Supplemental Material

This Supplemental Material provides additional details for the methods MICC.

Contents

Supplemental Methods:

1. Raw ChIA-PET sequencing data pre-processing

The linker filtered PETs (Li, *et al*., 2010) were mapped to genome by Bowtie (Langmead, *et al*., 2009). Two PET ends were mapped separately and then paired together. The duplicated PETs were merged to reduce PCR bias. To determine genomic distance threshold between self-ligation PETs and intra-chromosomal inter-ligation PETs, we computed the PET-count ratio between two types of PETs: one with two PET ends mapped to different strand of the genome (denoted as dsPETs), and the other with two PET ends mapped to the same strand (denoted as ssPETs). Intra-chromosomal inter-ligation PETs should be equiprobable to be dsPETs or ssPETs, while all self-ligation PETs must belong to dsPETs. Thus PET-count ratio between dsPETs and ssPETs would decrease along with the span of PETs and become constant once most of the PETs consist of intra-chromosomal inter-ligation PETs. Therefore, we classified dsPETs with genomic distance less than the distance threshold as self-ligation PETs and all PETs with distance larger than the distance threshold as intra-chromosomal inter-ligation PETs. After PETs classification, we used MACS (Zhang, *et al*., 2008) to call protein binding peaks from self-ligation PETs. These peaks were used as anchor regions to call chromatin interactions. The total PETs that linked two different anchor regions were defined as a PET cluster. It should be noticed that we did not incorporate these raw data pre-processing steps into the MICC R package, since users might prefer a different mapping tool or peak calling tool or even a different way to cluster the raw PETs. These alternatives do not affect the use of MICC R package. Thus instead, the programs for raw data pre-processing were provided as independent Perl scripts.

2. A three-component mixture model to call chromatin interactions.

We use a three-component mixture model to describe conditional distribution of PET-count from all the PET clusters. One component represents true interaction PET cluster (TiPC), and the other two for random collision PET cluster (RcPC) and random ligation PET cluster (RlPC), respectively. Let c_{AB} denote the PET-count between two anchor regions for PET cluster (A, B), and $c_A(c_B)$ be the total PET-count in anchor region A (B), and $\mathrm{d_{AB}}$ be the genomic distance between anchor A and B ($d_{AB} = +\infty$ if A and B are in two different chromosomes). Let $I_{AB} = 1$ denote (A, B) as a TiPC, $I_{AB} = 2$ as an RcPC and $I_{AB} = 3$ as an RlPC. Then the full model can be described as:

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

The prior probability for each component is not constant but depend on total PET-count c_A, c_B and distance d_{AB}. For simplicity, we can remove some of the dependencies, such that

$$P(c_{AB}|c_A, c_B, d_{AB}, I_{AB} = 1) = P(c_{AB}|d_{AB}, I_{AB} = 1) \quad (2\text{-}2)$$

$$P(c_{AB}|c_A, c_B, d_{AB}, I_{AB} = 2) = P(c_{AB}|d_{AB}, I_{AB} = 2) \quad (2\text{-}3)$$

$$P(c_{AB}|c_A, c_B, d_{AB}, I_{AB} = 3) = P(c_{AB}|c_A, c_B, I_{AB} = 3) \quad (2\text{-}4)$$

$$P(I_{AB} = 3|c_A, c_B, d_{AB}) = P(I_{AB} = 3|d_{AB}) \quad (2\text{-}5)$$

And the other two priors can be spitted as

$$\begin{aligned} P(I_{AB}=1|c_A,c_B,d_{AB}) &= P(I_{AB}=1|c_A,c_B,d_{AB},I_{AB}\neq 3)P(I_{AB}\neq 3|c_A,c_B,d_{AB}) \\ &= P(I_{AB}=1|c_A,c_B,I_{AB}\neq 3)\big(1-P(I_{AB}=3|c_A,c_B,d_{AB})\big) \\ &= P(I_{AB}=1|c_A,c_B,I_{AB}\neq 3)\big(1-P(I_{AB}=3|d_{AB})\big) \end{aligned} \quad (2\text{-}6)$$

$$\begin{aligned} P(I_{AB}=2|c_A,c_B,d_{AB}) &= P(I_{AB}=2|c_A,c_B,d_{AB},I_{AB}\neq 3)P(I_{AB}\neq 3|c_A,c_B,d_{AB}) \\ &= P(I_{AB}=2|c_A,c_B,I_{AB}\neq 3)\big(1-P(I_{AB}=3|c_A,c_B,d_{AB})\big) \\ &= P(I_{AB}=2|c_A,c_B,I_{AB}\neq 3)\big(1-P(I_{AB}=3|d_{AB})\big) \end{aligned} \quad (2\text{-}7)$$

The functional form for each item will be discussed in the following sections.

2.1 PET-count distribution for PET clusters derived from true interactions, random collision and random ligation.

We take the probability function of each component as the following distributions:

$$P(c_{AB}|I_{AB}=1,d_{AB}) = \frac{1}{\zeta\big(\theta_1(d_{AB})\big)(c_{AB})^{\theta_1(d_{AB})}} \quad (2\text{-}8)$$

$$P(c_{AB}|I_{AB}=2,d_{AB}) = \frac{1}{\zeta(\theta_2(d_{AB}))(c_{AB})^{\theta_2(d_{AB})}} \quad (2\text{-}9)$$

$$P(c_{AB}|I_{AB}=3,c_A,c_B) = \frac{dhyper(c_{AB},c_A,2N-c_A,c_B)}{1-dhyper(0,\ c_A,2N-c_A,c_B)} \quad (2\text{-}10)$$

where $\zeta(\theta)=\sum_{k=1}^{+\infty}1/k^{\theta}$ is Remann Zeta function (Jessen and Winter, 1935), and $dhyper(c_{AB},c_A,2N-c_A,c_B)=\frac{\binom{c_A}{c_{AB}}\binom{2N-c_A}{c_B-c_{AB}}}{\binom{2N}{c_B}}$ is probability mass function of hyper-geometric distribution and N is the sum of PET-count from all included PET clusters.

We use hyper-geometric distribution to describe count distribution of PETs derived from random ligations, as suggested by ChIA-PET tool (Li, *et al*., 2010). Since we could only observe the PET clusters with $c_{AB}>0$, the hyper-geometric distribution is zero-truncated. As the random ligation events happen in solution, they should be independent with the genomic distance between two anchors. For TiPC and RcPC, we find number of non-chimeric PET clusters that have PET-count c_{AB} (denoted as $n_{c_{AB}}$) is log-log linearly correlated with c_{AB} when c_{AB} is sufficiently large ($c_{AB}\geq 3$ for non-chimeric PET clusters, Figure S1), which follows power-law. Thus we model each of them as a discrete Pareto distribution (Newman, 2005), i.e., Zeta distribution, respectively. The parameters of these Zeta distributions, θ_1 and θ_2, should be related with genomic distance d_{AB}. Their functional forms will be discussed later.

2.2 Prior probability of random ligation PET clusters

As we can observe in Figure S4, log10-distance distribution of intra-chromosomal chimeric PETs, which is definitely random ligation products, has only one peak at a genomic distance larger than 1Mbps. In the meanwhile, log10-distance distribution of intra-chromosomal non-chimeric PETs presents a bimodal distribution, with the first peak located at ~50kbps and second peak at the position very similar to that of chimeric PETs. It indicates that random ligation events can be well separated by genomic distance. Thus we give a distance related prior to describe probability to observe a PET cluster with specific genomic distance span, i.e.,

$$P(I_{AB}=3|d_{AB}) = \lambda_0\frac{1}{1+e^{b_1 log(d_{AB})+b_2}}, 0<\lambda_0<1 \quad (2\text{-}11)$$

Denote

$$\lambda(d_{AB}) = P(I_{AB}=3|d_{AB}) \quad (2\text{-}12)$$

Then we have

$$P(I_{AB}\neq 3|d_{AB}) = 1-\lambda(d_{AB}) \quad (2\text{-}13)$$

For inter-chromosomal PET clusters, d_{AB} is set to be $+\infty$ and $\lambda(+\infty)=\lambda_0$

2.3 Prior probability of random collision PET clusters

It is supposed that RcPC should be less likely to reproduce between two experimental replicates than

TiPC. Thus we take a look at $log(c_A c_B)$ for two groups of PET 3+ clusters: one is shared between two replicates while the other is not. Taking PET 3+ clusters is to make sure that most of these PET clusters are not RlPCs. As is shown in Figure S3, $log(c_A c_B)$ is significantly larger for shared PET 3+ clusters. Thus total PET-count in anchor regions can act as an important feature to discriminate TiPCs and RcPCs. We incorporate it into prior probability to describe random collision events. More specifically, we set the prior probability of random collision PET clusters conditioned on $I_{AB} \neq 3$ as

$$P(I_{AB} = 2|c_A, c_B, I_{AB} \neq 3) = \frac{1 + e^{c_1 log(c_A)+c_2} + e^{c_1 log(c_B)+c_2}}{(1 + e^{c_1 log(c_A)+c_2})(1 + e^{c_1 log(c_B)+c_2})} \quad (2\text{-}14)$$

Denote

$$\mu(c_A, c_B) = P(I_{AB} = 2|c_A, c_B, I_{AB} \neq 3) \quad (2\text{-}15)$$

Then we have

$$P(I_{AB} = 1|c_A, c_B, I_{AB} \neq 3) = 1 - \mu(c_A, c_B) \quad (2\text{-}16)$$

2.4 Functional form for parameters of the Zeta distributions

We have noticed that most of the PET clusters with d_{AB} less than 1Mbps are not RlPCs as the analysis in section 1.2. Among them, most of the PET clusters with only one PET are probably RcPCs. Thus the fraction of PET 1 clusters in all PET clusters at specific distance d_{AB} ($d_{AB} < 1$ Mbps) is approximately equal to $\frac{1}{\zeta(\theta_2(d_{AB}))}$, which is positively correlated with $\theta_2(d_{AB})$. Therefore, as the observation in Figure S4, we try to set a function $\theta_2(d_{AB})$ to satisfy the following four conditions:

- $\theta_2(d_{AB})$ is bounded when d_{AB} goes to $+\infty$
- $\theta_2(d_{AB})$ goes to infinity when d_{AB} goes to 0
- $\theta_2(d_{AB})$ is first decreasing and then increasing
- $\theta_2(d_{AB})$ is approximately linear when d_{AB} is not large

The quadratic fractional function is one of the simplest functions that satisfy these four conditions simultaneously. There is an assumption that the average interaction frequency of TiPC at specific d_{AB} should always be higher than that of RcPC. It results that $\theta_1(d_{AB}) < \theta_2(d_{AB})$, i.e., $\theta_1(d_{AB}) + \theta_0 = \theta_2(d_{AB}), \theta_0 > 0$. For the sake of simplicity, θ_0 is set to be a constant independent with d_{AB}.

Hence we can set

$$\theta_1(d_{AB}) = \frac{a_1 d_{AB} + a_2 a_3}{d_{AB} + a_2} + \frac{a_4}{d_{AB}},\ a_1 > 1, a_2, a_3, a_4 > 0 \quad (2\text{-}17)$$

And $\theta_2(d_{AB})$ as the same shape with $\theta_1(d_{AB})$, but larger than it at anywhere.

$$\theta_2(d_{AB}) = \theta_1(d_{AB}) + \theta_0,\ \theta_0 > 0 \quad (2\text{-}18)$$

2.5 Full model to call true interaction clusters

For a specific PET cluster (A, B), we have

$$\begin{aligned}
&P(c_{AB}|c_A, c_B, d_{AB}) \\
&= P(c_{AB}|d_{AB}, I_{AB} = 1)P(I_{AB} = 1|c_A, c_B, d_{AB}) \\
&\quad + P(c_{AB}|d_{AB}, I_{AB} = 2)P(I_{AB} = 2|c_A, c_B, d_{AB}) \\
&\quad + P(c_{AB}|c_A, c_B, I_{AB} = 3)P(I_{AB} = 3|d_{AB}) \\
&= P(c_{AB}|d_{AB}, I_{AB} = 1)P(I_{AB} = 1|c_A, c_B, I_{AB} \neq 3)\big(1 - P(I_{AB} = 3|d_{AB})\big) \text{ by (2-6)} \\
&\quad + P(c_{AB}|d_{AB}, I_{AB} = 2)P(I_{AB} = 2|c_A, c_B, I_{AB} \neq 3)\big(1 - P(I_{AB} = 3|d_{AB})\big) \text{ by (2-7)} \\
&\quad + P(c_{AB}|c_A, c_B, I_{AB} = 3)\lambda(d_{AB}) \text{ by (2-12)} \\
&= P(c_{AB}|d_{AB}, I_{AB} = 1)\big(1 - \mu(c_A, c_B)\big)(1 - \lambda(d_{AB})) \\
&\quad + P(c_{AB}|d_{AB}, I_{AB} = 2)\mu(c_A, c_B)(1 - \lambda(d_{AB})) \\
&\quad + P(c_{AB}|c_A, c_B, I_{AB} = 3)\lambda(d_{AB})
\end{aligned} \quad (2\text{-}19)$$

Each item above can be substituted with a specific probability distribution function described in section 1.2-1.4

Parameters of the model are fitted by EM algorithm (Dempster, *et al*., 1977) on all PET clusters.

3. FDR estimation

We randomly generate PET-count c_{AB} and labels of PET cluster (A, B), i.e., I_{AB}, from the trained

model. Denote $Post^{(n)}$ and $rndPost^{(n)}$ as the posterior probability of the original dataset and randomly generated dataset, respectively. Then the FDR can be estimated as

$$FDR(p) = \frac{\#\{rndPost^{(n)} > p \ \& \ I_{AB} \neq 1\}}{\#\{rndPost^{(n)} > p\}} \quad (3\text{-}1)$$

Where #{} is a counting function.

4. Real data implementation

For ER ChIA-PET data (MCF7 cell), MICC is applied on the PET clusters derived from the original ChIA-PET paper (Fullwood, *et al*., 2009). This makes the comparison with ChIA-PET tool as objective as possible. The two higher-depth sequencing libraries are IHM001F and IHH015F, and the lower-depth sequencing library is IHH015M.

For Pol2 ChIA-PET data (K562 cell), MICC is applied on the PET clusters derived from our processing steps since ChiaSig and ChIA-PET tool use different methods to cluster the PETs.

Supplemental Figures:

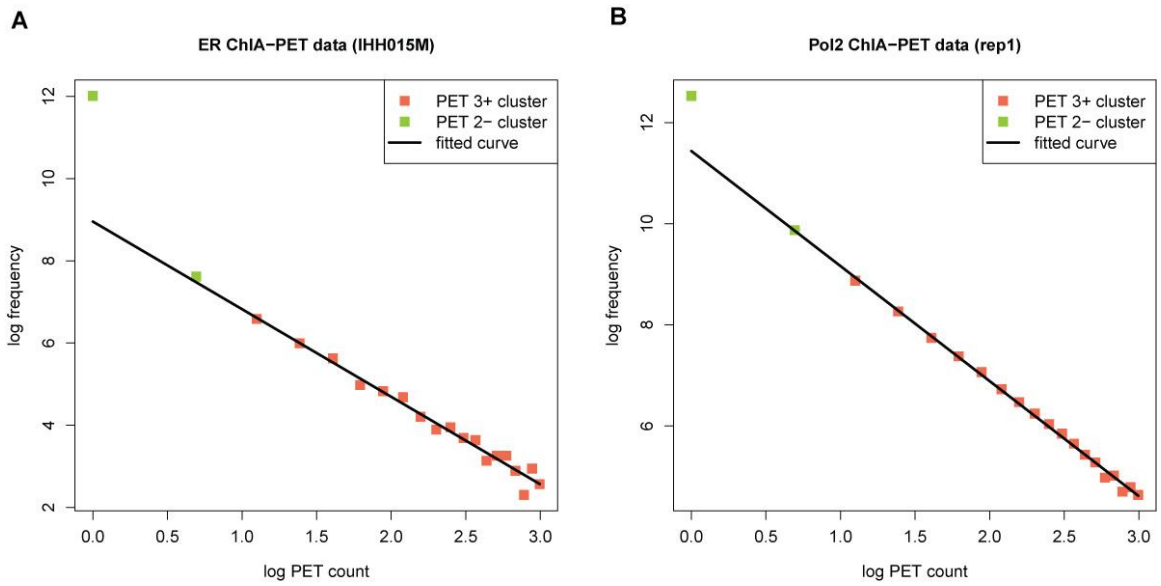

Fig. S1. PET-count c_{AB} and number of PET clusters $n_{c_{AB}}$ is log-log linearly correlated when $c_{AB} > 2$. **(A)** IHH015M library of ER ChIA-PET data. **(B)** Replicate 1 library of Pol2 ChIA-PET data.

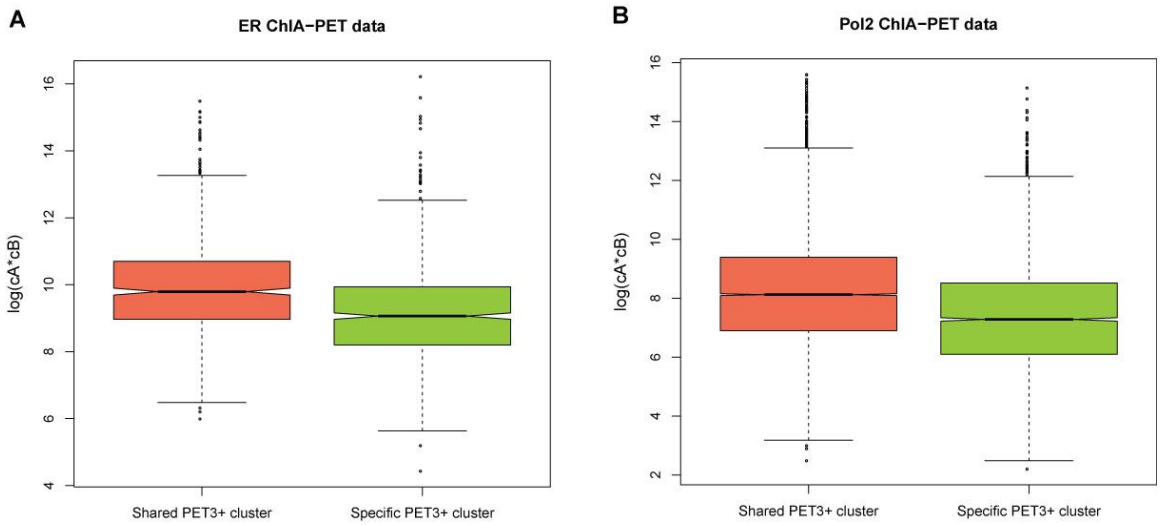

Fig. S2. Boxplot of $log(c_A c_B)$ for shared PET3+ clusters and specific PET3+ clusters between two replicates. **(A)** Shared PET3+ clusters are defined as PET3+ clusters from IHM001F library that overlap with PET3+ clusters from IHH015F library, and specific PET3+ clusters are defined as PET3+ clusters only present in IHM001F library. The p-value for this comparison is 2.91e-27, given by Wilcox test. The data is MCF7 ER ChIA-PET data. **(B)** Shared PET3+ clusters are defined as PET3+ clusters from Replicate 1 that overlap with PET3+ clusters from Replicated 2, and specific PET3+ clusters are defined as PET3+ clusters only present in Replicate 1. The p-value for this comparison is 1.07e-151, given by Wilcox test. The data is K562 Pol2 ChIA-PET data.

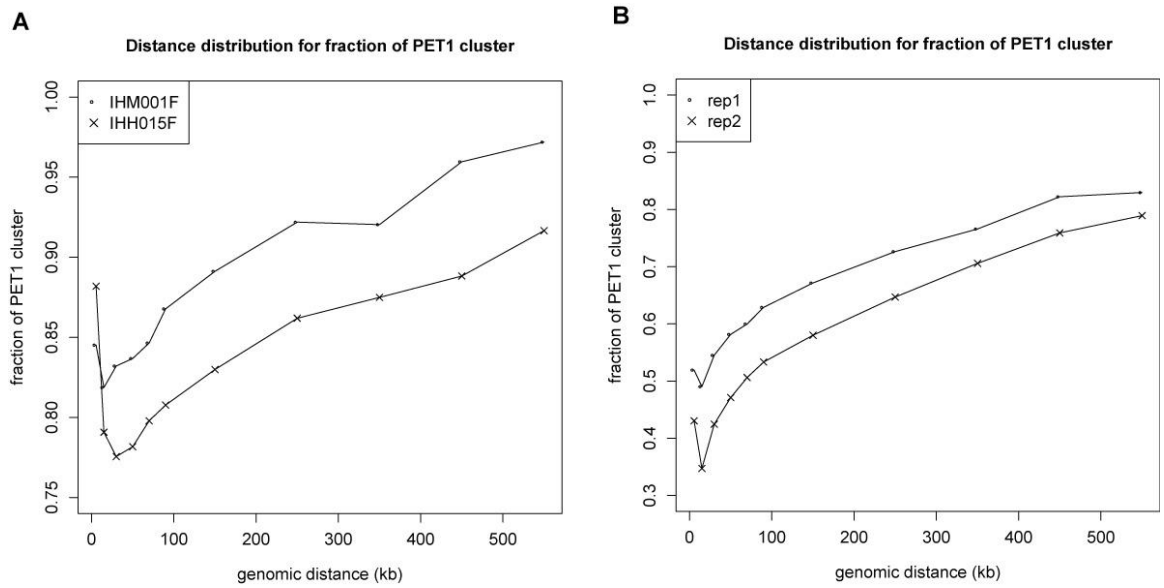

Fig. S3. Fraction of PET 1 clusters to all PET clusters at specific distance. **(A)** The two datasets are IHH001F and IHH015F libraries of ER ChIA-PET data (MCF7 cell). **(B)** The two datasets are two replicates of Pol2 ChIA-PET data (K562 cell).

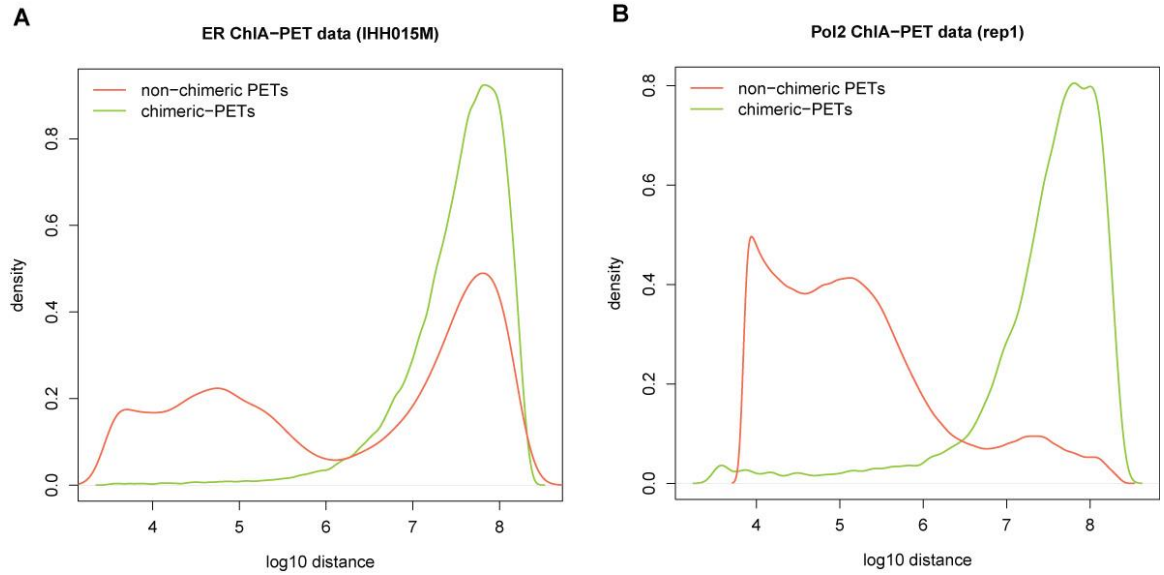

Fig. S4. PET span distribution of intra-chromosomal non-chimeric and chimeric PETs. **(A)** IHH015M library of ER ChIA-PET data (MCF7 cell). **(B)** Replicate 1 library of Pol2 ChIA-PET data (K562 cell).

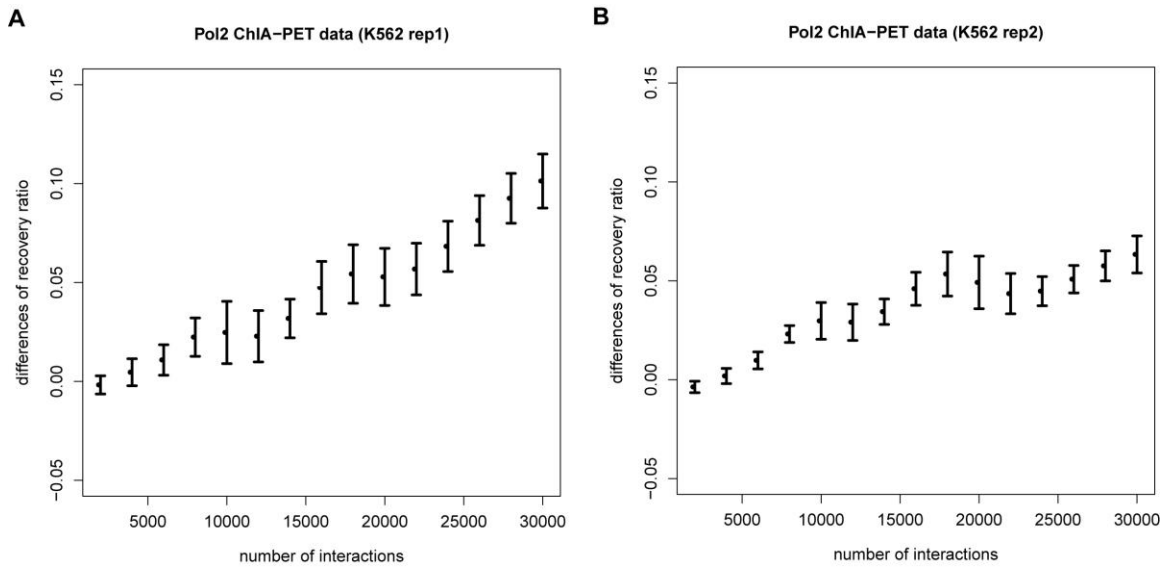

Fig. S5. Differences of fraction of interactions in higher-sequenced libraries recovered from lower-sequenced libraries between MICC and ChIA-PET tool ($Recovery\_Ratio_{MICC}$ - $Recovery\_Ratio_{ChIA\text{-}PETtool}$). The lower-sequenced libraries were selected by randomly sampling 50% PETs from each replicate for 100 times. Error bar in the figure marks the standard deviation of 100 times sampling for K562 Pol2 ChIA-PET data replicate 1 **(A)** and replicate 2 **(B)**. These results show that MICC can give more consistent predictions between lower-depth and higher-depth sequencing libraries.

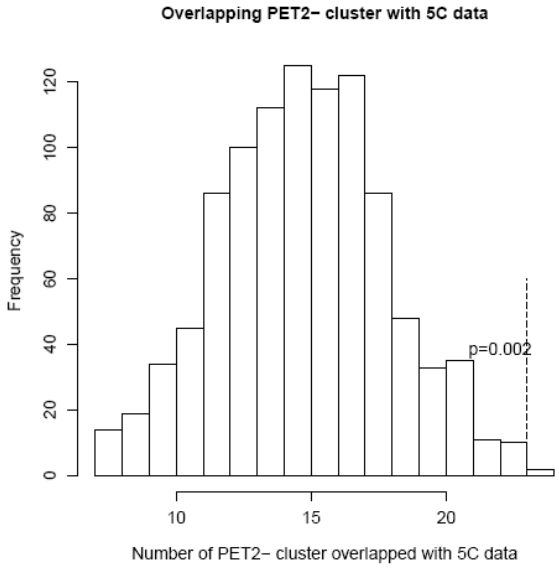

Fig. S6. Distribution for number of randomly sampled PET 2- clusters that overlap with 5C significant interactions. Dashed line marks the number of 5C validated interactions called by MICC.

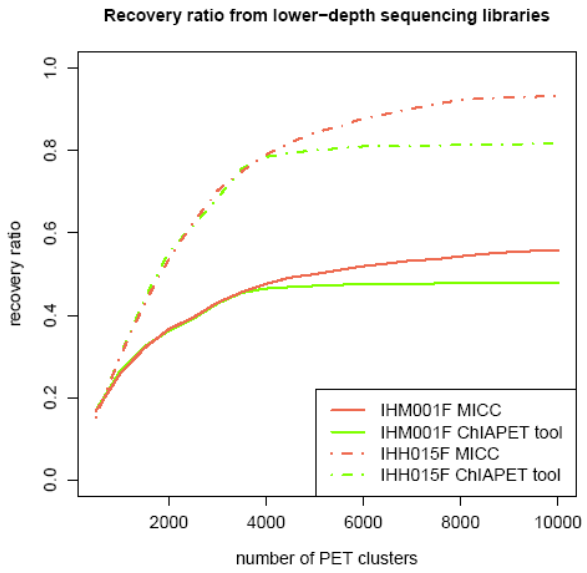

Fig. S7. Fraction of interactions in two higher-depth sequencing libraries (IHM001F and IHH015F) recovered from lower-depth sequencing library IHH015M. The data is ER ChIA-PET data (MCF7 cell).

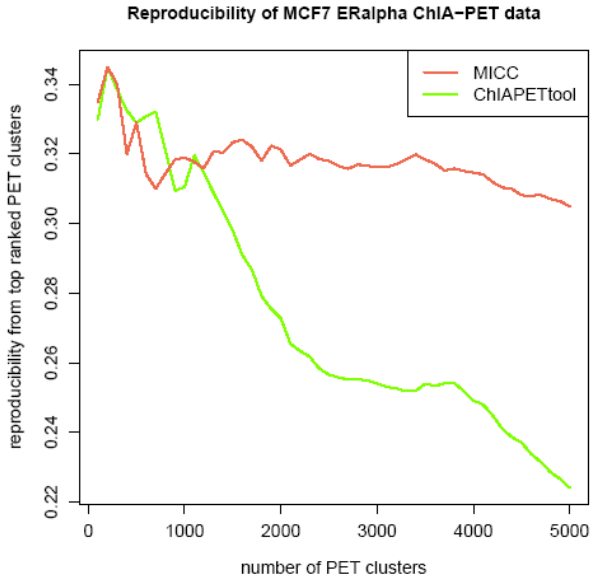

Fig. S8. Fraction of interactions overlapped between top-ranked interactions from two ER ChIA-PET replicates detected by ChIA-PET tool and MICC, respectively. The number of ChiaSig detected interactions is very small, thus it is not used for this comparison.

Supplemental Tables:

Library Name	Number of PET clusters	Time cost
K562 Pol2 ChIA-PET rep1	130,973	44m
K562 Pol2 ChIA-PET rep2	135,214	52m
MCF7 ERalpha ChIA-PET IHM001F	1,097,288	6h 59m
MCF7 ERalpha ChIA-PET IHH015F	1,777,454	10h 56m
MCF7 ERalpha ChIA-PET IHH015M	169,067	69m

Table S1. Time cost of MICC for ChIA-PET data used in the paper. The system to run MICC is Linux 2.6.18-274.el5 and the CPU is AuthenticAMD with 2000 MHz**.**

Supplemental References:

Dempster, P. *et al*.. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal statistical Society*, **39**, 1–38.

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. *Transactions of the American Mathematical Society*, **38**, 48-88.

Langmead, B. *et al*.. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*. **10**, R25.

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

Newman, E.. (2005). Power laws, pareto distributions and zipf's law. *Contemporary physics*, **46**, 323–351.

Zhang, Y. *et al*.. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biology*. 9, R137.