Quantitative Biology 2016, 4(3): 217–225
DOI 10.1007/s40484-016-0080-3

REVIEW

Advances in computational ChIA-PET data analysis

Chao He[1,†], Guipeng Li[1,†], Diekidel M. Nadhir[1], Yang Chen[1,*], Xiaowo Wang[1,*] and Michael Q. Zhang[2,1,*]

[1] MOE Key Laboratory of Bioinformatics and Bioinformatics Division, TNLIST, Center for Synthetic and Systems Biology/ Department of Automation, Tsinghua University, Beijing 100084, China
[2] Department of Biological Sciences, Center for Systems Biology, the University of Texas at Dallas, Richardson, TX 75080-3021, USA
* Correspondence: yc@tsinghua.edu.cn, xwwang@tsinghua.edu.cn, michael.zhang@utdallas.edu

Received May 14, 2016; Revised July 7, 2016; Accepted July 19, 2016

Genome-wide chromatin interaction analysis has become important for understanding 3D topological structure of a genome as well as for linking distal cis-regulatory elements to their target genes. Compared to the Hi-C method, chromatin interaction analysis by paired-end tag sequencing (ChIA-PET) is unique, in that one can interrogate thousands of chromatin interactions (in a genome) mediated by a specific protein of interest at high resolution and reasonable cost. However, because of the noisy nature of the data, efficient analytical tools have become necessary. Here, we review some new computational methods recently developed by us and compare them with other existing methods. Our intention is to help readers to better understand ChIA-PET results and to guide the users on selection of the most appropriate tools for their own projects.

INTRODUCTION

Recently, ChIA-PET [1,2] became one of the few exciting high-throughput genomic technologies that can detect global chromatin interactions at the genome scale at high resolution. Compared to other methods (e.g., Hi-C [3,4]), ChIA-PET allows for mapping of almost all *in vivo* chromatin interactions mediated by a specific protein of interest. At a similar sequencing depth, ChIA-PET can detect chromatin interactions at much higher resolution, which is necessary for studying gene regulation, such as long-range interactions of enhancers and their target gene promoters. Therefore, ChIA-PET is suitable for studying regulatory function of a specific protein factor via its chromatin interactions.

However, a ChIA-PET experiment is relatively complicated and has some characteristics and inevitable noisy sources. It requires effective computational tools to remove this noise in order to detect interactions with high confidence. Below we will first review our new method MICC [5] which involves a hierarchical mixed probability model for calling statistically significant interactions. Then we will describe how to build a machine learning model that can use multiple 1D ChIP-Seq data to predict 3D chromatin interactions [6] and hence can help to improve ChIA-PET data analysis or to rescue missed/novel interactions. Finally, we will introduce our method 3CPET [7], which allows users to identify possible major protein complexes mediating the ChIA-PET-detected interactions at different loci in the context of 3D genome.

IMPROVED ChIA-PET NOISE MODELS

A common framework for processing ChIA-PET raw data contains five steps. First, one must filter out linkers. For the half-linkers ChIA-PET protocol [1], one has to remove the two types of half-linkers contained in the sequenced paired-end-tags (PETs). The PETs with two different types of half-linkers are designated as chimeric PETs which are used to estimate the frequency of random ligations in solution, while those with the same types of half-linkers are non-chimeric PETs containing both noise and the true signal. For the bridge-linker ChIA-PET protocol [8], the

[†] These authors contributed equally to this work.

This article is dedicated to the Special Collection of Recent Advances in Next-Generation Bioinformatics (Ed. Xuegong Zhang).

bridge linkers also need to be trimmed. Second, one has to map PETs to the reference genome. The two ends of a PET mark the location pair of the interaction sites. Third, one carefully classifies the PETs. Because there is a considerable proportion of PETs formed by self-ligation of single DNA fragments, these self-ligation PETs have relatively smaller spans between the two ends than intrachromosomal inter-ligation PETs. They should be filtered out to enrich functional PETs. Fourth, one has to cluster the PETs. Those PETs that link the same two anchor sites (generally protein-binding sites) are merged to form a PET cluster. Lastly, one can detect potentially true interactions (e.g., with high likelihood) among PET clusters by means of computational models. There are two main types of noise in all ChIA-PET interaction data: chromatin random ligation in solution and chromatin random collision during cross-linking. The major difference between our method and the existing methods lies in the model used to remove these types of noise at this last step. To further improve ChIA-PET data analysis, one would have to integrate other data types.

There are a few methods for processing of ChIA-PET data. Three of them are used for comparison with our MICC method in this article. They are the ChIA-PET Tool [9], ChiaSig [10], and Mango [11]. The ChIA-PET Tool is the first freely available software to deal with ChIA-PET data. It implies that there is only random-ligation noise and uses a PET count (no less than 3) to filter out interactions [9]. This strategy, however, misses many weaker but probably true interactions. ChiaSig takes the two types of noise into consideration and shows an improvement over the ChIA-PET Tool via comparison with 5C data [10]. However, ChiaSig is much too conservative and thus has a high false negative rate [5]. Mango models the probability of observing an interaction between genomic loci as a function of both genomic distance and peak depth and then uses this model to estimate statistical confidence of the interactions [11]. MICC, like ChiaSig, is aimed at removal of both random-ligation noise and random-collision noise simultaneously. It can detect chromatin interactions with high sensitivity while controlling the false discovery rate (FDR) at a reasonable level.

Here we take a brief look at the model underlying each of these four methods, and the details of the methods can be found in the original papers. For a PET cluster (A, B), where A and B are two anchor sites, we denote i) c_{AB} as the count of PETs that link anchor A and B, ii) $c_A(c_B)$ as the total PET count linking anchor A (B), and iii) d_{AB} as the distance between two anchor regions ($d_{AB} = +\infty$ if A and B are in two different chromosomes).

The assumption in the ChIA-PET Tool is that most of the noise derives from random ligations in solution between two ends of different DNA-protein complexes. The ligation is completely random; thus, the ligation probability is associated with the total PET count of either anchor. Specifically, they use a hypergeometric distribution to describe the probability of PET cluster (A, B) as a random ligation PET cluster (RlPC), i.e.,

$$P(c_{AB}|c_A,c_B) = dhyper(c_{AB},c_A,2N-c_A,c_B) = \frac{\binom{c_A}{c_{AB}}\binom{2N-c_A}{c_B-c_{AB}}}{\binom{2N}{c_B}},$$

where N is the sum of all PETs and $dhyper()$ is the probability mass function of the hypergeometric distribution. Therefore, the significance of a PET cluster as a True interaction PET Cluster (TiPC) can be evaluated by means of the hypergeometric p-value(s). The FDR is estimated via a comparison with permutated p-value(s) from a randomly shuffled dataset. Because the hypergeometric p-value(s) tends to be optimistic for such data, the ChIA-PET Tool requires TiPCs to have a PET count no less than 3 for predictions with higher confidence.

ChiaSig adds the description of random-collision noise and a newly developed non-central hyper-geometric test based on the ChIA-PET Tool. ChiaSig implies that the probability of random collision events positively correlates with the distance between two ends of the PET. Specifically, the probability of a PET cluster that is derived from noise is given by

$$P(c_{AB}|c_A,c_B,d_{AB}) = \frac{\binom{c_A}{c_{AB}}\binom{2N-c_A}{c_B-c_{AB}}[\omega(d_{AB})]^{c_{AB}}}{\Sigma_{(A,B)}\binom{c_A}{c_{AB}}\binom{2N-c_A}{c_B-c_{AB}}[\omega(d_{AB})]^{c_{AB}}},$$

where N is the sum of all PETs and $\omega(d_{AB})$ is a distance-related function that is used to describe the probability of a PET with span d_{AB} that is derived from a random collision. The significance of a PET cluster as a TiPC can be evaluated by means of the p-value(s) of this non-central hyper-geometric test. The FDRs are estimated using a discrete FDR procedure [12]. Predictions with good confidence also require to have at least three PETs.

Mango first models the probability of observing a single PET linking two loci as a function of their genomic distance d_{AB} and of the product of their read depths $R_{AB} = c_A \times c_B$ as

$$P(I) = \frac{P(I|d_{AB}) \times P(I|R_{AB})}{P(C|d_{AB}) \times P(C|R_{AB}) \times C_T},$$

where $P(I|d_{AB})$ represents the probability of observing a PET that links loci at distance d_{AB}, $P(I|R_{AB})$ is the probability of observing a PET linking loci with depth

R_{AB}, $P(C|d_{AB})$ denotes the probability of observing a pair of loci at distance d_{AB} (regardless of whether any PETs link the two loci), $P(C|R_{AB})$ means the probability of observing a pair of loci with depth R_{AB} (regardless of whether any PETs link the two loci), and C_T is the total number of pairwise combinations of the loci. All the terms on the right side of the equation can be estimated from the data. Note that this equation is based on the assumption that the PET genomic distance of separation and the product of the corresponding read depths are independent. Then, according to the binomial distribution, the probability of observing exactly c_{AB} PETs can be determined as

$$P(K=c_{AB})=\binom{N}{c_{AB}}P(I)^{c_{AB}}(1-P(I))^{N-c_{AB}},$$

where N is the total number of PETs. Finally, p-value(s) $[P(K\geqslant c_{AB})=\Sigma_{i=c_{AB}}^{N}P(K=i)]$ for all possible pairs of interacting loci are corrected for multiple hypothesis testing using the Benjamini-Hochberg procedure [13].

Our MICC method builds a mixture model with three components to distinguish TiPCs from noise. It uses the Zeta distribution [14] to describe the probability of PET count distribution for TiPCs and Random collision PET Clusters (RcPCs), and uses hypergeometric distribution to describe the PET count of RlPCs. The coefficients of the Zeta distribution for TiPCs and RcPCs are different, depending on distance between the anchors. Besides, it is assumed that the prior probability of RlPCs also depends on the distance of PET clusters, according to the observation that chimeric PETs and non-chimeric PETs are well separated by the PET spans. The prior probability of RcPCs is also associated with the total PET count in anchor regions. Let $I_{AB}=1$ denote (A, B) as a TiPC, as $I_{AB}=2$ an RcPC and $I_{AB}=3$ as an RlPC. The full probability to observe a PET cluster (A, B) is therefore given by

$$\begin{aligned}
&P(c_{AB}|c_A,c_B,d_{AB})\\
&=P(c_{AB}|d_{AB},I_{AB}=1)P(I_{AB}=1|c_A,c_B,d_{AB})\\
&\quad+P(c_{AB}|d_{AB},I_{AB}=2)P(I_{AB}=2|c_A,c_B,d_{AB})\\
&\quad+P(c_{AB}|c_A,c_B,I_{AB}=3)P(I_{AB}=3|d_{AB})\\
&=P(c_{AB}|d_{AB},I_{AB}=1)P(I_{AB}=1|c_A,c_B,I_{AB}\neq 3)\\
&\quad P(I_{AB}\neq 3|d_{AB})\\
&\quad+P(c_{AB}|d_{AB},I_{AB}=2)P(I_{AB}=2|c_A,c_B,I_{AB}\neq 3)\\
&\quad P(I_{AB}\neq 3|d_{AB})\\
&\quad+P(c_{AB}|c_A,c_B,I_{AB}=3)P(I_{AB}=3|d_{AB})\\
&=P(c_{AB}|d_{AB},I_{AB}=1)(1-\mu(c_A,c_B))(1-\lambda(d_{AB}))\\
&\quad+P(c_{AB}|d_{AB},I_{AB}=2)\mu(c_A,c_B)(1-\lambda(d_{AB}))\\
&\quad+P(c_{AB}|c_A,c_B,I_{AB}=3)\lambda(d_{AB}).
\end{aligned}$$

The details for each term can be found in the supplementary material of the MICC article [5]. The FDR is estimated by comparing the posterior probabilities from the original dataset to those from a randomly generated dataset.

The comparison of performance among these four methods was conducted by applying them to real-life datasets. Although MICC and Mango were compared with ChiaSig and the ChIA-PET Tool respectively, in the original research papers, there is no direct comparison between MICC and Mango yet. To fill this gap, we evaluated their performance on K562 Pol2 ChIA-PET dataset GSE33664 [2]. The PET clusters as input of MICC were obtained from the Mango pipeline, which makes the comparison between MICC and Mango fairer. On the basis of the comparison results shown in Figure 1 and those in MICC and Mango original papers, we can summarize the results as follows. First, MICC can produce the greatest number of interactions and ChiaSig the smallest number. Via a sampling strategy, one can see that MICC can recover more interactions in the original datasets from less sampled sequencing libraries than other methods can. This means that MICC can yield more consistent results between deeply sequenced and shallowly sequenced libraries. Second, MICC shows the best reproducibility between two experimental replicates, especially when we select the same number of top-ranked interactions. This finding suggests that MICC can remove experimental noise in a more consistent manner. Third, although the proportion of MICC-detected significant interactions that can be validated by 5C experiments [15] is similar to that of ChiaSig and Mango, MICC can statistically significantly detect more 5C-validated interactions in ChIA-PET data. This result indicates that MICC has a higher sensitivity at a similar FDR as compared with the other methods. Lastly, the time cost for calling significant interactions in PET clusters, can be ranked as follows (from the lowest to highest): the ChIA-PET Tool, MICC, Mango and ChiaSig. All of them require less than 24 hours to run one current dataset as input of PET clusters. The ChIA-PET Tool and Mango provide more complete pipelines for processing half-linkers in ChIA-PET raw datasets, including all five steps namely linker removal, read mapping, PETs filtering, PET clustering and significant interaction calling. ChiaSig and MICC simply focus on modeling the randomness of ligation and collision to detect the significant chromatin interactions. Table 1 shows a brief comparison of these methods. Users can choose one of the methods in accordance with the requirements of a practical application.

PREDICTION OF 3D INTERACTIONS BY INTEGRATING 1D DATA

Although the ChIA-PET experiment can detect chromatin interactions genome wide, the list of interactions

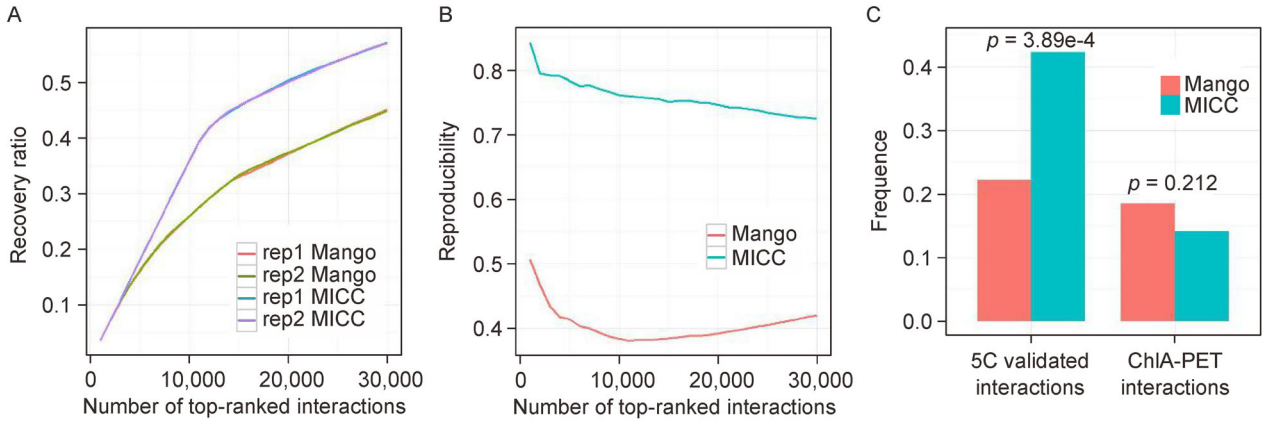

Figure 1. Performance comparison between MICC and Mango. (A) The proportion of interactions called in combined libraries of two Pol2 ChIA-PET replicates recovered by a single replicate. (B) The proportion of interactions overlapping between the sets of top-ranked interactions from two Pol2 ChIA-PET replicates detected by Mango and MICC respectively. (C) The proportion of all 5C-validated ChIA-PET interactions that are predicted by either the computational method (left) and the proportion of predicted ChIA-PET interactions validated by 5C (right). *p*-value(s) were obtained by Fisher's exact test.

Table 1. Comparison of the ChIA-PET Tool, ChiaSig, Mango and MICC.

Methods	ChIA-PET Tool	ChiaSig	Mango	MICC
Model	Hyper-geometric test	Non-central hyper-geometric test	Binomial model	Mixture model
Consider random ligation	Yes	Yes	Yes	Yes
Consider random collision	No	Yes	Yes	Yes
Applicable to inter-chromosomal interactions	Yes	No	No	Yes
Filter interactions by PET-count	Yes	Yes	Yes	No
Number of interactions	Moderate	Least	Moderate	Most
Reproducibility	Moderate	Worst	Moderate	Best
Time cost	Least	Most	Moderate	Moderate
Complete pipeline	Yes	No	Yes	No

identified from currently available ChIA-PET data is far from comprehensive. For example, many 5C-verified interactions cannot be recovered by ChIA-PET, whereas a considerable proportion of ChIA-PET predictions cannot be validated by 3C or 5C experiment [5,15]. Along with the availability of abundant genomic and epigenomic high throughput data, we tested whether we can integrate the information from multiple 1D omics data like transcription factor (TF) binding profiles and histone modification profiles to better study long-range chromatin interactions.

Under the assumption that long-range interactions can be determined by means of local chromatin states which are defined by histone modification and a TF-binding pattern, we collected the publicly available TF and histone modification ChIP-Seq data, and DNase-Seq data from the same experimental conditions to predict ERα-associated interactions in MCF7 cells [6].

Using a ChIA-PET experiment, Fullwood *et al.* [1,16] detected thousands of ERα associated interactions in MCF7 cells. Many of these interactions link distal ERα binding sites (ERBSes) to target gene promoters. However, a larger proportion of ERBSes was not found to be associated with any long-range interactions. We determined whether there are any features that can separate these loop-associated ERBSes (laERBSes) and non-loop-associated ERBSes (nlaERBSes). We found that the average ChIP-Seq signals of multiple histone modifications (such as H3K4me1/3, H3K9ac) are significantly depleted in the center of laERBSes than those in nlaERBSes. In the meantime, the average DNase-Seq signal, which reflects the state of open chromatin, is significantly enriched in the center of laERBSes than in the center of nlaERBSes. Furthermore, we found that ERα cofactors such as FoxA1, GATA3, AP2γ, and p300 are significantly more enriched in laERBSes than in nlaERBSes. Taken together, these data suggest that

laERBSes are more likely to be located in the nucleosome-free region, binding to co-factors, though the binding of ERα itself to DNA does not require nucleosome depletion [17]. Furthermore, the ERBSes with a neighboring ERBS (distance less than 3 kb) are significantly more likely to form loops with some other ERBSes.

According to the features listed above, we designed a two-step classifier to predict ERα associated interactions using multiple 1D omics data. We selected 903 high-confidence ERα-associated interaction as a foreground and randomly sampled ERBSes that are not associated with any interactions as the background to compile the training set. The first step was to find a possible laERBS among the ~15,000 ERBSes. Three types of features were computed:

(i) $F_{i,j}$: log-transformed read counts for ChIP-Seq data on protein j (or DNase-Seq data) in ERBS i, with a window size of 400 bp centered around the ERBS peak summit.

(ii) D_i: log-transformed distance between the ERBS neighboring ERBS i and ERBS i itself.

(iii) $H_{i,j}$: differences between log-transformed ratios of read coverage against input in the central region (± 100 bp region relative to the peak summit) to the average of read coverage against input in the two flanking regions (-400 bp to -200 bp and $+200$ bp to $+400$ bp relative to the peak summit) of ERBS i for histone modification j.

The logistical classifier is designed as

$$P(E_i=1) = \left\{1+exp\left(-k_0-\sum\nolimits_j a_j F_{i,j}-bD_i-\sum\nolimits_j c_j H_{i,j}\right)\right\}^{-1}$$

where E_i is the indicator of whether ERBS i is a laERBS; $P()$ is a probability function; and $k_0, a_j, b,$ *and* c_j are the model parameters. After training, ERBSes were filtered by setting the threshold of the logistic classifier to 0.2 in order to find putative laERBSes. After this step, 96% (869 out of 903) training foreground interactions passed the threshold and ~11,000 ERBSs were kept for further analysis.

The second step was to predict interactions between laERBS pairs. Recently, Hi-C experiments showed that chromosomes are organized as large topological domains, and interactions between regulatory elements largely take place within these domains. Such domain structure is highly conserved among different cell types [18]. We noticed that the use of the domain information can greatly reduce the number of candidate pairs but keep most of the true interactions. Thus we restricted the predictions to the laERBSes pairs within each topological domain. This filter can retain the majority of true interactions in the training set (800 out of 869 training interactions that pass the first step) but ruled out more than 98% random assortment ERBS pairs. Next, two types of features were computed for each candidate pair:

(i) $PF_{i_1 i_2, j}=F_{i_1, j}+F_{i_2, j}$: the sum of log-transformed jth ChIP-Seq (or DNase-Seq) read counts for each candidate ERBS pair (i_1, i_2), with a window size of 3 kb for each end (the region ± 1.5 kb relative to the peak summit).

(ii) $PD_{i_1 i_2, j}$: the log-distance and inverse distance between each ERBS pair (i_1, i_2)

The logistical classifier is expressed as

$$P(PE_{i_1 i_2}=1) = \left\{1+exp\left(-k_0-\sum_j a_j PE_{i_1 i_2, j}-\sum_j b_j PD_{i_1 i_2, j}\right)\right\}^{-1}$$

where $PE_{i_1 i_2}$ is the indicator of whether ERBS pair (i_1, i_2) formed an interaction, $P()$ is a probability function, and $k_0, a_j,$ *and* b_j are the model parameters. The average true positive rate (TPR) was 93% and average false positive rate (FPRs) was 8% for five-fold cross validation for the training set.

Over all, the two-step classifier recovered a large proportion (2,356 of 3,527) of ERα interactions revealed by ChIA-PET experiments. Meanwhile, 8,805 unreported putative ERα-associated interactions were predicted, many of which can be validated by independent 3C or Pol II ChIA-PET experiments. This work suggested that the chromatin interactions are determined or largely influenced by local genetic and epigenetic status of the anchor sites. The use of a machine learning model integrating multiple 1D ChIP-Seq data can predict the majority of ChIA-PET-identified interactions. Such computational approaches may be a valuable addition to a ChIA-PET experiment.

PREDICTION OF PROTEIN COMPLEXES MEDIATING ChIA-PET DETECTED INTERACTIONS

Different studies showed that proteins preferentially bind to specific loci to participate in the establishment and maintenance of chromatin interactions [19]. On the other hand, existing studies are mainly focused on studying some specific proteins such as the role of architectural proteins like CTCF and Cohesin [20,21], even though it is known that other proteins are also involved in the chromatin interaction [22,23]. This drawback is mainly due to the limitations of existing experimental methods; as for ChIP-Seq studies, it is hard to design sensitive and specific antibodies, whereas for mass-spectrometry, the obtained data cannot help us to distinguish between proteins that are involved in chromatin interactions and those that are not.

Among the existing experimental assays, only ChIA-PET can give researchers information about chromatin interactions involving a certain protein of interest. The value of this information can be leveraged further after its integration with various ChIP-Seq and protein-interaction data to detect the partners of the ChIA-PET target protein.

Thus, to fill this gap, we developed 3CPET [7], a tool based on a nonparametric Bayesian approach, to infer the set of the most probable protein complexes involved in the maintenance of chromatin interactions. The rationale behind 3CPET is that transcription factors can use a distinct combination of collaborating proteins, depending on the genomic and spatial context. To detect these protein complexes, 3CPET builds for each ChIA-PET interaction a local protein-protein interaction (PPI) network connecting its interacting chromatin segments. First, a TF-related ChIP-Seq signal is used to detect the TFs involved at the anchors of each interaction with DNA. Then, we build the local PPIs by connecting each TF on one side of the interaction to the other TFs on the other side by taking the shortest path between them in the background PPI.

Because the existing PPI networks are not condition specific, a filtering step is needed to remove all noisy interactions. First, we consider only nuclear proteins. Second, we filter all the non-expressed proteins. Third, we consider only the proteins that are co-expressed and show a physical interaction in the PPI network. These three filters help us limit the false positive predictions. Ideally, a cell type specific PPI should be used.

Given the generated corpus of protein-protein interactions, we aimed to find the set of the most enriched protein subnetworks. Here, we refer to these subnetworks as chromatin maintainer networks (CMNs). In our corpus, we can see that the edges of each protein network can be considered sampled from a mixture of different CMNs in different mixture proportions in each network. To detect these CMNs, 3CPET uses the hierarchical Dirichlet process (HDP) model [24]. There are two reasons for the usage of HDP. First, we noticed that the distribution of CMNs shows a kind of hierarchy: at the corpus level, the different CMNs show different enrichment proportions and at the chromatin interaction level, each chromatin interaction is enriched for different CMNs and in different proportions (Figure 2A). Second, instead of assuming a fixed number of CMNs, we wanted to automatically infer the latent enriched CMNs depending on the data at hand.

Because the number of CMNs is not known, it is allowed to grow infinitely. However, because the Dirichlet distribution is defined on the basis of a k-dimensional simplex, the HDP model uses the Dirichlet process(DP) as an extension of the Dirichlet distribution into the continuous space. If $G_j \sim DP(\alpha, G_0)$ with base distribution G_0 and concentration parameter α, we can express G_j using the stick-breaking representation $G_j = \Sigma_{k=1}^{\infty} \alpha_k \delta \beta_k$ with β_k representing the CMNs (Figure 2B). Using the stick-breaking representation and introducing label z_{jn} for each edge e_{jn}, the model can be expressed as:

$$H \sim Dir(\eta), \qquad \beta_k | H \sim H,$$

$$\pi | \gamma \sim Dir(\gamma_1, \gamma_2, ..., \gamma_K, \gamma_u), \qquad \theta_n | \alpha, \pi \sim Dir(\alpha, \pi),$$

$$z_{jn} | \theta_n \sim Mult(\theta_n), \qquad e_{jn} | z_{jn}, (\beta_k)_{k=1}^{L} \sim F(\beta_{z_{jn}}),$$

where $\gamma_k = \frac{\gamma}{L}, k = 1 \ldots K$ and $\gamma_u = 1 - \Sigma_{k=1}^{K} \gamma_k$ and $F(\beta_k)$ indicates the probability distribution of a CMN over all the possible edges in our vocabulary.

Because the HDP algorithm is based on counting element frequency, each PPI in our corpus is converted into a bag of edges in which the frequency of each edge is equal to the number of the shortest paths in which it participates. This approach can help us prioritize important edges in each local PPI. To avoid the bias introduced by hub proteins in the background PPI and rare interactions, we filter interactions that appear to be more or less than a maximal threshold and minimal threshold.

3CPET was tested on ER-α and Pol-II associated interactions from ChIA-PET. In our study, we showed that 3CPET can predict known ER-α co-factors and yield a significant overlap with other previously reported experimental results detected in the RIME experiment [25]. Moreover, we showed that ER-α associates with different partner proteins in different genomic contexts. Similarly, using the RNAP-II associated interactions, we were able to predict cofactors known to be a part of known transcription-related complexes such as Mediator and showed variability in RNAP-II co-factors. Simulation tests also indicated high robustness of the method. 3CPET was not designed solely for protein interactions, other elements such as non-coding RNAs can be integrated into the PPI network.

CONCLUDING REMARKS AND FUTURE PERSPECTIVES

In this review, we first introduce the basic computational procedures of analyzing ChIA-PET raw data to get candidate chromatin interactions. Since ChIA-PET data may be quite noisy, it is important to distinguish the genuine chromatin interactions from the random noise. Several computational methods were proposed to address this problem in the recent years. We compare these existing statistical models which estimate the confidence of chromatin interactions to help users better understand the rationale behind these models and choose appropriate tools in their own study. In the case of no ChIA-PET data

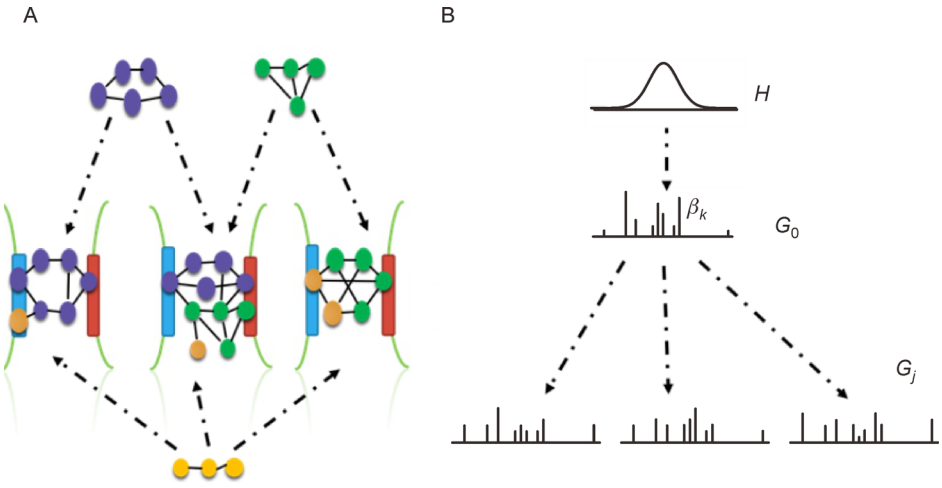

Figure 2. 3CPET model. (A) In 3CPET we model each edge as a sample from a co-factor complex, thus, each PPI associated with a DNA interaction can be considered as mixture of some co-factor complexes shared among our PPI corpus. (B) To allow the sharing of CMNs between the different DNA interactions, a hierarchical sampling is introduced. Instead of sampling directly from the continues base distribution H, an intermediate distribution G_0 is introduced to first sample a discrete number of CMNs atoms from distribution H then to use them as a basis to sample the labels of the different edges in our PPI networks collection, thus, ensuring the sharing of CMNs as the number of labels is less than the number of edges.

at hand, one can predict the chromatin interactions from 1D data, such as ChIP-Seq data. The predicted chromatin interactions may be a good starting point for further study but users should also be aware of the false positive discovery of the predictive method. Since ChIA-PET was designed to detect chromatin interactions mediated by one specific protein, it is hardly to know the protein complexes which mediate the chromatin interactions with only ChIA-PET data. So we introduced a hierarchical Dirichlet process model which leverage ChIA-PET, ChIP-Seq and PPI network to predict the protein complexes mediating ChIA-PET detected interactions. These computational methods will be helpful in the ChIA-PET data analysis.

Besides the perspectives we mentioned above, there are also other attractive topics in the ChIA-PET data analysis, e.g. integrating ChIA-PET chromatin interactions with genetic variants. ChIA-PET has been mainly popular among researchers who want to link genetic mutations to target genes. Lately, many investigators have been using ChIA-PET information as an additional layer to select strong SNP-to-gene associations. This kind of studies can be classified into two types: i) integrative studies that are focused on generation of global datasets and ii) specific studies where ChIA-PET data are used to explore SNP-gene associations in a specific case.

Among the integrative studies on SNP-gene associations, we can list GWAS3D, which integrates chromatin interaction data, chromatin state, functional genomic and conservation data on top of GWAS studies designed to associate genetic variance with regulatory pathways [26]. Another work of interest is a study by the Snyder lab where they generated a collection of histone quantitative trait loci (hQTLs) and chromatin contacts from ChIA-PET to identify connections between genetic variations and histone modifications changed at the distal elements [27].

Other groups used ChIA-PET to focus on study of specific diseases. Gerald *et al*. developed a multilevel mapping method where they used ChIA-PET and Hi-C data to screen novel noncoding SNPs associated with a psychotropic drug response, e.g., disruption of a lithium response correlates with the SNP in the promoter of the *SLC1A2* gene [28]. Smemo *et al*. used ChIA-PET to uncover the mechanisms of action of a noncoding SNP in the intron of the *FTO* gene and showed that this SNP interacts with the promoter of the *IRX3* gene, thus identifying *IRX3* as a new candidate for an obesity gene [29]. Hnisz *et al.* found that perturbation of CTCF-CTCF loops in nonmalignant cells is sufficient to activate proto-oncogenes, e.g., *TAL1* gene [30]. For more applications and comparison of ChIA-PET and Hi-C results, we refer readers to other articles [8,31,32]. We expect that integration analysis will be an attractive topic regarding to ChIA-PET data analysis in the near future, especially when more ChIA-PET datasets are coming out.

ACKNOWLEDGEMENTS

This work is supported by the National Basic Research Program of China (No. 2012CB316503), the National Nature Science Foundation of China (Nos. 91519326, 31361163004 and 31301044) and Tsinghua National

Laboratory for Information Science and Technology Cross-discipline Foundation. We thank Zhenyu Liang for help of cover figure design.

COMPLIANCE WITH ETHICS GUIDELINES

The authors Chao He, Guipeng Li, Diekidel M. Nadhir, Yang Chen, Xiaowo Wang and Michael Q. Zhang declare that they have no conflict of interests.

This article does not contain any studies with human or animal subjects performed by any of the authors.

REFERENCES

1. Fullwood, M. J., Liu, M. H., Pan, Y. F., Liu, J., Xu, H., Mohamed, Y. B., Orlov, Y. L., Velkov, S., Ho, A., Mei, P. H., *et al.* (2009) An oestrogen-receptor-alpha-bound human chromatin interactome. Nature, 462, 58–64
2. Li, G., Ruan, X., Auerbach, R. K., Sandhu, K. S., Zheng, M., Wang, P., Poh, H. M., Goh, Y., Lim, J., Zhang, J., *et al.* (2012) Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. Cell, 148, 84–98
3. Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science, 326, 289–293.
4. Rao, S. S., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., Sanborn, A. L., Machol, I., Omer, A. D., Lander, E. S., *et al.* (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell, 159, 1665–1680.
5. He, C., Zhang, M. Q. and Wang, X. (2015) MICC: an R package for identifying chromatin interactions from ChIA-PET data. Bioinformatics, 31, 3832–3834
6. He, C., Wang, X. and Zhang, M. Q. (2014) Nucleosome eviction and multiple co-factor binding predict estrogen-receptor-alpha-associated long-range interactions. Nucleic Acids Res., 42, 6935–6944
7. Djekidel, M. N., Liang, Z., Wang, Q., Hu, Z., Li, G., Chen, Y. and Zhang, M. Q. (2015) 3CPET: finding co-factor complexes from ChIA-PET data using a hierarchical Dirichlet process. Genome Biol., 16, 288
8. Tang, Z., Luo, O. J., Li, X., Zheng, M., Zhu, J. J., 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
9. Li, G., Fullwood, M. J., Xu, H., Mulawadi, F. H., Velkov, S., Vega, V., Ariyaratne, P. N., Mohamed, Y. B., Ooi, H. S., Tennakoon, C., *et al.* (2010) ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing. Genome Biol., 11, R22
10. Paulsen, J., Rødland, E. A., Holden, L., Holden, M. and Hovig, E. (2014) A statistical model of ChIA-PET data for accurate detection of chromatin 3D interactions. Nucleic Acids Res., 42, e143
11. Phanstiel, D. H., Boyle, A. P., Heidari, N. and Snyder, M. P. (2015) Mango: a bias-correcting ChIA-PET analysis pipeline. Bioinformatics, 31, 3092–3098
12. Heyse, J. (2011) A false discovery rate procedure for categorical data. In Resent Advances in Biostatistics: False Discovery Rates, Survival Analysis, and Related Topics, 43–58, World Scientific Publishing Company
13. Benjamini YaH, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. B Methodol. 57, 289–300
14. Jessen, B. and Wintner, A. (1935) Distribution functions and the Riemann Zeta function. Trans. Am. Math. Soc., 38, 48–88
15. Sanyal, A., Lajoie, B. R., Jain, G. and Dekker, J. (2012) The long-range interaction landscape of gene promoters. Nature, 489, 109–113
16. Fullwood, M. J., Wei, C. L., Liu, E. T. and Ruan, Y. (2009) Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. Genome Res., 19, 521–532
17. He, H. H., Meyer, C. A., Chen, M. W., Jordan, V. C., Brown, M. and Liu, X. S. (2012) Differential DNase I hypersensitivity reveals factor-dependent chromatin dynamics. Genome Res., 22, 1015–1025
18. Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S. and Ren, B. (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature, 485, 376–380
19. Marsman J., Horsfield, J.(2012) Long distance relationships: enhancer–promoter communication and dynamic gene transcription. Biochim. Biophys. Acta, 1819:1217–1227
20. Phillips-Cremins, J. E., Sauria, M. E., Sanyal, A., Gerasimova, T. I., Lajoie, B. R., Bell, J. S., Ong, C. T., Hookway, T. A., Guo, C., Sun, Y., *et al.* (2013) Architectural protein subclasses shape 3D organization of genomes during lineage commitment. Cell, 153, 1281–1295
21. Kagey, M. H., Newman, J. J., Bilodeau, S., Zhan, Y., Orlando, D. A., van Berkum, N. L., Ebmeier, C. C., Goossens, J., Rahl, P. B., Levine, S. S., *et al.* (2010) Mediator and cohesin connect gene expression and chromatin architecture. Nature, 467, 430–435
22. Lan, X., Witt, H., Katsumura, K., Ye, Z., Wang, Q., Bresnick, E. H., Farnham, P. J. and Jin, V. X. (2012) Integration of Hi-C and ChIP-seq data reveals distinct types of chromatin linkages. Nucleic Acids Res., 40, 7690–7704
23. Deng, W., Lee, J., Wang, H., Miller, J., Reik, A., Gregory, P. D., Dean, A. and Blobel, G. A. (2012) Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. Cell, 149, 1233–1244
24. Teha, Y.W., Jordana, M. I., Beala, M. J. and Bleia, D. M. (2006) Hierarchical Dirichlet processes. J. Am. Stat. Assoc., 101, 1566–1581
25. Mohammed, H., D'Santos, C., Serandour, A. A., Ali, H. R., Brown, G. D., Atkins, A., Rueda, O. M., Holmes, K. A., Theodorou, V., Robinson, J. L., *et al.* (2013) Endogenous purification reveals GREB1 as a key estrogen receptor regulatory factor. Cell Reports, 3, 342–349
26. Li, M. J., Wang, L.Y., Xia, Z., Sham, P.C., Wang, J. (2013) GWAS3D: Detecting human regulatory variants by integrative analysis of genome-wide associations, chromosome interactions and histone modifications. Nucl. Acids Res. 41, W150–W158
27. Grubert, F., Zaugg, J. B., Kasowski, M., Ursu, O., Spacek, D. V., Martin, A. R., Greenside, P., Srivas, R., Phanstiel, D. H., Pekowska, A., *et al.* (2015) Genetic control of chromatin states in humans involves local and distal chromosomal interactions. Cell, 162, 1051–1065
28. Higgins, G. A., Allyn-Feuer, A. and Athey, B. D. (2015) Epigenomic mapping and effect sizes of noncoding variants associated with psychotropic drug response. Pharmacogenomics, 16, 1565–1583
29. Smemo, S., Tena, J. J., Kim, K. H., Gamazon, E. R., Sakabe, N. J., Gómez-Marín, C., Aneas, I., Credidio, F. L., Sobreira, D. R., Wasserman, N. F., *et al.* (2014) Obesity-associated variants within *FTO* form long-range functional connections with *IRX3*. Nature, 507, 371–375

30. Hnisz, D., Weintraub, A. S., Day, D. S., Valton, A. L., Bak, R. O., Li, C. H., Goldmann, J., Lajoie, B. R., Fan, Z. P., Sigova, A. A., *et al.* (2016) Activation of proto-oncogenes by disruption of chromosome neighborhoods. Science, 351, 1454–1458
31. Heidari, N., Phanstiel, D. H., He, C., Grubert, F., Jahanbani, F., Kasowski, M., Zhang, M. Q. and Snyder, M. P. (2014) Genome-wide map of regulatory interactions in the human genome. Genome Res., 24, 1905–1917
32. Li, G., Cai, L., Chang, H., Hong, P., Zhou, Q., Kulakova, E. V., Kolchanov, N. A. and Ruan, Y. (2014) Chromatin interaction analysis with paired-end tag (ChIA-PET) sequencing technology and application. BMC Genomics, 15, S11

