Quantitative Biology 2015, 3(2): 81–89
DOI 10.1007/s40484-015-0047-9

RESEARCH ARTICLE

A novel method to identify topological domains using Hi-C data

Yang Wang[1,†], Yanjian Li[1,2,†], Juntao Gao[1,*] and Michael Q. Zhang[1,3,*]

[1] MOE Key Laboratory of Bioinformatics; Bioinformatics Division and Center for Synthetic & Systems Biology, TNLIST; Department of Automation, Tsinghua University, Beijing 100084, China
[2] Department of Basic Medical Sciences, School of Medicine, Tsinghua University, Beijing 100084, China
[3] Department of Molecular and Cell Biology, Center for Systems Biology, the University of Texas at Dallas, Richardson, TX 75080, USA
* Correspondence: jtgao@biomed.tsinghua.edu.cn, michaelzhang@tsinghua.edu.cn

Received April 10, 2015; Revised June 20, 2015; Accepted July 6, 2015

Over the last decade the 3C-based (Chromosome Conformation Capture, 3C) approaches have been developed to describe the frequency of chromatin interaction. The invention of Hi-C allows us to obtain genome-wide chromatin interaction map. However, it is challenging to develop efficient and robust analytical tools to interpret the Hi-C data. Here we present a new method called Clustering based Hi-C Domain Finder (CHDF), which is based on the difference of interaction intensity inside/outside domains, to identify Hi-C domains. We also compared CHDF with existing methods including Direction Index (DI) and HiCseg. CHDF can define more chromatin domains validated by higher resolution local chromatin structure data (Chromosome Conformation Capture Carbon Copy (5C) data). Using Hi-C data of lower sequencing depth, chromatin structure identified by CHDF is closer to that discovered by data of higher sequencing depth. Furthermore, the implement of CHDF is faster than the other two. Using CHDF, we are potentially able to discover more hints and clues about chromatin structural elements at domain level.

Keywords: chromatin domain; Hi-C; dynamic programming

INTRODUCTION

During interphase, chromosomes in eukaryotic cells do not mix randomly but occupy separate areas called chromosome territories [1]. Nucleus is spatially compartmentalized and gene expression is correlated with gene position in the nucleus [2]. Although advances in microscopes and imaging methods, such as super-resolution microscopy [3,4] and fluorescence *in situ* hybridization (FISH) [5,6] allow people to visualize targeted genes or chromosomes at increasing resolution through base pairing of nucleic acid probes, these methods do not offer positioning information of chromatins at genome-wide scale. Fortunately, advances of the 3C-based techniques have surpassed and complemented the imaging approach, especially demonstrating that the target genes and their regulatory elements are in close spatial proximity [2], and indicating that the topological associated chromatin domains contain clusters of genes that are co-regulated [7].

To study chromatin structure with 3C-based data, people developed different kinds of algorithms, such as domain calling method based Direction Index [8], HiCseg [9], three-dimensional reconstructions and molecular dynamic simulations [10–14], among which bottom-up restraint-based three-dimensional approaches proved to be useful for rather stable chromosomal domains, while top-down polymer-based biophysical models present the statistical organizational features of folding states of chromosomes, in order to rationalize the measured probabilities of 3D chromatin interactions.

Studies using these approaches have discovered that in metazoan genomes, chromatin itself is further packaged into ~5 Mb-sized compartment A and B and spatially segregated megabase-sized domains [2]. In *Saccharomyces cerevisiae*, the clustering of the rDNA locus on

† These authors contributed equally to this work

chromosome XII [15], the clustering of tRNAs [16,17] and telomeres, and the known rosette organization of centromeric regions as regions of early replication, where chromosome arms extend from a centromeric cluster near one spindle pole, were confirmed [12,18]. Many of these genomes are predicted to have fractal globule conformations [19,20]. For example, metazoan organism has a higher degree of organizational complexity. It was discovered that there are discrete megabase-sized domains [8,13], where gene regulatory elements such as promoters and enhancers can be brought in contact by chromatin loop. Direction Index (DI) method [8] and HiCseg method [9] offer such solutions to identify domains systematically in mammalian cells.

Here we present a new method, Clustering based Hi-C Domain Finder (CHDF), to analyze Hi-C data based on clustering. With this method we systematically identified domains using Hi-C data from three cell lines. We compared CHDF with existing methods such as DI and HiCseg. CHDF has four advantages: (i) CHDF model is based on the knowledge, which is easy to understand, that chromatin interaction domains are the regions which have higher intensity of chromatin contacts inside, lower outside. (ii) CHDF can define chromatin domains at smaller scales and provide a finer chromatin structure, which can be verified by other kinds of experiment datasets. (iii) The boundary regions of CHDF results are more enriched with CTCF binding sites and active histone modification marks which are highly characteristic with known topological associated domain boundaries. (iv) Last but not the least, the implement of CHDF is faster than the other two methods, especially for the Hi-C matrix of large dimension.

RESULTS

Application to synthetic data

To compare the performance of CHDF with DI and HiCseg, we first applied three methods to simulated Hi-C contacts matrices generated by ourselves. The domains called by DI method in IMR90 cell line were used to generate simulated matrices. For each domain called by DI, we calculated the means and variances of normalized Hi-C contacts intra-domain (refers region D_i in Figure 1A) and region between itself and its 5′ upstream domain (refers region A_i in Figure 1A). For each chromosome, the mean and variance of normalized Hi-C contacts in the regions of residual (refers region R in Figure 1A) were calculated as background. Next, we randomly selected 100 domains and one background and generated simulated Hi-C contact matrix by the means and variances using Gaussian distribution and maintained the size of each domain in DI results in IMR90 cell line. This step was repeated 50 times and 50 simulated Hi-C contact matrices were generated. CHDF, DI and HiCseg methods were applied to these 50 simulated Hi-C contact matrices. To evaluate the each result, we compared the domain borders (see the definition of borders in Supplementary Materials) called by each method to the real borders (designed in simulated matrices). The sensitivity (the ratio of the number of true positive borders versus the number of all real borders) and the false discovery rate (the ratio of the number of false borders versus the number of borders called by each method) were calculated for each result (Figure 1B). Obviously, it was CHDF method that owned

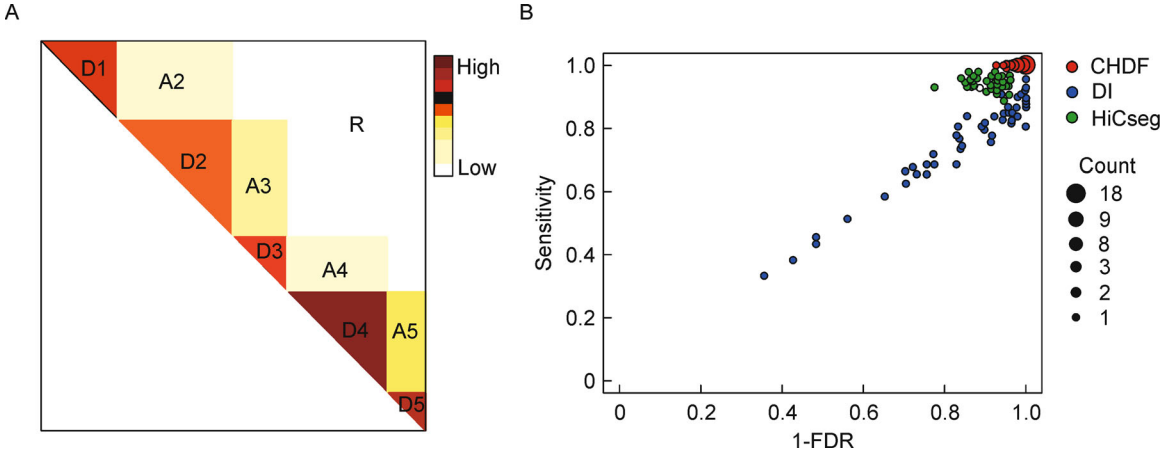

Figure 1. Schematic graph of CHDF model and results of simulation studies. (A) Examples of domains (D_i), adjacent regions (A_i) and residuals (R) were illustrated respectively. (B) The comparison of the sensitivity and false discovery rate (FDR) of three methods on simulated Hi-C data. For the overlapped points in the figure, the size of the points were used to illustrate the number of points.

the highest sensitivities and the lowest false discovery rate among the three methods, indicating that CHDF performed best on simulated data.

Application to real Hi-C data

We identified 5,715 domains with 6,067 borders in human IMR90 cell line (Figure S2A), and 5,326 domains with 5,504 borders in mouse ES cells (Figure S2B) using CHDF. We calculated the domain number distribution in human IMR90 cell line and mouse ES cells, using three different methods. The distribution of domain number in every chromosome of human IMR90 cell line (Figure S3A) and mouse ES cells (Figure S3B) was listed. The size distributions of domains in mouse ES cells were shown (Figure S4). We used several datasets (Table S1) to evaluate the results called by these three methods.

Smaller-scaled chromatin structure defined by CHDF can be validated by 5C data

5C [21] is another 3C-based method which provides the map of chromatin structure at higher resolution in specific chromatin regions. Phillips-Cremins and colleagues used 5C and high-throughput sequencing to map higher-order chromatins in six gene regions (*Oct4, Nanog, Nestin, Sox2, Klf4, Olig1-Olig2*) in mouse ES cells [22]. There are other two 5C experiments in mouse ES cells, designed in chromosome X (base-position from 98,831,149 to 103,404,509) [7] and *HoxD* cluster region [23]. In these gene regions, we employed the method to call domain using 5C data in Phillips-Cremins's paper to verify the domains called by Hi-C data using three methods.

We showed the heatmaps of Hi-C and 5C data, Hi-C domains called by three methods and 5C domains in *Klf4* and *Sox2* gene regions (Figure 2A). From the heatmap of Hi-C data in *Sox2* gene region, obviously the Hi-C interaction intensity was changed around base-position 34,520,000, where a border was identified by 5C data (the third border in *Sox2* region). CHDF identified a border at base-position 34,560,000, while HiCseg and DI method did not. In *Klf4* gene region, CHDF identified 3 borders and 2 of them can be aligned to the borders called by 5C data. The only border that cannot be aligned was too close to the beginning of *Klf4* gene region, which lacks upstream 5C data. In contrast, DI method identified just two borders of which only one border can be aligned, while HiCseg method identified only one border which cannot be aligned at all.

Next, we compared all the borders called by three methods to borders called by 5C data in these eight regions (Figures 2B and S5). CHDF can not only identify more borders but also increase the ratio of the borders which can be aligned to borders called by 5C data. The false discovery rate (the ratio of the number of the Hi-C domain borders which could not be aligned to 5C domain borders versus total number of Hi-C domain borders) of CHDF result was 0.4762, while 0.5625 for DI and 0.5833 for HiCseg. The false discovery rate of CHDF was slightly lower than that of the other two methods. The sensitivity (the ratio of the number of the Hi-C domain borders which could be aligned to 5C domain borders versus total number of 5C domain borders) of CHDF result was 0.3235, while 0.1029 for DI and 0.1471 for HiCseg. The sensitivity of CHDF was significantly higher than that of the other two methods (*P*-value 0.002932 to DI and *P*-value 0.02523 to HiCseg; Fisher's exact test). Comparing the Hi-C domains called by three methods to domains called by 5C data in these eight regions, we concluded that CHDF can identify more precise domains and provide chromatin structure at smaller size.

More CHDF domain borders can be validated by high-resolution Hi-C interaction map

Deeper sequencing in Hi-C experiment could probe the 3D architecture of genomes at higher resolution (i.e., smaller bin size). Rao et al. [13] constructed 10 kb resolution map in IMR90 cell line, providing finer chromatin structure than Dixon's Hi-C data, and partitioned the genome into contact domains (called arrowhead domains which have different definition from non-overlap domains called by three methods). Domains called by Hi-C contact matrix of small size bins will be small, which were more approximate to real domain-structure. Arrowhead domains were generated by Hi-C contact matrix with bin size of 10 kb. They also used much *priori* biological knowledge combined with their model to find domains and employed many other experiments to validate their domains. Based on these two reasons, we considered arrowhead domains could provide more precise chromatin structure than all domains called by Hi-C contact matrix with bin size of 40 kb. So arrowhead domains were employed to evaluated domains called by three methods in IMR90 cell line (Figure 2C).

Large parts of borders identified by three methods were aligned to the borders of arrowhead domains (Figure 2C). The false discovery rate (the ratio of borders which could not be aligned to arrowhead domain borders versus number of all borders called by each method) of CHDF result was 0.3737, while 0.4032 for DI and 0.3779 for HiCseg. The false discovery rate of CHDF was slightly lower than that of the other two methods. The sensitivity (the ratio of aligned borders which can be aligned to arrowhead domain borders versus number of all arrowhead domain borders) of CHDF result was 0.3298, while 0.1862 for DI and 0.2339 for HiCseg. The sensitivity of CHDF was significantly higher than that of the other two

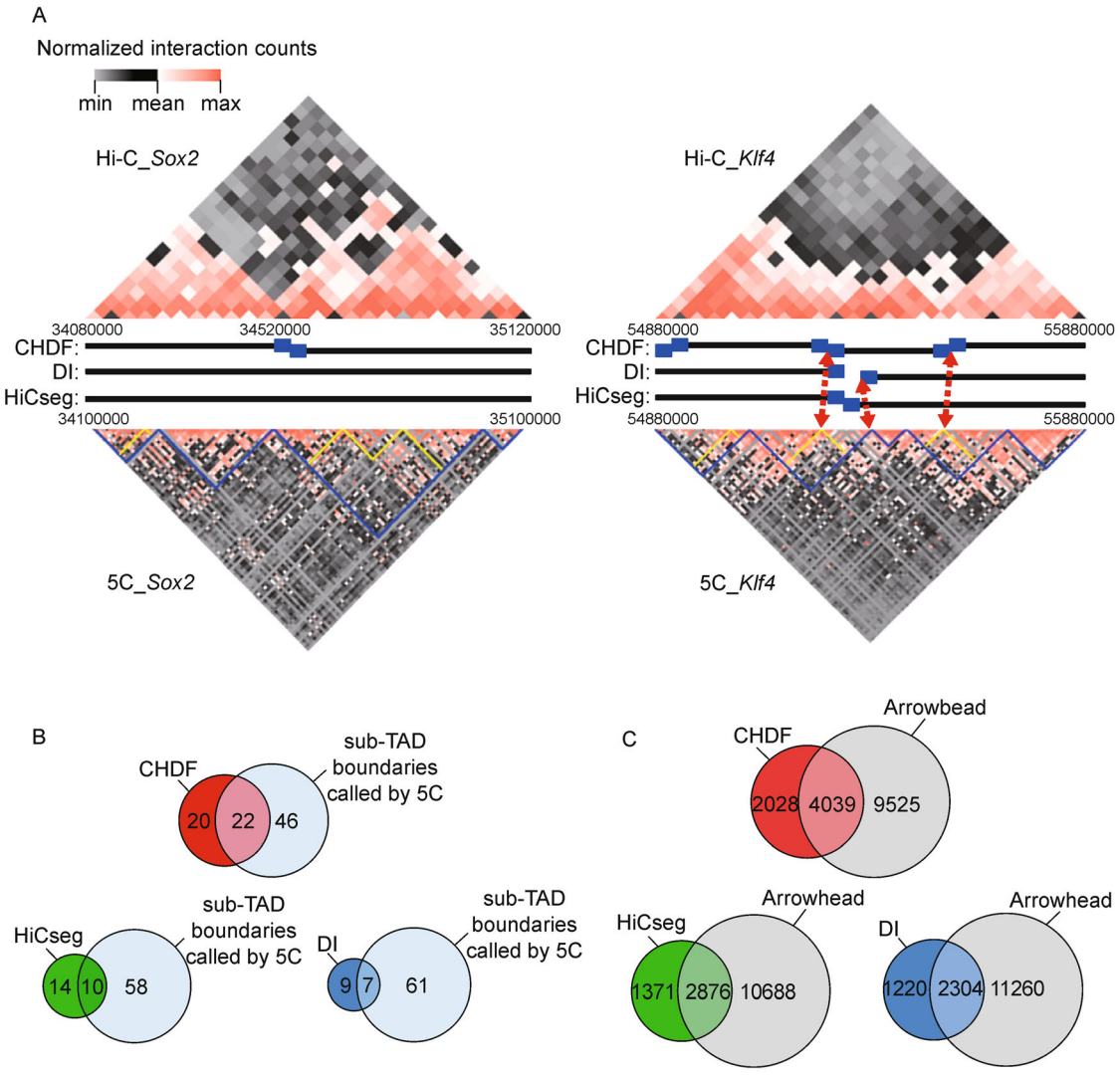

Figure 2. 5C data and arrowhead domains were used to validate the results called by three methods. (A) The boundaries and domains around gene *Sox2* and *Klf4* identified by CHDF method in mouse ES cell. Top: the chromatin interactions in Hi-C data after normalization. Middle: blue bars were the domain borders identified using three different methods (The width of border was the size of one bin, 40 kb). Bottom: the matrix of 5C interactions, with bin size of 10 kb. 5C domains were the lines marked in yellow and blue with different resolution. If a Hi-C border can be aligned to a 5C domain border, then a red dash line with arrows was used to connect these two borders. (B) The comparison of Hi-C domain borders called by three methods with borders called by 5C data in these eight genomic regions in mouse ES cells. Red: borders called by CHDF method. Blue: borders called by DI method, and green: borders called by HiCseg method. (C) The comparison of Hi-C domain borders called by three method with the borders of arrowhead domains (grey): CHDF method (red); HiCseg method (green); DI method (blue).

methods (P-value $< 2.2 \times 10^{-16}$ to DI and P-value $< 2.2 \times 10^{-16}$ to HiCseg; Fisher's exact test). Therefore, CHDF owns almost the same false positive rate, but higher sensitivity.

CHDF results were more coincident with arrowhead domains which provide chromatin structure at higher resolution, indicating that we could get more precise chromatin structure using Hi-C data matrix of limited resolution comparing with other two methods.

CHDF boundary regions enrich more CTCF binding sites, active histone marks and cis-eQTLs

The insulator binding protein CTCF is known to demarcate most boundaries between euchromatin and heterochromatin, thus important for the maintenance of

chromatin structure [8,24]. CTCF binding sites are enriched at the topological boundary regions (see the definition of boundary in Supplementary Materials) [8]. When comparing the results of every two methods, we called all regions within +/−20 kb of the boundaries. Among these regions, the overlapped regions between two results were removed, and the average peak number of CTCF binding site was calculated in the un-overlapped regions (Figure 3). Comparing to other two results, CHDF boundary regions were enriched more CTCF binding sites, especially in IMR90 cell (Figure 3).

Although most topological boundaries are enriched for the binding of CTCF, CTCF binding alone is insufficient to form domain boundaries. Active histone modification marks, combining with CTCF, demarcate domain boundaries [8]. Here we used four active histone modification marks (H3K4me1, H3K4me3, H3K27ac and H3K9ac) to compare the transcription activity in boundary regions

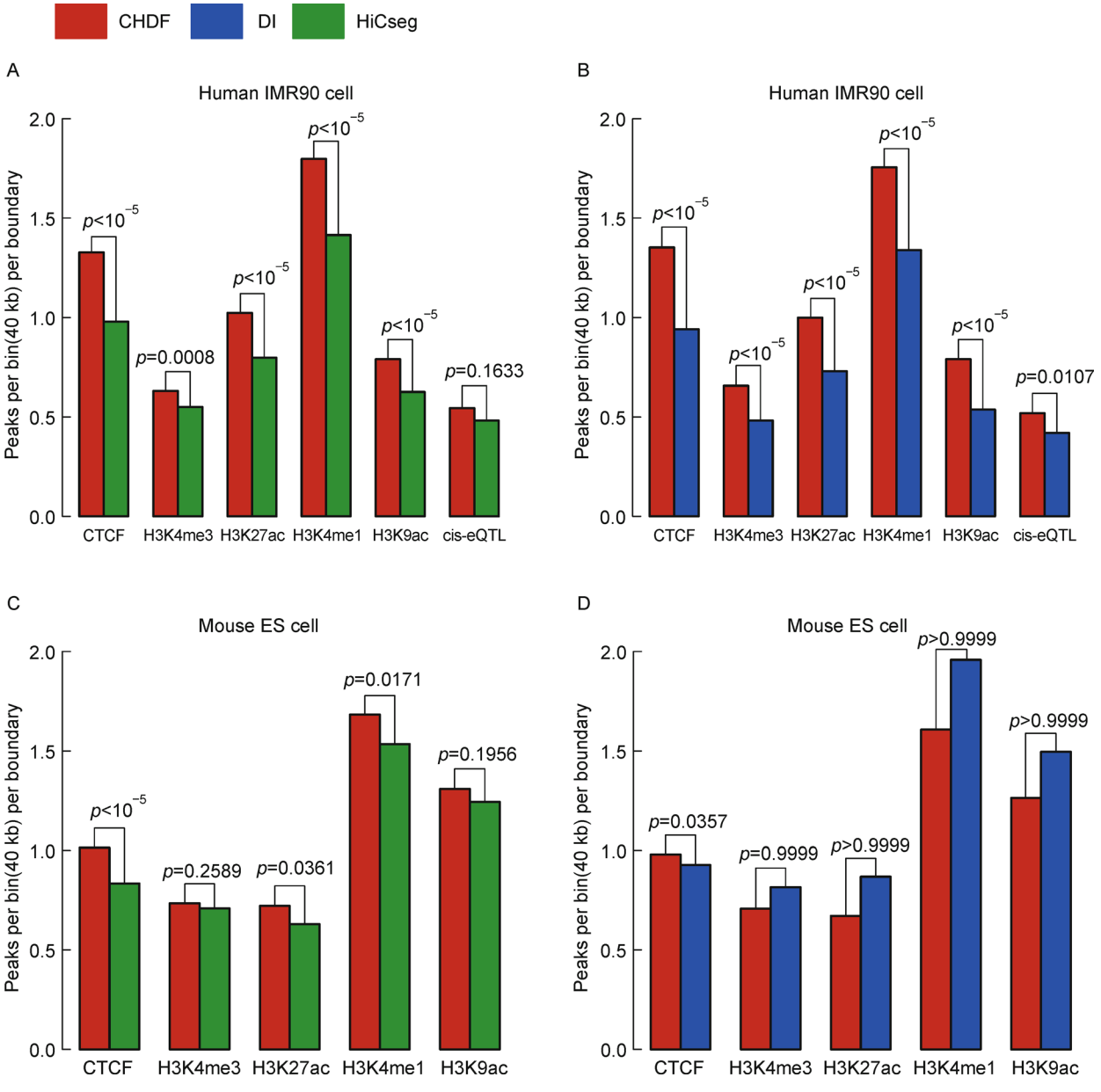

Figure 3. The average peak number of CTCF, active histone modification marks and number of cis-eQTLs at the different boundary regions. (A) Comparison of CHDF method with HiCseg method in human IMR90 cell line. (B) Comparison of CHDF method with DI method in human IMR90 cell line. (C) Comparison of CHDF method with HiCseg method in mouse ES cells. (D) Comparison of CHDF method with DI method in mouse ES cells.

identified by three methods (Figure 3). In IMR90 cell line, CHDF boundary regions were significantly enriched of more active histone modification marks (Figure 3A, 3B). In mouse ES cells, the number of average peaks of active marks in CHDF boundary regions was higher than that of HiCseg results, but lower than that of DI results (Figure 3C, 3D).

Cis-eQTLs are genomic loci that regulate expression levels of genomically approximate target gene. Latest study found that cis-eQTLs were genomically close to topological domain boundaries [25]. We collected cis-eQTLs from a database of eQTLs (eQTL Browser, https://eqtl.uchicago.edu) in six cell types (Table S1) [26–34]. Of these, we selected 24,757 cis-eQTLs that were at least 40 kb from the boundary of their associated genes because Hi-C data have an inherent resolution limit (40 kb in the data we used). In IMR90 cell line, the average number of cis-eQTLs was calculated in the un-overlapped regions (Figure 3A, 3B). CHDF boundary regions were enriched of more eQTLs.

Although the enrichments of CTCF binding sites, active histone modification marks or cis-eQTLs are not the criterion of topological domain boundaries, the regions with higher enrichments of these marks are more likely to be boundaries, indicating that the boundaries called by CHDF were more likely to be the real boundary regions comparing to boundaries called by the other two methods.

Comparison of robustness and speed of three methods

Since random ligation of DNA fragment in Hi-C experiment may generated much noise in the final data, the robustness of analysis method must be considered. First, to evaluate the consistency of three methods on biological replicate Hi-C datasets, we applied each method to two replicates and combined Hi-C data in IMR90 cell line. For each method, we named the domains called by these three datasets replicate1-domains, replicate2-domains and combined-domains respectively. The ratio of the number of same domains between replicate1-domains and replicate2-domains versus the mean number of replicate1-domains and replicate2-domains was 0.6668 in CHDF result, while 0.5915 for DI and 0.6515 for HiCseg. The ratio of the number of same domains among replicate1-domains, replicate2-domains and combined-domains versus the mean number of replicate1-domains, replicate2-domains and combined-domains was 0.5970 in CHDF result, while 0.3466 for DI and 0.4499 for HiCseg (Figures 4C and S6). Comparing the three methods using these two ratios, the domains called by CHDF using different replicate data variated least. We concluded that consistency of CHDF performed the best on the biological replicates among the three methods.

Next, to evaluate the robustness to technical noise, we compared these three methods by sub-sampling the original data. Here robustness was measured by two parameters P_1 and P_2:

$$P_1 = \text{the number of C} / \text{the number of B},$$

$$P_2 = \text{the number of C} / \text{the number of S},$$

where S is the domains identified in sub-sampled Hi-C data, B is the original domains, and C is the same domains between S and B.

The robustness to the technical noise of CHDF was better than that of DI method, though a little lower than that of HiCseg (Figure 4A and 4B).

As human genome size (3 billion base pairs) is larger than that of mouse (2.5 billion base pairs), we used human IMR90 cell line for speed comparison. In the optimal solution, the maximum size of domain was less than 200 bins. When implementing CHDF, we set a parameter to limit the maximum size of domain to find the optimal solution, which reduced the computational cost. When the maximum size of domain was set to 200 bins, the time cost of CHDF implement was the shortest among three methods (Figure 4D). The memory cost of CHDF was $O(n^2)$ (n refers to the dimension of Hi-C contact matrix), while $O(n^2)$ for HiCseg and $O(n)$ for DI. Since $O(n^2)$ memory was needed to read Hi-C contact matrix into memory, we considered that the memory cost of CHDF was acceptable.

DISCUSSION

Chromatin DNA is organized into hierarchical modules spatially, therefore it is of vital importance to identify modules or domains at genomic level. It is well accepted that chromatin domains are regions with higher contact intensity inside than outside [8]. Based on this knowledge, we developed a novel domain detection method and compared the domains called by this new method and by other two published methods in human IMR90 cell line and mouse ES cells. Independent biological judgment is needed to verify the domains identified. We used 5C data, CTCF binding sites, histone modification marks, cis-eQTLs and higher resolution map of Hi-C data to evaluate these three methods.

So far there is no gold standard to define the size of domains. Based on the public data, people tend to define compartment at mega-base size (about 5 Mb) while TAD at hundred kb (about 800 kb) [35]. Finer structure smaller than TAD is called sub-TAD. When people use Hi-C data to call domains, the size of the domain is largely dependent on the size of bins which is determined by

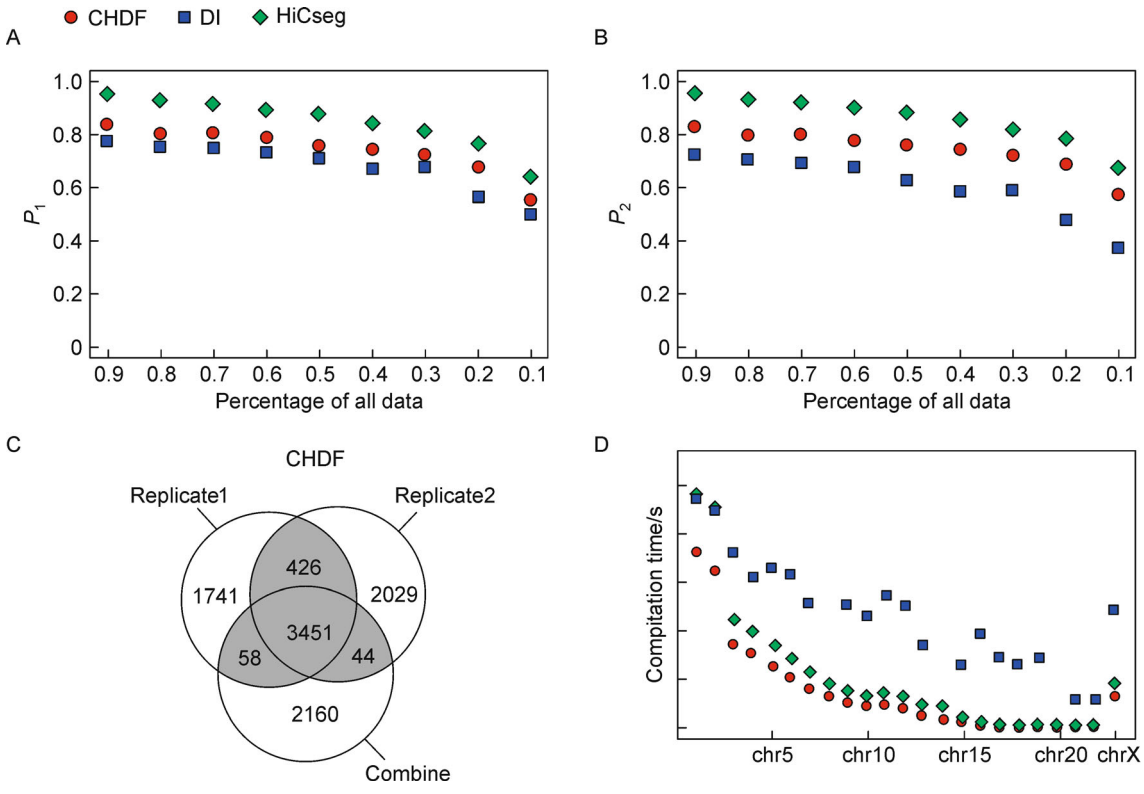

Figure 4. Robustness and time cost of the three methods. (A, B) The robustness to technical noise of three methods was compared by applying three methods to subsampled data in mouse ES cell. Here robustness was measured by two parameters P_1 (A) and P_2 (B). (C) Number of domains called by CHDF in two replicate and combined datasets. The intersection represented the same domains called in different Hi-C datasets. (D) The comparison of computational time cost of three methods in human IMR90 cell line. For the time cost comparison, a single 3.2 GHz CPU was used.

the number of Hi-C contact counts in the experiment. When the size of bins is chosen, the average size of Hi-C domains is also determined. Generally, restricted by sequencing depth, the size of bins ranges from 10 kb to 100 kb, which makes the size of Hi-C domains ranges from 0.5 Mb to 1 Mb. So Hi-C data can only define the structure at the level of TAD size range from 0.5 Mb to 1 Mb. Using the Hi-C contact matrix of same bin size, CHDF method can define chromatin domains at smaller scales which provided finer chromatin structure.

With the reduction of sequencing cost, Hi-C data become larger and larger and provide higher resolution chromatin contact map. When calling domains in Hi-C contact matrix of large dimension, time cost must be considered. Here we consider the maximum size of domains in the optimal solution of CHDF method Max bins. When dynamic programming is used to achieve the optimal solution of CHDF model, the cases that maximum domain size is larger than Max could not be the optimal solution and should not be considered. To reduce computation cost and save time, we defined a parameter P which limits domain size and should be larger than Max. Users can use their prior knowledge to set the P to avoid the unnecessary computation.

The further improvement of CHDF method should be: First, as chromatin architecture displays a hierarchical structure at different scales, we can improve CHDF model to find hierarchical domain structure. Second, as the calculation of interaction matrix of larger dimension is time-consuming, we need to improve the speed of CHDF method. Furthermore, new strategy should be applied to sparse matrix, to reduce the cost of memory. Fourth, many inter-chromosome chromatin interactions in Hi-C data, which might provide more information to identify domains, should be considered as well.

METHODS

We suppose that three kinds of regions exist in a Hi-C interaction matrix (Figure 1): domain regions (D), the regions between two adjacent domains (A) and the residuals (R). For domain k, the starting and ending positions are $b1_k$ and $b2_k$, respectively. So regions of D_k, A_k and R in the N-dimensional matrix:

$$D_k = \{(i, j) : b1_k \leqslant i < j \leqslant b2_k\},$$

$$A_k = \{(i, j) : b1_k \leqslant i \leqslant b2_k, b1_{k-1} \leqslant j \leqslant b2_{k-1}\},$$

$$R = \{(i, j) : 1 \leqslant i < j \leqslant N\} \cap \overline{A} \cap \overline{D}.$$

Every domain or region between two adjacent domains (corresponding to the interaction of the two domains) is regarded as a cluster. We do clustering to find the borders of domains. The border of a domain is defined as the two end points of a domain.

Our goal is to identify a set of non-overlapping domains D. The sum-of-squared-error criterion is used here as it is the most widely used and the simplest criterion function for clustering. For each cluster, let x be the sample of Hi-C interaction data between two bins (see the definition of bins in Supplementary Materials), n the number of the sample and m the mean of the sample. The sum-of-squared-error S in a cluster is:

$$S = \sum_{k=1}^{n} (x_k - m)^2.$$

When the number of domains is K, the sum-of-squared-error criterion J_e is:

$$J_e = \sum_{k=1}^{K} S_{D_k} + \sum_{k=2}^{K} S_{A_k} + S_{D_R}.$$

We not only need to minimize the sum-of-squared-error, but also find the domains where the Hi-C interaction intensity is much higher than that of the regions between two adjacent domains. So a penalty term P was added:

$$P_k = \text{sign}(m_{D_k} - m_{A_k})(m_{D_k} - m_{A_k})^2 l_{D_k},$$

where l is the length of the domain.

The goal is to identify a set of non-overlapping domains D_k that optimizes the following objective:

$$O = \min\left(J_e - \sum_{k=2}^{N} P\right).$$

Dynamic programming algorithm [36] is applied here to identify optimal solutions efficiently. The details of using dynamic programming algorithm was provided in Supplemental materials.

SUPPLEMENTARY MATERIALS

The supplementary materials can be found online with this article at DOI 10.1007/s40484-015-0047-9.

ACKNOWLEDGEMENTS

This work is supported by National Key Basic Research Project (973 program, 2012CB316503) and the National Natural Science Foundation of China (Nos. 31361163004 and 91019016).

COMPLIANCE WITH ETHICS GUIDELINES

The authors Yang Wang, Yanjian Li, Juntao Gao and Michael Q. Zhang declare they have no conflict of interest.

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

REFERENCES

1. Cremer, T. and Cremer, C. (2001) Chromosome territories, nuclear architecture and gene regulation in mammalian cells. Nat. Rev. Genet., 2, 292–301
2. Dekker, J., Marti-Renom, M. A. and Mirny, L. A. (2013) Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. Nat. Rev. Genet., 14, 390–403
3. Betzig, E., Trautman, J. K., Harris, T. D., Weiner, J. S. and Kostelak, R. L. (1991) Breaking the diffraction barrier: optical microscopy on a nanometric scale. Science, 251, 1468–1470
4. Bretschneider, S., Eggeling, C. and Hell, S. W. (2007) Breaking the diffraction barrier in fluorescence microscopy by optical shelving. Phys. Rev. Lett., 98, 218103
5. Langer-Safer, P. R., Levine, M. and Ward, D. C. (1982) Immunological method for mapping genes on *Drosophila* polytene chromosomes. Proc. Natl. Acad. Sci. USA, 79, 4381–4385
6. Lichter, P., Tang, C. J., Call, K., Hermanson, G., Evans, G. A., Housman, D. and Ward, D. C. (1990) High-resolution mapping of human chromosome 11 by *in situ* hybridization with cosmid clones. Science, 247, 64–69
7. Nora, E. P., Lajoie, B. R., Schulz, E. G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., van Berkum, N. L., Meisig, J., Sedat, J., et al. (2012) Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature, 485, 381–385
8. 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
9. Lévy-Leduc, C., Delattre, M., Mary-Huard, T. and Robin, S. (2014) Two-dimensional segmentation for analyzing Hi-C data. Bioinformatics, 30, i386–i392
10. Hu, M., Deng, K., Qin, Z., Dixon, J., Selvaraj, S., Fang, J., Ren, B. and Liu, J. S. (2013) Bayesian inference of spatial organizations of chromosomes. PLoS Comput. Biol., 9, e1002893
11. Tanizawa, H., Iwasaki, O., Tanaka, A., Capizzi, J. R., Wickramasinghe, P., Lee, M., Fu, Z. and Noma, K. (2010) Mapping of long-range associations throughout the fission yeast genome reveals global genome organization linked to transcriptional regulation. Nucleic Acids Res., 38, 8164–8177
12. Duan, Z., Andronescu, M., Schutz, K., McIlwain, S., Kim, Y. J., Lee, C., Shendure, J., Fields, S., Blau, C. A. and Noble, W. S. (2010) A three-dimensional model of the yeast genome. Nature, 465, 363–367
13. Rao, S. S. P., 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
14. Varoquaux, N., Ay, F., Noble, W. S. and Vert, J. P. (2014) A statistical approach for inferring the 3D structure of the genome. Bioinformatics, 30, i26–i33
15. Léger-Silvestre, I., Trumtel, S., Noaillac-Depeyre, J. and Gas, N. (1999)

Functional compartmentalization of the nucleus in the budding yeast *Saccharomyces cerevisiae*. Chromosoma, 108, 103–113

16. Thompson, M., Haeusler, R. A., Good, P. D. and Engelke, D. R. (2003) Nucleolar clustering of dispersed tRNA genes. Science, 302, 1399–1401
17. Haeusler, R. A., Pratt-Hyatt, M., Good, P. D., Gipson, T. A. and Engelke, D. R. (2008) Clustering of yeast tRNA genes is mediated by specific association of condensin with tRNA gene transcription complexes. Genes Dev., 22, 2204–2214
18. Hoang, S. A. and Bekiranov, S. (2013) The network architecture of the *Saccharomyces cerevisiae* genome. PLoS One, 8, e81972
19. 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
20. Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., Parrinello, H., Tanay, A. and Cavalli, G. (2012) Three-dimensional folding and functional organization principles of the *Drosophila* genome. Cell, 148, 458–472
21. Dostie, J., Richmond, T. A., Arnaout, R. A., Selzer, R. R., Lee, W. L., Honan, T. A., Rubio, E. D., Krumm, A., Lamb, J., Nusbaum, C., et al. (2006) Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. Genome Res., 16, 1299–1309
22. Phillips-Cremins, J. E., Sauria, M. E. G., Sanyal, A., Gerasimova, T. I., Lajoie, B. R., Bell, J. S. K., 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
23. Williamson, I., Berlivet, S., Eskeland, R., Boyle, S., Illingworth, R. S., Paquette, D., Dostie, J. and Bickmore, W. A. (2014) Spatial genome organization: contrasting views from chromosome conformation capture and fluorescence *in situ* hybridization. Genes Dev., 28, 2778–2791
24. Handoko, L., Xu, H., Li, G., Ngan, C. Y., Chew, E., Schnapp, M., Lee, C. W. H., Ye, C., Ping, J. L. H., Mulawadi, F., et al. (2011) CTCF-mediated functional chromatin interactome in pluripotent cells. Nat. Genet., 43, 630–638
25. Duggal, G., Wang, H. and Kingsford, C. (2014) Higher-order chromatin domains link eQTLs with the expression of far-away genes. Nucleic Acids Res., 42, 87–96
26. Gaffney, D. J., Veyrieras, J. B., Degner, J. F., Pique-Regi, R., Pai, A. A., Crawford, G. E., Stephens, M., Gilad, Y. and Pritchard, J. K. (2012) Dissecting the regulatory architecture of gene expression QTLs. Genome Biol., 13, R7
27. Dimas, A. S., Deutsch, S., Stranger, B. E., Montgomery, S. B., Borel, C., Attar-Cohen, H., Ingle, C., Beazley, C., Gutierrez Arcelus, M., Sekowska, M., et al. (2009) Common regulatory variation impacts gene expression in a cell type-dependent manner. Science, 325, 1246–1250
28. Veyrieras, J. B., Kudaravalli, S., Kim, S. Y., Dermitzakis, E. T., Gilad, Y., Stephens, M. and Pritchard, J. K. (2008) High-resolution mapping of expression-QTLs yields insight into human gene regulation. PLoS Genet., 4, e1000214
29. Zeller, T., Wild, P., Szymczak, S., Rotival, M., Schillert, A., Castagne, R., Maouche, S., Germain, M., Lackner, K., Rossmann, H., et al. (2010) Genetics and beyond—the transcriptome of human monocytes and disease susceptibility. PLoS One, 5, e10693
30. Myers, A. J., Gibbs, J. R., Webster, J. A., Rohrer, K., Zhao, A., Marlowe, L., Kaleem, M., Leung, D., Bryden, L., Nath, P., et al. (2007) A survey of genetic human cortical gene expression. Nat. Genet., 39, 1494–1499
31. Schadt, E. E., Molony, C., Chudin, E., Hao, K., Yang, X., Lum, P. Y., Kasarskis, A., Zhang, B., Wang, S., Suver, C., et al. (2008) Mapping the genetic architecture of gene expression in human liver. PLoS Biol., 6, e107
32. Pickrell, J. K., Marioni, J. C., Pai, A. A., Degner, J. F., Engelhardt, B. E., Nkadori, E., Veyrieras, J. B., Stephens, M., Gilad, Y. and Pritchard, J. K. (2010) Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature, 464, 768–772
33. Stranger, B. E., Nica, A. C., Forrest, M. S., Dimas, A., Bird, C. P., Beazley, C., Ingle, C. E., Dunning, M., Flicek, P., Koller, D., et al. (2007) Population genomics of human gene expression. Nat. Genet., 39, 1217–1224
34. Innocenti, F., Cooper, G. M., Stanaway, I. B., Gamazon, E. R., Smith, J. D., Mirkov, S., Ramirez, J., Liu, W., Lin, Y. S., Moloney, C., et al. (2011) Identification, replication, and functional fine-mapping of expression quantitative trait loci in primary human liver tissue. PLoS Genet., 7, e1002078
35. Phillips-Cremins, J. E. (2014) Unraveling architecture of the pluripotent genome. Curr. Opin. Cell Biol., 28, 96–104
36. Bellman, R. and Kotkin, B. (1961) On the Approximation of Curves by Line Segments Using Dynamic Programming. Commun. ACM, 4, 284
37. Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W., et al. (2008) Model-based analysis of ChIP-Seq (MACS). Genome Biol., 9, R137
38. Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M. and Haussler, D. (2002) The human genome browser at UCSC. Genome Res., 12, 996–1006
39. Ziebarth, J. D., Bhattacharya, A. and Cui, Y. (2013) CTCFBSDB 2.0: a database for CTCF-binding sites and genome organization. Nucleic Acids Res., 41, D188–D194
40. Yue, F., Cheng, Y., Breschi, A., Vierstra, J., Wu, W., Ryba, T., Sandstrom, R., Ma, Z., Davis, C., Pope, B. D., et al. (2014) A comparative encyclopedia of DNA elements in the mouse genome. Nature, 515, 355–364
41. Bernstein, B. E., Stamatoyannopoulos, J. A., Costello, J. F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M. A., Beaudet, A. L., Ecker, J. R., et al. (2010) The NIH roadmap epigenomics mapping consortium. Nat. Biotechnol., 28, 1045–1048
42. Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 25, 1754–1760
43. Yaffe, E. and Tanay, A. (2011) Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. Nat. Genet., 43, 1059–1065

