1 **HBP: an integrative and flexible pipeline for the interaction analysis of Hi-C dataset**

2

3 Chao He¹, Ping Li¹, Minglei Shi¹, Yan Zhang¹, Bingyu Ye¹, Dejian Xie¹, Wenlong Shen^{1,*} and Zhihu Zhao^{1,*}

4 ¹Beijing Institute of Biotechnology, Beijing 100071, China.

5 *Correspondence: shenwl1988@gmail.com, zhaozh@bmi.ac.cn

6 Email address:

7 Chao He: hechao2010@tsinghua.org.cn

8 Ping Li: lipingtry@163.com

9 Minglei Shi: shimi79@126.com

10 Yan Zhang: zany1983@163.com

11 Bingyu Ye: bby0804@gmail.com

12 Dejian Xie: ncuskxiedejian@163.com

13 Wenlong Shen: shenwl1988@gmail.com

14 Zhihu Zhao: zhaozh@bmi.ac.cn

15

16

17

18

19

20

21

22

1 **Abstract**

2 **Background:** The spatial organization of interphase chromatin in the nucleus play an important role in
 3 gene expression regulation and function. With the rapid development of revolutionized chromosome
 4 conformation capture technology and its genome-wide derivatives such as Hi-C, investigation of the
 5 genome folding becomes more efficient and convenient. How to robustly deal with these massive
 6 datasets and infer accurate 3D model and within-nucleus compartmentalization of chromosomes
 7 becomes a new challenge.

8 **Result:** The implemented pipeline HBP (Hi-C BED file analysis Pipeline) integrates existing pipelines
 9 focusing on individual steps of Hi-C data processing into an all-in-one package with adjustable
 10 parameters to infer the consensus 3D structure of genome from raw Hi-C sequencing data. What's
 11 more, HBP could assign statistical confidence estimation for chromatin interactions, and clustering
 12 interaction loci according to enrichment tracks or topological structure automatically.

13 **Conclusion:** The freely available HBP is an optimized and flexible pipeline for analyzing the folding of
 14 whole chromosome and interactions between some specific sites from the Hi-C raw sequencing reads
 15 to the partially processed datasets. The other complex genetic and epigenetic datasets from public
 16 sources such as GWAS, ENCODE consortiums etc. will also easily be integrated into HBP, hence the
 17 final output results of HBP could provide a comprehensive in-depth understanding for the specific
 18 chromatin interactions, potential molecular mechanisms and biological significance. We believe that
 19 HBP is a reliable tool for the rapidly analysis of Hi-C data and will be very useful for a wide range of
 20 researchers, particularly those who lack of background in computational biology. HBP is freely accessible at
 21 https://github.com/hechao0407/HBP/blob/master/HBP_1.0.tar.gz.

Keywords: Algorithms, Hi-C, Visualization, Chromatin Conformation

Background

Numerous light microscopy studies have shown that chromatin is tightly packaged and organized into higher-order dynamic structures, which is essential for the genome function[1-3]. Indeed, the organization of chromatin was found to play important regulation roles in many biological processes such as transcription, DNA replication, repair and even translation [4-6]. With the occurrence and rapid development of DNA-FISH[3], other super-resolution imaging[7-9] and molecular techniques, particularly newly developed 3C (Chromosome Conformation Capture)[10] and its genome-wide derivatives such as 4C[11, 12], 5C[13], ChIA-PET[14], Hi-C[15], series of Capture-C[16] etc., the study of the chromosome folding and dynamics has made great progress in recent years and the resolution achieves an unprecedented high-level[17, 18]. In fact, the resolution of Hi-C for human genome already achieved to 1-kilobase [19], which could even be reached at the 64bp theoretically as the very recent work demonstrated [20].

Combined with the numerous newly developed computational modeling and data processing methods[21], the conventional or 'ensemble' C-series studies of cell populations irrefutably revealed that the interphase chromatin is hierarchically and dynamic folded to form the higher-order fractal globule structure and chromosome territories, which is coincided with the results of the DNA-FISH at the single cell level in most cases[17, 18]. It also revealed the TAD (Topological Associated Domain) is the conserved structural and functional units of the chromatin folding [22, 23]. Each TAD is composed of sub-TADs and numerous loop interactions that within the same TADs are much higher than those

1 across TADs. The different TADs could further cluster into larger-scale compartments corresponding to
2 transcription active or inactive chromatin[24]. Although numerous different computational modeling
3 methods such as GenomicInteractions[25], HiTC[26], hiclib[27], ChromContact[28], HiCPlotter[29],
4 HiC-Pro[21, 30] etc. have been developed and promote the systematic, global analysis of Hi-C results,
5 unfortunately, we still do not have an integrated general platform to analyze interactions of global and
6 specific sequences from mapping, filtering, interaction calling, statistical evaluation and visualization
7 of interactions automatically. Recently, Cournac and colleagues developed some program to analyze
8 interactions between repetitive elements[31], but the program only calculate co-localization scores to
9 evaluate genomic imprints of the 3D folding, and thus cannot be used in a single command-line
10 manner either. The GenomicInteractions [25] is also able to analyze the interaction network of specific
11 sites. However, it can neither process data from raw Hi-C reads, add tracks to investigate the
12 enrichment of histones or transcription factor binding sites, nor perform the statistical analysis.
13 HiCPlotter focuses on the association between global Hi-C interaction heatmap and other epigenetic
14 features such as histone modification[29]. But it cannot investigate interactions between specific
15 types of chromatin sequences. ChromContact is a web tool for analyzing spatial contact chromosomes.
16 It can retrieve all the information of loci of interest from the UCSC browser directly[28]. However, it is
17 incompatible with researcher's user-supplied data, and cannot provide information such as interaction
18 frequency or probability etc.

19 To fulfill the great challenges and provide an integrative resolution for Hi-C data processing, we
20 present HBP, an open-source, easy-to-use complete pipeline to provide the tools for the sequence
21 mapping of raw reads, extraction and normalization of chromosome interactions, heatmap
22 visualization of interaction frequency, network analysis of specific interactions, statistical evaluation

with adjustable parameters. HBP is an intact, integrative all-in-one workflow for the whole process of Hi-C data mapping, filtering and visualizing. HBP can start either from Hi-C raw dataset or other existing interaction frequency matrix. Here, we illustrate features of HBP by applying it to some publicly available datasets, and also show some interesting findings.

Implementation

HBP is a publicly available R package for processing Hi-C datasets. It uses Hi-C raw datasets or interaction matrix files from hiclib[27] or HiC-Pro [30], a BED format file containing specific sites information. HBP is also compatible with the tracks files of histone modification or transcription factors DNA binding motif enrichment information of ChIP-seq which can be downloaded from the UCSC [32] or user-own-supplied files. Users can examine the interaction frequency distribution, explore the network topology information such as the list of name, degree, closeness, betweenness, local cluster coefficient, eigenvector centrality of network nodes etc. Besides providing the global scatter diagram to pinpoint locations of specific sites, circos[33] pictures with different tracks, network clusters according to these tracks or topological structure of the global genome and statistical evaluation, HBP could also show the frequency, number and distribution of interactions between specific sites in the genome. These information are very important for the understanding of the chromatin interaction networks. Here we provide vignettes detailing the use of HBP in Hi-C data.

Data import

HBP can import different categories datasets such as the raw Hi-C sequencing files or the output files from HiC-Pro [30], specific sites information of standard BED format and tracks information of WIG

format etc. This flexible feature allows users to easily integrate the datasets from many different sources such as ENCODE[34], modENCODE [35], GWAS [36]consortium and UCSC [32] etc.

Interaction frequency distribution

Following the Gu's work [37], HBP sets up and down thresholds, and discards the interaction-pairs whose frequencies are lower than the down threshold or higher than the up threshold. Then HBP separates chromatin interaction frequencies into many bins (referred to as frequency bin), and computes density for specific sites in each frequency bins. Meanwhile, HBP will compare the results with that of random control groups which were generated according to the distribution of these imported specific sites. Based on these imported specific region, every member of random control groups was calculated and made in the similar region. This comparison results make it easier for users to ascertain whether the observed interactions were biologically significant and whether these specific sites have effect on the conformation of chromatin or not.

Interaction network topological analysis

HBP maps sequence sites of interest to the probability matrix, and calculate the interaction frequency and pinpoint the interaction networks mediated by these sites. The igraph package was used to plot an interaction network graph and make topological cluster analysis. According to the clustering results, users can classify all sites of interest into different types or clusters, and examine their potential differential properties respectively.

Visualization of interactions and tracks

The package can visualize the interaction of interested sequences. It is compatible with the adding of some tracks such as histone modifications or transcription factors binding sites enrichment. From the graph, users can easily find out the association rules of the chromatin interactions and other genetic and epigenetic features of the interested sequences, put forward hypotheses and further design related experiments to verify.

Statistical significance tests

HBP can also perform cluster analysis using histone modifications or transcription factors binding tracks, and make Kruskal-Wallis rank sum or T test to evaluate the statistical difference of interaction frequency, or the interaction strength of the specific sites respectively. These tests will tell us whether there exist significant differences, suggesting different properties of these sites and the interactions.

RESULT

Usage examples

Investigating effect of different repeat elements on the chromatin conformation

The repeat sequence accounts of around 50% of human genome and is reported to be related to many different physiological process of genome [34, 38]. We and others previously reported some repeat sequences are involved in chromosome interactions [11, 31, 37]. Here we examine whether all the different kinds of repeat sequence might mediate extensive chromatin interactions. To this aim, we provide a demonstration of using HBP to perform an example analysis of Hi-C data from human IMR90 cells (GEO dataset GSM1551600) [19]. With the built-in function to import data from HiC-Pro interaction file format, HBP pre-processed the Hi-C data by HiC-Pro package, and the result was

normalized by the iterative correction algorithm of HiC-Pro to remove systematic noise and bias. After downloaded from Repbase in the form of FASTA file [39], and aligned to hg19 by Blat [40], the datasets of repeat elements of L1, LTR retrotransposon, SINE, satellite, ERV2 and DNA transposons etc. were inputted into our pipeline. Then the interactions frequency distribution associated with these repeat elements was estimated. To confirm the difference, HBP also set up random control groups corresponding to the distribution of different repeat elements. Figure 2 shows the distribution of six repeat elements families. While the interaction frequency of satellite and SINE involved is much lower than random controls, that of L1 is higher than random groups. These results indicated as a family, L1 but not the satellite and SINE repeat elements, is globally mediated or maintained chromatin interactions, which is agreement with a previous model [41]. Compared with control, the frequency of other elements is fluctuated, suggesting these elements maybe need to be classified into different sub-families for further accurate analysis.

Investigating the CTCF interaction network in human GM12878 cells

Alone or in combination with other protein and even RNA regulators, CTCF mediated extensive chromatin interactions [42, 43], but the topological features of CTCF-involved chromatin interaction networks remain elusive. Here we examined this issue by inputting the human CTCF DNA binding peaks of interphase chromatin into our pipeline firstly. GM12878 CTCF peaks were downloaded from ENCODE project [34, 38], GM12878 Hi-C data was downloaded from Rao's work (GEO dataset GSM1551591) [19]. HBP processed the Hi-C dataset by HiC-Pro, invalid pairs and duplicates were removed automatically.

1 All CTCF-involved interactions, defined as both sites of interactions with CTCF-binding peaks around,
2 were filtered from the dataset, and the interaction network and interaction frequency distribution
3 were checked by HBP. As an example, we just examined the chosen region of
4 chr8:69,300,000-74,300,000. Firstly, it was found that both the enrichments (Figure 3 a, b) and peaks
5 (Figure 3 c, d) of the global chromosome interactions correlated with CTCF-involved specific
6 interactions very well. These results, along with the previous comparison of ChIA-PET with Hi-C [44],
7 strongly indicated CTCF is the main architectural protein bridging genome topology[42]. Secondly, it
8 was found all CTCF peaks within this region (98 in total) were involved in the interaction network and
9 have either higher or lower interactions with each other (Figure 4 a), additionally, the interaction
10 network plotted by HBP could be sorted into two clusters (contained 46, 52 nodes respectively)
11 according to network topological structure (Figure 4 b, c). These interaction clusters were defined as
12 of TSC (Topological Structure Cluster), which had different distribution relative to genes position. For
13 example, while majority of TSC1 peaks from intron, TSC2 didn't have peaks in the 3'UTR and exons
14 regions (Fig 4 b). The different distribution of two TSCs relative to gene position implied each of these
15 TSCs might has different regulation function, which needs further examinations. What's more, just like
16 Fig 3 d showed, the two clusters just belonged to two different peaks respectively, and they were
17 separated by the TAD boundary at chr8:71,760,000-71,800,000. Additionally, we also found the node
18 in chr8:73,199,355-73,200,024 had the highest degree (degree = 79) and connected with more than
19 half nodes, which suggested that this node might function as a connector or hub of the whole
20 network. To further characterize this node, we looked into this area by UCSC genome browser. We
21 found that it was a DNase I Hypersensitivity sites cluster with many transcription factors binding sites
22 of EBF1, IRF4, ZNF143, RAD21, SMC3, etc. within (Figure 4 c). This result provide a potential molecular

1 mechanism explanation for the network formation.

2

3 In order to learn more about the characteristics of CTCF involved interactions, we used HBP to plot a

4 circos picture (Figure 5 a) by adding 6 histone modification features of H3K4me1, H3K4me2, H3K4me3,

5 H3K27ac, H3K36me3 and H3K79me2. In the picture, the interaction stronger, the color deeper. We

6 could clearly found that some interaction nodes have H3K4me1 and H3K4me2 enrichment from the

7 tracks heatmap, which are posttranslational marks associated with enhancer regions. To know more

8 details, all above mentioned six histone modification datasets were inputted and histone modification

9 clusters were made by HBP, a cluster tree and a heatmap sorted by these clusters was plotted (Figure

10 5 b). These clusters had different modification enrichment according to the heatmap. While cluster 2

11 has very lower even no any enrichment of all 6 histone modification, cluster 1 and 3 have relative

12 higher enrichment of all 6 histone modifications. Additionally, compared with cluster 3, cluster1 had

13 higher H3K27ac enrichment but lower H3K36me3 enrichment, cluster3 was just on the contrary. As

14 the Histone H3K27ac separates active from poised enhancers[45, 46] and higher H3K36me3

15 enrichment marks the active exon[46, 47], these results strongly suggested CTCF cluster 1 is active

16 enhancer whereas cluster 3 is active expressed gene bodies, especially exons.

17

18 To further quantitatively determine whether there existed statistical differences in these histone

19 modifications between these CTCF binding sites and other sites, Kruskal-Wallis rank sum tests were

20 made by HBP. The test shown that H3k4me1, H3k4me2, H3k4me3, H3K27ac, H3K36me3 and

21 H3K79me2 markers existed significant differences ($p = 4.17e-05, 4.93e-03, 1.67e-07, 1.21e-03,$

22 $4.55e-05, 5.30e-04$ respectively). Then T Test was carried out to estimate histone modification level of

1 each binding site or node, and a node list containing information about modification level higher or
2 lower than average level was obtained (Supple Table 1). What's interesting is that 61.2% of the nodes
3 had the coherent tendency of H3K4me1, H3K4me2 and H3K27ac relative high enrichment level, which
4 is the typical markers of active enhancers [45]. We chose one node for further analysis. It indeed just
5 located within a super-enhancer (Figure 5 c). Super-enhancers are clusters of transcriptional
6 enhancers that usually drive expression of genes important for defining cell identity during
7 development[48]. What's more, the H3K36me3 and H3K79me2 modification level in 73.5% nodes
8 showed similar tendency, it was suggested that these site might work together in the boundary of
9 exon-intron regions [45]. At last, T Test was also used to evaluate the interaction number between
10 CTCF binding sites. HBP selected 100 random group with the similar distribution of CTCF binding sites
11 to set up control. It was found that while the interaction number of random group was 1393 ~ 1409 at
12 95 percent confidence interval, that of the CTCF binding sites involved in this region is 1423 ($p =$
13 $1.21e-05$). These results showed that in chr8:69,300,000-74,300,000, CTCF involved interaction
14 number was much higher than random group, in other words, CTCF binding sites were more likely to
15 interact with each other in this region.

17 **Conclusions**

18 We proposed a new pipeline, HBP, to analyze global and defined interactions between specific sites
19 from Hi-C data. Compared with other software, HBP is more optimized and flexible by providing many
20 user-adjustable parameters, and can process Hi-C data from raw sequence or already partially
21 processed data in an all-in-one manner. All these unique features make HBP will be a robust and
22 useful tools for the investigation of genome hierarchical folding and functions, particularly for

1 researchers without strong background in computational biology.

2

3

4 **Availability and requirements**

5 HBP is a publicly available R package available from <https://github.com/hechao0407/HBP/blob/master>

6 [/HBP_1.0.tar.gz](#). This package requires R > = 3.3.0 and depend on several R packages including

7 optparse, grid, lattice, IDPmisc, OmicCircos, stringr, ggplot2, igraph, reshape2, pgirmess, coin,

8 multcomp, flexclust, HiTC, rtracklayer and gplots. All of the analyses and figures presented in the

9 paper can be reproduced by using HBP.

10

11 **Ethics statement**

12 Ethics statement is not applicable to our study as this study only uses publicly available data.

13

14 **Competing interests**

15 The authors declare that they have no competing interests.

16

17 **Authors' contributions**

18 CH, WLS and ZhZ conceived the research. CH designed the software. CH, PL, MIS, YZ, WIS and ZhZ

19 contributed to the development of software and documentation. CH is responsible for the

20 maintenance of software. All authors contributed to the writing of the manuscript. All authors read

21 and approved the final manuscript.

22

Acknowledgements

The authors would like to thank Zhihu Zhao, Wenlong Shen, and Jing Zeng for their helpful discussions during the development of this package. This work was supported by National Basic Research Program of China [2013CB966802], National Natural Science Foundation of China [31370762, 31030026, 31272416, 81372218] and National Key Technology R&D Program of China [2012BAI01B07].

References

1. Cremer T, Cremer C: **Rise, fall and resurrection of chromosome territories: a historical perspective. Part I. The rise of chromosome territories.** *Eur J Histochem* 2006, **50**(3):161-176.
2. Cremer T, Cremer C: **Rise, fall and resurrection of chromosome territories: a historical perspective. Part II. Fall and resurrection of chromosome territories during the 1950s to 1980s. Part III. Chromosome territories and the functional nuclear architecture: experiments and models from the 1990s to the present.** *Eur J Histochem* 2006, **50**(4):223-272.
3. Rouquette J, Cremer C, Cremer T, Fakan S: **Functional nuclear architecture studied by microscopy: present and future.** *Int Rev Cell Mol Biol* 2010, **282**:1-90.
4. Diamant A, Pinter RY, Tuller T: **Three-dimensional eukaryotic genomic organization is strongly correlated with codon usage expression and function.** *Nat Commun* 2014, **5**:5876.
5. Shen W, Wang D, Ye B, Shi M, Ma L, Zhang Y, Zhao Z: **GC3-biased gene domains in mammalian genomes.** *Bioinformatics* 2015, **31**(19):3081-3084.
6. Cavalli G, Misteli T: **Functional implications of genome topology.** *Nat Struct Mol Biol* 2013, **20**(3):290-299.
7. Klar TA, Jakobs S, Dyba M, Egner A, Hell SW: **Fluorescence microscopy with diffraction resolution barrier broken by stimulated emission.** *Proc Natl Acad Sci U S A* 2000, **97**(15):8206-8210.
8. Rust MJ, Bates M, Zhuang X: **Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM).** *Nat Methods* 2006, **3**(10):793-795.

- 1 9. Boettiger AN, Bintu B, Moffitt JR, Wang S, Beliveau BJ, Fudenberg G, Imakaev M, Mirny LA,
2 Wu CT, Zhuang X: **Super-resolution imaging reveals distinct chromatin folding for different**
3 **epigenetic states**. *Nature* 2016, **529**(7586):418-422.
- 4 10. Dekker J, Rippe K, Dekker M, Kleckner N: **Capturing chromosome conformation**. *Science*
5 2002, **295**(5558):1306-1311.
- 6 11. Zhao Z, Tavoosidana G, Sjolinder M, Gondor A, Mariano P, Wang S, Kanduri C, Lezcano M,
7 Sandhu KS, Singh U *et al*: **Circular chromosome conformation capture (4C) uncovers**
8 **extensive networks of epigenetically regulated intra- and interchromosomal interactions**.
9 *Nat Genet* 2006, **38**(11):1341-1347.
- 10 12. Simonis M, Klous P, Splinter E, Moshkin Y, Willemsen R, de Wit E, van Steensel B, de Laat W:
11 **Nuclear organization of active and inactive chromatin domains uncovered by chromosome**
12 **conformation capture-on-chip (4C)**. *Nat Genet* 2006, **38**(11):1348-1354.
- 13 13. Dostie J, Richmond TA, Arnaout RA, Selzer RR, Lee WL, Honan TA, Rubio ED, Krumm A, Lamb J,
14 Nusbaum C *et al*: **Chromosome Conformation Capture Carbon Copy (5C): a massively**
15 **parallel solution for mapping interactions between genomic elements**. *Genome Res* 2006,
16 **16**(10):1299-1309.
- 17 14. Fullwood MJ, Liu MH, Pan YF, Liu J, Xu H, Mohamed YB, Orlov YL, Velkov S, Ho A, Mei PH *et al*:
18 **An oestrogen-receptor-alpha-bound human chromatin interactome**. *Nature* 2009,
19 **462**(7269):58-64.
- 20 15. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie
21 BR, Sabo PJ, Dorschner MO *et al*: **Comprehensive mapping of long-range interactions reveals**
22 **folding principles of the human genome**. *Science* 2009, **326**(5950):289-293.
- 23 16. Mifsud B, Tavares-Cadete F, Young AN, Sugar R, Schoenfelder S, Ferreira L, Wingett SW,
24 Andrews S, Grey W, Ewels PA *et al*: **Mapping long-range promoter contacts in human cells**
25 **with high-resolution capture Hi-C**. *Nat Genet* 2015, **47**(6):598-606.
- 26 17. Dekker J, Mirny L: **The 3D Genome as Moderator of Chromosomal Communication**. *Cell*
27 2016, **164**(6):1110-1121.
- 28 18. Denker A, de Laat W: **The second decade of 3C technologies: detailed insights into nuclear**
29 **organization**. *Genes Dev* 2016, **30**(12):1357-1382.
- 30 19. Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol
31 I, Omer AD, Lander ES *et al*: **A 3D map of the human genome at kilobase resolution reveals**
32 **principles of chromatin looping**. *Cell* 2014, **159**(7):1665-1680.
- 33 20. Darrow EM, Huntley MH, Dudchenko O, Stamenova EK, Durand NC, Sun Z, Huang SC, Sanborn
34 AL, Machol I, Shamim M *et al*: **Deletion of DXZ4 on the human inactive X chromosome alters**
35 **higher-order genome architecture**. *Proc Natl Acad Sci U S A* 2016, **113**(31):E4504-4512.
- 36 21. Ay F, Noble WS: **Analysis methods for studying the 3D architecture of the genome**. *Genome*
37 *Biol* 2015, **16**:183.
- 38 22. Nora EP, Lajoie BR, Schulz EG, Giorgetti L, Okamoto I, Servant N, Piolot T, van Berkum NL,
39 Meisig J, Sedat J *et al*: **Spatial partitioning of the regulatory landscape of the X-inactivation**
40 **centre**. *Nature* 2012, **485**(7398):381-385.
- 41 23. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B: **Topological domains in**
42 **mammalian genomes identified by analysis of chromatin interactions**. *Nature* 2012,
43 **485**(7398):376-380.

- 1 24. Gibcus JH, Dekker J: **The hierarchy of the 3D genome.** *Mol Cell* 2013, **49**(5):773-782.
- 2 25. Harmston N, Ing-Simmons E, Perry M, Baresic A, Lenhard B: **GenomicInteractions: An**
3 **R/Bioconductor package for manipulating and investigating chromatin interaction data.**
4 *BMC Genomics* 2015, **16**(1):963.
- 5 26. Servant N, Lajoie BR, Nora EP, Giorgetti L, Chen CJ, Heard E, Dekker J, Barillot E: **HiTC:**
6 **exploration of high-throughput 'C' experiments.** *Bioinformatics* 2012, **28**(21):2843-2844.
- 7 27. Naumova N, Imakaev M, Fudenberg G, Zhan Y, Lajoie BR, Mirny LA, Dekker J: **Organization of**
8 **the mitotic chromosome.** *Science* 2013, **342**(6161):948-953.
- 9 28. Sato T, Suyama M: **ChromContact: A web tool for analyzing spatial contact of chromosomes**
10 **from Hi-C data.** *BMC Genomics* 2015, **16**:1060.
- 11 29. Akdemir KC, Chin L: **HiCPlotter integrates genomic data with interaction matrices.** *Genome*
12 *Biol* 2015, **16**:198.
- 13 30. Servant N, Varoquaux N, Lajoie BR, Viara E, Chen CJ, Vert JP, Heard E, Dekker J, Barillot E: **HiC-Pro: an**
14 **optimized and flexible pipeline for Hi-C data processing.** *Genome Biol* 2015,
15 **16**:259.
- 16 31. Cournac A, Koszul R, Mozziconacci J: **The 3D folding of metazoan genomes correlates with**
17 **the association of similar repetitive elements.** *Nucleic Acids Res* 2016, **44**(1):245-255.
- 18 32. Rosenbloom KR, Armstrong J, Barber GP, Casper J, Clawson H, Diekhans M, Dreszer TR, Fujita
19 PA, Guruvadoo L, Haeussler M *et al*: **The UCSC Genome Browser database: 2015 update.**
20 *Nucleic Acids Res* 2015, **43**(Database issue):D670-681.
- 21 33. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA: **Circos:**
22 **an information aesthetic for comparative genomics.** *Genome Res* 2009, **19**(9):1639-1645.
- 23 34. Consortium EP: **An integrated encyclopedia of DNA elements in the human genome.** *Nature*
24 2012, **489**(7414):57-74.
- 25 35. mod EC, Roy S, Ernst J, Kharchenko PV, Kheradpour P, Negre N, Eaton ML, Landolin JM,
26 Bristow CA, Ma L *et al*: **Identification of functional elements and regulatory circuits by**
27 **Drosophila modENCODE.** *Science* 2010, **330**(6012):1787-1797.
- 28 36. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T,
29 Hindorff L *et al*: **The NHGRI GWAS Catalog, a curated resource of SNP-trait associations.**
30 *Nucleic Acids Res* 2014, **42**(Database issue):D1001-1006.
- 31 37. Gu Z, Jin K, Crabbe MJ, Zhang Y, Liu X, Huang Y, Hua M, Nan P, Zhang Z, Zhong Y: **Enrichment**
32 **analysis of Alu elements with different spatial chromatin proximity in the human genome.**
33 *Protein Cell* 2016, **7**(4):250-266.
- 34 38. Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan KK, Cheng C, Mu XJ, Khurana E,
35 Rozowsky J, Alexander R *et al*: **Architecture of the human regulatory network derived from**
36 **ENCODE data.** *Nature* 2012, **489**(7414):91-100.
- 37 39. Kohany O, Gentles AJ, Hankus L, Jurka J: **Annotation, submission and screening of repetitive**
38 **elements in Repbase: RepbaseSubmitter and Censor.** *BMC Bioinformatics* 2006, **7**:474.
- 39 40. Kent WJ: **BLAT--the BLAST-like alignment tool.** *Genome Res* 2002, **12**(4):656-664.
- 40 41. Tang SJ: **A Model of Repetitive-DNA-Organized Chromatin Network of Interphase**
41 **Chromosomes.** *Genes (Basel)* 2012, **3**(1):167-175.
- 42 42. Ong CT, Corces VG: **CTCF: an architectural protein bridging genome topology and function.**
43 *Nat Rev Genet* 2014, **15**(4):234-246.

- 1 43. Ye BY, Shen WL, Wang D, Li P, Zhang Z, Shi ML, Zhang Y, Zhang FX, Zhao ZH: **ZNF143 is**
2 **involved in CTCF-mediated chromatin interactions by cooperation with cohesin and other**
3 **partners**. *Mol Biol (Mosk)* 2016, **50**(3):496-503.
- 4 44. Tang Z, Luo OJ, Li X, Zheng M, Zhu JJ, Szalaj P, Trzaskoma P, Magalska A, Wlodarczyk J,
5 Ruszczycki B *et al*: **CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin**
6 **Topology for Transcription**. *Cell* 2015, **163**(7):1611-1627.
- 7 45. Creighton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, Hanna J, Lodato MA,
8 Frampton GM, Sharp PA *et al*: **Histone H3K27ac separates active from poised enhancers and**
9 **predicts developmental state**. *Proc Natl Acad Sci U S A* 2010, **107**(50):21931-21936.
- 10 46. Zhou VW, Goren A, Bernstein BE: **Charting histone modifications and the functional**
11 **organization of mammalian genomes**. *Nat Rev Genet* 2011, **12**(1):7-18.
- 12 47. Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W,
13 Kim TK, Koche RP *et al*: **Genome-wide maps of chromatin state in pluripotent and**
14 **lineage-committed cells**. *Nature* 2007, **448**(7153):553-560.
- 15 48. Hnisz D, Abraham BJ, Lee TI, Lau A, Saint-Andre V, Sigova AA, Hoke HA, Young RA:
16 **Super-enhancers in the control of cell identity and disease**. *Cell* 2013, **155**(4):934-947.

17

18

19 Additional files

20 Figure1. The workflow of HBP. This workflow contains the main content of the software, can be used
21 to analyze the interaction between some specific sites.

22

23 Figure 2. Interactions frequency distribution of different repeat elements. The red line represents
24 interactions frequency distribution of repeat element, and the box plot represents distribution of
25 random control.

26

27 Figure 3. Interaction distribution of chr8: 69,300,000-74,300,000. a Hi-C interaction heatmap of this
28 region. b CTCF involved interaction distribution of this region. c The reads counts of each bin in Hi-C
29 interaction heatmap. d The reads counts of each bin in CTCF involved interaction distribution.

30

1 Figure 4. Interaction network of chr8:69,300,000-74,300,000. a The network picture without cluster.
2 The color of interaction line represent the read counts, the size of each node represent the degree,
3 and the color of each node represent topological structure clusters. b c The distribution of TSC1 and
4 TSC2. The two clusters have different characteristics in relation to genes. d The highest degree site in
5 chr8:69,300,000-74,300,000 plotted by UCSC. It is a DNase I Hypersensitivity sites cluster with many
6 transcription factors binding sites of EBF1, IRF4, ZNF143, RAD21, SMC3, etc.

7
8 Figure 5. Circos plot and tracks clusters heatmap of chr8:69,300,000-74,300,000. a Circos plot of the
9 defined region. From outside to inside, the six tracks are H3K27ac, H3K4me1, H3K4me2, H3K4me3,
10 H3K9ac, and H3K9me3 respectively. b Tracks clusters heatmap of the defined region. The three
11 clusters are separated by the black line in this picture. The blue line represents the value of
12 corresponding site. c The histone modification picture plotted by UCSC. An example node from cluster
13 1 of b with the black square line. It is a strong active enhancer with H3K4me1, H3K4me2, H3K27ac
14 enrichment peaks.

15
16 Supple Table 1. The list of CTCF binding sites in chr8:69,300,000-74,300,000. It contains information
17 about degree, histone modification level, cluster membership, etc.

L1

SINE

satellite

DNA transposon

ERV2

LTR retrotransposon

a

b

c

a

b

c