Bioinformatics, 31(23), 2015, 3832–3834
doi: 10.1093/bioinformatics/btv445
Advance Access Publication Date: 31 July 2015
Applications Note

OXFORD

Genome analysis

MICC: an R package for identifying chromatin interactions from ChIA-PET data

Chao He[1], Michael Q. Zhang[2,1,*] and Xiaowo Wang[1,*]

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

*To whom correspondence should be addressed.

Associate Editor: Inanc Birol

Received on April 20, 2015; revised on July 5, 2015; accepted on July 27, 2015

Abstract

Summary: ChIA-PET is rapidly emerging as an important experimental approach to detect chromatin long-range interactions at high resolution. Here, we present *M*odel based *I*nteraction *C*alling from *C*hIA-PET data (MICC), an easy-to-use R package to detect chromatin interactions from ChIA-PET sequencing data. By applying a Bayesian mixture model to systematically remove random ligation and random collision noise, MICC could identify chromatin interactions with a significantly higher sensitivity than existing methods at the same false discovery rate.
Availability and implementation: http://bioinfo.au.tsinghua.edu.cn/member/xwwang/MICCusage
Contact: michael.zhang@utdallas.edu or xwwang@tsinghua.edu.cn
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

High-throughput sequencing technologies aiming to detect chromatin interactions are rapidly developing these years. Among them, ChIA-PET (Fullwood *et al.*, 2009) can genome-widely detect chromatin interactions that are associated with the protein of interest. It is suitable for studying physical interactions between regulatory elements, such as enhancer–promoter interactions. However, ChIA-PET experiment suffers from high levels of noises caused by chromatin random collision events and random ligation in solution. Therefore, it needs effective computational methods to process raw ChIA-PET data.

Up to now, there are few tools in hand to deal with ChIA-PET data. ChIA-PET tool (Li *et al.*, 2010) is the first one freely available to the public, which combines *P*-values from hyper-geometric test and an arbitrary and strict threshold for no less than 3 *P*aired-*e*nd *t*ags (PETs) to identify chromatin interactions. However, this strategy does not consider random collision noise, and the cut-off for PET-count may result in losing some relatively weaker but non-random interactions. Recently, another method called ChiaSig (Paulsen *et al.*, 2015) tries to take chromatin random collision events into account. The authors showed an improvement than ChIA-PET tool by comparing with 5C data. However, ChiaSig is much too conservative, thus suffers from high false negative rate. Here we present MICC, an easy-to-use R package to process ChIA-PET data. It aims to detect chromatin interactions at a high sensitivity while controlling false discovery rate (FDR) at a reasonable level. The input of MICC is raw PET clusters derived from ChIA-PET data. The final output of MICC includes: (i) a list of posterior probabilities that describe the PET clusters as true interaction clusters and (ii) the corresponding FDR. MICC can always detect significantly more interactions than ChIA-PET tool and ChiaSig at the same FDR on different datasets. The interactions detected by MICC are also more consistent between biological replicates.

2 Methods

Detailed description of MICC method could be found in supplementary methods. Here we briefly describe the principles. We first used self-ligation PETs to call protein binding peaks and set them as anchor regions. Then inter-ligation PETs linking anchor regions were grouped as PET clusters. For the sake of simplicity, from here on, the phrase PET mentioned later is only referred to inter-ligation PET. To infer a PET cluster (A, B), where A and B are two anchor regions linked by at least one PET, to be a *T*rue interaction *P*ET

cluster (TiPC), a Random collision PET cluster (RcPC) or a Random ligation PET cluster (RlPC), we used three types of features: (i) PET-count c_{AB} between anchor regions A and B, (ii) total PET-count c_A (c_B) in anchor region A (B) and (iii) genomic distance d_{AB} between anchor regions A and B ($d_{AB} = +\infty$ if A and B are in different chromosomes). If (A, B) is an RlPC, c_{AB} is modeled to follow a hyper-geometric distribution (Li *et al.*, 2010). If (A, B) is a TiPC or RcPC, we modeled it as a discrete Pareto distribution, i.e. Zeta distribution (Jessen and Winter, 1935) since c_{AB} follows power-law when c_{AB} is sufficiently large ($c_{AB} \geq 3$) (Supplementary Fig. S1). The parameters of the Zeta distributions depend on d_{AB} (Supplementary Fig. S2) and we set it as a quadratic fractional function. It is noticed that $log(c_A c_B)$ is significantly larger for reproducible PET 3+ clusters between replicates than that of those non-reproducible ones (Supplementary Fig. S3). Thus we used c_A and c_B as features to estimate the prior probability of an RcPC. The feature d_{AB} is then used to describe the prior probability for observing an RlPC to filter out random ligation noise (Supplementary Fig. S4). The full model consists of three components, each of which is the conditional probability distribution of PET-count for TiPC, RcPC and RlPC, respectively. The prior probability and parameters for each component can be described by total PET-count in anchor regions and the genomic distance between two anchors.

3 Application examples and comparison with previous methods

We applied MICC on K562 Pol2 ChIA-PET data (Li et al., 2012). First, we checked the performance to recover interactions detected in higher-depth sequencing libraries from lower-depth sequencing libraries between MICC and ChIA-PET tool (ChiaSig was not included in this comparison as it detected much less interactions). The lower-depth data were selected by randomly sampling 50% PETs from each replicate for 100 times. For each higher-depth data, interactions identified by both MICC and ChIA-PET tool were defined as the total interaction set. As is shown in Figure 1A and Supplementary Figure S5, top-ranked predictions by ChIA-PET tool and MICC recovered similar amount of high-confidence interactions, but MICC detected more true positives from weaker signals. This suggests that MICC can give a more consistent performance between lower-depth and higher-depth sequencing libraries.

Next, the reproducibility between biological replicates was compared among MICC, ChIA-PET tool and ChiaSig. We evaluated reproducibility by overlapping top-ranked interactions from two replicates for these methods. Inter-chromosomal PET clusters were removed at first since ChiaSig could not deal with them. Again, MICC shows the best performance, while reproducibility between two replicates decreases very quickly for ChiaSig (Fig. 1B). These observations suggest that MICC can remove ChIA-PET noises in a more consistent way, thus improve the reproducibility between biological replicates.

Lastly, we made a further comparison between ChiaSig and MICC by overlapping with 5C results, since ChiaSig paper (Paulsen *et al.*, 2015) showed that the method gives more precise results than ChIA-PET tool by comparing with 5C data (Amartya *et al.*, 2012). Here PET clusters were derived from the original ChiaSig paper, which mixed two replicates of K562 Pol2 ChIA-PET data. For both methods, we used FDR 0.05 to call significant interactions. Among 267 MICC significant interactions that overlap with 5C anchors at both ends, 53 can be validated by 5C significant interactions. For ChiaSig, there is only 9 interactions can be validated by 5C while the number of ChiaSig significant interactions that overlap with 5C anchor regions at both ends is 41. The fraction of 5C validated interactions is very similar between the two methods (P-value = 0.834), but MICC can call significantly more interactions (P-value = 2.82e − 10) (Fig. 1C). Furthermore, we checked the significance of MICC called PET 2-clusters (PET clusters with one or two PETs) that overlap with 5C significant interactions. There are 24 MICC called significant interactions with PET-count less than 3 that can be validated by 5C data. This number is significantly higher than that of the randomly sampled PET 2-clusters (P-value = 0.002, Supplementary Fig. S6). It suggests that many of MICC detected weaker interactions are likely true interactions.

Comparisons on MCF7 ER ChIA-PET data (Fullwood *et al.*, 2009) also showed MICC gave the best performance. (Supplementary Fig. S7, S8).

4 Discussion

We proposed a new method, MICC, to call significant chromatin interactions from ChIA-PET data. Compared with ChIA-PET tool, MICC recovered a significantly greater fraction of interactions detected in higher-depth sequencing library using a much lower-depth sequencing library. It also gives a more consistent ranking for the PET clusters, thus can improve the reproducibility between experimental replicates. By comparing with 5C data, we showed that MICC can detect significantly more validated interactions than ChiaSig. Besides, the interactions with low PET-count detected by

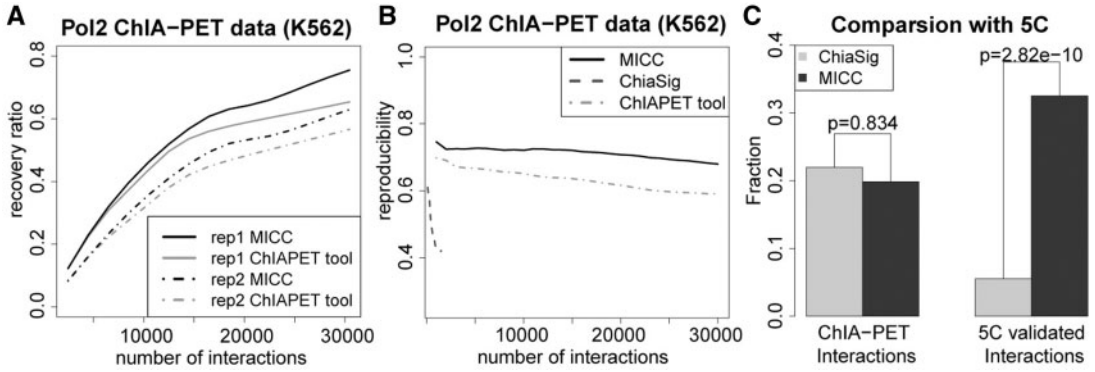

Fig. 1. (**A**) Average fraction of interactions in two original sequencing libraries from lower-sampled libraries (average of 100 times). (**B**) Fraction of interactions overlapped between top-ranked interactions from two Pol2 ChIA-PET replicates detected by ChIA-PET tool, ChiaSig and MICC, respectively. (**C**) Fraction of ChIA-PET interactions validated by 5C (left), and fraction of total 5C validated ChIA-PET interactions that are predicted by either computational methods (right)

MICC have a significant fraction of overlapping with 5C data, suggesting MICC is feasible to search for weak interactions. These features make MICC superior over other existing tools especially when processing ChIA-PET data with less sequencing depth.

Funding

National Basic Research Program of China (Grant Number 2012316503); National Natural Science Foundation of China (Grant Number 91019016, 31371341, 31361163004), and Tsinghua University Initiative Scientific Research Program.

Conflict of Interest: none declared.

References

Amartya,S. *et al*. (2012) The long-range interaction landscape of gene promoters. *Nature*, **489**, 109–113.

Fullwood,M. *et al*. (2009) An oestrogen-receptor-α-bound human chromatin interactome. *Nature*, **462**, 58–64.

Jessen,B. and Winter,A. (1935) Distribution functions and the Riemann zeta function. *Trans. Am. Math. Soc.*, **38**, 48–88.

Li,G. *et al*. (2010) ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing. *Genome Biol.*, **11**, 1–13.

Li,G. *et al*. (2012) Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell*, **148**, 84–98.

Paulsen,J. *et al*. (2015) A statistical model of ChIA-PET data for accurate detection of chromatin 3D interactions. *Nucleic Acids Res.*, **42**, e143.