connection between the pairs under consideration, and therefore we modified the definition by eliminating the double counting. Using the same simulated data but with the new mc definition, we achieved identical results. This exercise reinforce our prior notation that mc adds an abundance of information but its specific definition does not affect the outcome as long as it captures the essence of dependency.

Our final additional simulation study is to investigate the stability of MC_DIST under different simulation settings. We first consider the effect of the sampling probabilities on the performance of MC_DIST. To this end, we reset the ratio of the sampling probabilities of each pair in the three groups mentioned in the Subsection 3.1 from the original 6:3:1 to 8:4:1 and 10:5:1 and simulated two new ChIA-PET count data sets. We ran MC_DIST on the two new data sets and found that, as the ratio becomes larger, the type I error rate of MC_DIST decreases and the power of MC_DIST increases. Specifically, for the data set with the sampling ratio of 8:4:1 and the data set with the sampling ratio of 10:5:1, the type I error rates are 0 and 0 and the powers are 0.885 and 0.904, respectively. We then consider the effect of the composition of true pairs and false pairs used for the simulation on MC_DIST. To this end, we increased the number of pairs of DNA loci in group 3, i.e., the group of pairs that were designated as pairs without interaction, from the original 4232 to 6348, and simulated another three ChIA-PET count data sets by setting the ratio of the sampling probabilities of each pair in the three groups be 6:3:1, 8:4:1 and 10:5:1. The results showed that MC_DIST also performed well. The type I error rates are 0.0005, 0.0005 and 0.0004; and the powers are 0.895, 0.899, and 0.913, respectively. Again here we see that the ratio becomes larger, the type I error rate of MC_DIST decreases and the power of MC_DIST increases. To get the 6348 new false pairs of DNA loci, we randomly selected $23 \times 3 = 69$ non-specific loci (not a ER α binding site or a gene transcription start site, three on each human chromosome), and paired them with $TFBS_{ij}$ and TSS_j , where $1 \leq i \leq 23$ and $1 \leq j \leq 2$, to get $69 \times 92 = 6348$ pairs of DNA loci.

The main advantage of MC_DIST is that it makes use of the dependency and distance information of ChIA-PET data to reduce the false positives in distinguishing the true pairs from the false ones. As such, it has a much higher sensitivity and specificity than the HG model currently in use for ChIA-PET analysis. However, there are rooms for improvement. For example, instead of using distance information based on existing databases to weigh the relevance of a site in chromatin interaction, we may directly use data from ChIP-Seq and gene expression experiments performed on the same experimental units (e.g., cell lines) to gain “first-hand” rather than “second-hand” information, which may lead to greater information in specifying the prior probabilities.

Acknowledgments: We thank Guoliang Li for the useful discussion on the results of real data analysis.

Funding: National Science Foundation (Grant / Award Number: ‘DMS-1042946’).

References

- Anders, S. and W. Huber (2010): “Differential expression analysis for sequence count data,” *Genome Biol.*, 11, R106.
- Carroll, J. S., C. A. Meyer, J. Song, W. Li, T. R. Geistlinger, J. Eeckhoutte, A. S. Brodsky, E. K. K. Keeton, K. C. Fertuck, G. F. Hall, Q. Wang, S. Bekiranov, V. Sementchenko, E. A. Fox, P. A. Silver, T. R. Gingeras, X. S. Liu and M. Brown (2006): “Genome-wide analysis of estrogen receptor binding sites,” *Nat. Genet.*, 38, 1289–1297.
- de Wit, E. and W. de Laat (2012): “A decade of 3c technologies: insights into nuclear organization,” *Genes Develop.*, 26, 11–24.
- Dekker, J., K. Rippe, M. Dekker and N. Kleckner (2002): “Capturing chromosome conformation,” *Science*, 295, 1306–1311.
- Dostie, J., T. Richmond, R. Arnaout, R. Selzer, W. Lee, T. Honan, E. Rubio, A. Krumm, J. Lamb, C. Nusbaum, R. Green and J. Dekker (2006): “Chromosome conformation capture carbon copy (5c): A massively parallel solution for mapping interactions between genomic elements,” *Genome Res.*, 16, 1299–1309.
- Fullwood, M. J., M. H. Liu, Y. F. Pan, J. Liu, H. Xu, Y. B. Mohamed, Y. L. Orlov, S. Velkov, A. Ho, P. H. Mei, E. G. Chew, P. Y. Huang, W. J. Welboren, Y. Han, H. S. Ooi, P. N. Ariyaratne, V. B. Vega, Y. Luo, P. Y. Tan, P. Y. Choy, K. D. Wansa, B. Zhao, K. S. Lim, S. C. Leow, J. S. Yow, R. Joseph, H. Li, K. V. Desai, J. S. Thomsen, Y. K. Lee, R. K. Karuturi, T. Herve, G. Bourque, H. G. Stunnenberg, X. Ruan, V. Cacheux-Rataboul, W. K. Sung, E. T. Liu, C. L. Wei, E. Cheung and Y. Ruan (2009): “An oestrogen-receptor-alpha-bound human chromatin interactome,” *Nature*, 462, 58–64.
- Gelman, A. and D. B. Rubin (1992): “Inference from iterative simulation using multiple sequences,” *Stat. Sci.*, 7, 457–472.
- Horiike, S.-i., S. Cai, M. Miyano, J.-F. F. Cheng and T. Kohwi-Shigematsu (2005): “Loss of silent-chromatin looping and impaired imprinting of *dlx5* in rett syndrome,” *Nat. Genet.*, 37, 31–40.
- Jia, L., B. P. Berman, U. Jariwala, X. Yan, J. P. Cogan, A. Walters, T. Chen, G. Buchanan, B. Frenkel and G. A. Coetzee (2008): “Genomic androgen receptor-occupied regions with different functions, defined by histone acetylation, coregulators and transcriptional capacity,” *PLoS One*, 3, e3645.
- Jiang, H. and W. H. Wong (2009): “Statistical inferences for isoform expression in rna-seq,” *Bioinformatics*, 25, 1026–1032.

- Karolchik, D., G. P. Barber, J. Casper, H. Clawson, M. S. Cline, M. Diekhans, T. R. Dreszer, P. A. Fujita, L. Guruvadoo, M. Haeussler, R. A. Harte, S. Heitner, A. S. Hinrichs, K. Learned, B. T. Lee, C. H. Li, B. J. Raney, B. Rhead, K. R. Rosenbloom, C. A. Sloan, M. L. Speir, A. S. Zweig, D. Haussler, R. M. Kuhn and W. J. Kent. (2014): "The ucsc genome browser database: 2014 update," *Nucleic Acids Res.*, 42, D764–D770.
- Labhart, P., S. Karmakar, E. M. Salicru, B. S. Egan, V. Alexiadis, B. W. O'Malley and C. L. Smith (2005): "Identification of target genes in breast cancer cells directly regulated by the src-3/aib1 coactivator," *Proc. Natl. Acad. Sci. USA*, 102, 1339–1344.
- Li, G., M. J. Fullwood, H. Xu, F. H. H. Mulawadi, S. Velkov, V. Vega, P. N. N. Ariyaratne, Y. B. B. Mohamed, H.-S. S. Ooi, C. Tennakoon, C.-L. L. Wei, Y. Ruan and W.-K. K. Sung (2010): "Chia-pet tool for comprehensive chromatin interaction analysis with paired-end tag sequencing," *Genome Biol.*, 11, R22+.
- Lieberman-Aiden, E., N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander and J. Dekker (2009): "Comprehensive mapping of long-range interactions reveals folding principles of the human genome," *Science*, 326, 289–293.
- Marioni, J. C., C. E. Mason, S. M. Mane, M. Stephens and Y. Gilad (2008): "Rnaseq: an assessment of technical reproducibility and comparison with gene expression arrays," *Genome Res.*, 18, 1509–1517.
- Murrell, A., S. Heeson and W. Reik (2004): "Interaction between differentially methylated regions partitions the imprinted genes *igf2* and *h19* into parent-specific chromatin loops," *Nat. Genet.*, 36, 889–893.
- Parisi, F., B. Sonderegger, P. Wirapati, M. Delorenzi and F. Naef (2009): "Relationship between estrogen receptor α location and gene induction reveals the importance of downstream sites and cofactors," *BMC Genom.*, 10, 381.
- Pruitt, K. D., G. R. Brown, S. M. Hiatt, F. Thibaud-Nissen, A. Astashyn, O. Ermolaeva, C. M. Farrell, J. Hart, M. J. Landrum, K. M. McGarvey, M. R. Murphy, N. A. O'Leary, S. Pujar, B. Rajput, S. H. Rangwala, L. D. Riddick, A. Shkeda, H. Sun, P. Tamez, R. E. Tully, C. Wallin, D. Webb, J. Weber, W. Wu, M. DiCuccio, P. Kitts, D. R. Maglott, T. D. Murphy and J. M. Ostell (2014): "Refseq: an update on mammalian reference sequences," *Nucleic Acids Res.*, 42, D756–D763.
- Raftery, A. E. and S. M. Lewis (1995): The number of iterations, convergence diagnostics and generic metropolis algorithms. In: Gilks, W. R., Spiegelhalter, D. J. and Richardson S. (Eds.), *Practical Markov Chain Monte Carlo*, Chapman and Hall, pp. 115–130.
- Robinson, M. D., D. J. McCarthy and G. K. Smyth (2010): "Edger: a bioconductor package for differential expression analysis of digital gene expression data," *Bioinformatics*, 26, 139–140.
- Rousseau, M., J. Fraser, M. Ferraiuolo, J. Dostie and M. Blanchette (2011): "Three-dimensional modeling of chromatin structure from interaction frequency data using markov chain monte carlo sampling," *BMC Bioinform.*, 12, 414.
- Simonis, M., P. Klous, E. Splinter, Y. Moshkin, R. Willemsen, E. de Wit, B. van Steensel and W. de Laat (2006): "Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4c)," *Nat. Genet.*, 38, 1348–1354.
- Spilianakis, C. G. and R. A. Flavell (2004): "Long-range intrachromosomal interactions in the t helper type 2 cytokine locus," *Nat. Immunol.*, 5, 1017–1027.
- Srivastava, S. and L. Chen (2010): "A two-parameter generalized poisson model to improve the analysis of rna-seq data," *Nucleic Acids Res.*, 38, e170–e170.
- Tolhuis, B., R.-J. Palstra, E. Splinter, F. Grosveld and W. de Laat (2002): "Looping and interaction between hypersensitive sites in the active β -globin locus," *Mol. Cell.*, 10, 1453–1465.
- Vernimmen, D., M. De Gobbi, J. A. Sloane-Stanley, W. G. Wood and D. R. Higgs (2007): "Long-range chromosomal interactions regulate the timing of the transition between poised and active gene expression," *The EMBO J.*, 26, 2041–2051.
- Wang, Q., W. Li, Y. Zhang, X. Yuan, K. Xu, J. Yu, Z. Chen, R. Beroukhim, H. Wang, M. Lupien, T. Wu, M. M. Regan, C. A. Meyer, J. S. Carroll, A. K. K. Manrai, O. A. Jänne, S. P. Balk, R. Mehra, B. Han, A. M. Chinnaiyan, M. A. Rubin, L. True, M. Fiorentino, C. Fiore, M. Loda, P. W. Kantoff, X. S. Liu and M. Brown (2009): "Androgen receptor regulates a distinct transcription program in androgen-independent prostate cancer," *Cell*, 138, 245–256.
- Wright, P. K., F. E. May, S. Darby, R. Saif, T. W. Lennard and B. R. Westley (2009): "Estrogen regulates vesicle trafficking gene expression in eff-3, efm-19 and mcf-7 breast cancer cells," *Int. J. Clin. Exp. Pathol.*, 2, 463.
- Yang, J.-H., J.-H. Li, S. Jiang, H. Zhou and L.-H. Qu (2013): "Chipbase: a database for decoding the transcriptional regulation of long non-coding rna and microRNA genes from chip-seq data," *Nucleic Acids Res.*, 41, D177–D187.
- Yu, J., J. Yu, R.-S. Mani, Q. Cao, C. J. Brenner, X. Cao, X. Wang, L. Wu, J. Li, M. Hu, Y. Gong, H. Cheng, B. Laxman, A. Vellaichamy, S. Shankar, Y. Li, S. M. Dhanasekaran, R. Morey, T. Barrette, R. J. Lonigro, S. A. Tomlins, S. Varambally, Z. S. Qin and A. M. Chinnaiyan (2010): "An integrated network of androgen receptor, polycomb, and *tmprss2-erg* gene fusions in prostate cancer progression," *Cancer Cell*, 17, 443–454.
- Zhao, Z., G. Tavoosidana, M. Sjölander, A. Göndör, P. Mariano, S. Wang, C. Kanduri, M. Lezcano, K. S. S. Sandhu, U. Singh, V. Pant, V. Tiwari, S. Kurukuti and R. Ohlsson (2006): "Circular chromosome conformation capture (4c) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions," *Nat. Genet.*, 38, 1341–1347.