

## Supplementary Methods for Yang D. Jang I. et al, 3DIV: A 3D-genome Interaction Viewer and database

### Collection and processing of raw Hi-C sequencing data

In this database, we collected 349 Hi-C experiments performed on 96 samples including 18 human tissues and 78 human cell lines of various conditions [Table S1]. Raw sra files of 78 Hi-C samples were downloaded from Short Read Archive (SRA) and converted to fastq files using sra-toolkit. Remaining 18 Hi-C samples performed by ENCODE Project and Genomics of Gene Regulation (GGR) consortium were downloaded from ENCODE Data Coordinate Center (DCC) (1).

Each single-ended reads from paired-end Hi-C sequencing data were aligned to the human reference genome (hg19), separately, using BWA-mem (2) with default parameters, and then the pair was merged together as paired-end aligned bam file using in-house script in order to process chimeric reads which span the ligation site. These chimeric reads are the result of the ligation chemistry during Hi-C library construction, and they are not properly processed by paired-end BWA-mem. PCR duplicates were removed with Picard and low quality reads (MAPQ < 10) were discarded during downstream analysis.

As we focused on *cis*-chromosomal interactions, we discarded invalid Hi-C reads with the following conditions; inter-chromosomal interactions, putative self-ligations (distance between mates is shorter than 15kb), and non-restriction site specific ligations (distance to the nearest restriction site is longer than 500bp) as described in previous studies (3,4). We merged multiple biological replicates and processed samples that contain more than 5,000,000 valid *cis*-interactions, resulting in 80 samples to be used in downstream data processing. The average number of interactions for each 40kb bin was measured to demonstrate the data resolution.

### Processing to remove experimental biases

Raw Hi-C interaction frequencies are affected by multiple experimental biases caused by intrinsic DNA sequence, experimental artefacts, and et cetera. (5,6). In order to remove such biases, 3DIV applied negative binomial model-based implicit normalization approaches (3,4,7). Experimental biases were removed in 3DIV from 5kb resolution interaction frequency matrix by estimating negative binomial regression parameters based on 40kb resolution interaction frequency matrix. Then, 2D convolution smooth function was applied to bias-removed interaction frequency matrix  $I$ , in order to generate smoothed Hi-C interaction frequency matrix  $C$  as:

$$C = (I \otimes G_{5 \times 5}) \odot W$$

Where  $G_{5 \times 5}$  is 5 by 5 Gaussian matrix, and  $W = \{w_{i,j}\}$  is the weight matrix which compensated for the elimination of interactions shorter than 15kb.  $w_{i,j}$  is defined as:

$$w_{i,j} = \begin{cases} \frac{50}{d(17-d) - 22} & d < 8 \\ 1 & \text{otherwise} \end{cases}$$

Where  $d = |i - j|$ .

### **Evaluation of negative binomial model-based implicit normalization**

We validated our normalization pipeline compare to common pipelines. Common normalization pipelines, HiCNorm (5) and ICE (8), were used as the standard. First, reproducibility between biological replicates was measured along the distance. Mouse embryonic stem cell Hi-C data from different library construction conditions were used as the validation set (9). Correlation coefficient between normalized interaction frequencies was measured along the distance. Negative binomial model-based implicit normalization presented reliable reproducibility irrelevant to the genomic distance and resolution of interaction frequency matrix (Figure S1A). Similarly, TADs called from normalized interaction frequency matrices were compared. Normalized interaction frequency matrices in 40kb resolution were constructed from GM12878 in-situ Hi-C library using Mbol restriction enzyme (10). Then, DomainCaller (11) with default parameters identified TADs from normalized interaction frequency matrices. The assignment and the comparison of TAD boundaries were performed by scripts from HiC-bench (12). Most (92%) of the TAD boundaries called from our normalization pipeline were also identified in other normalized interaction frequency matrices (Figure S1B). This concordance of TAD boundaries is higher than those of TAD boundaries from several TAD calling algorithms (50~75%) (13). These benchmark results strongly support that negative binomial model-based implicit normalization is equivalent to the common Hi-C normalization pipelines.

### **Normalization against distance dependent background signals**

Bias-removed Hi-C interaction frequencies are still dominated by strong genomic distance-dependent background signals. These background signals make it difficult to identify significant long-range chromatin interactions. In order to remove such bias, LOWESS regression was applied to estimate the expected interaction frequency over the distance between loci. Then, the fold change between observed and expected interaction frequencies was assigned as the distance-normalized interaction frequency.

### **Interpretation of interaction frequency**

The significant chromatin interactions may have high interaction frequency in distance-normalized and bias-removed manner. 3DIV visualizes both distance-normalized and bias-removed interaction frequencies for this purpose. Users can identify significant chromatin interactions such as the interaction represented as an arc, supported by high bar graph in one-to-all interaction plot.

### **Scaling Hi-C data for the comparison mode**

Since 3DIV is a collection of various Hi-C data with different sequencing depth and library complexity, their interaction frequencies follow different scales. Thereby, scale adjustment of interaction frequencies is required to compare interaction frequencies between samples. To adjust the scale of interaction frequencies, 3DIV measured mean and standard deviation of bias-removed interaction frequency of each sample. The averages of mean and standard deviation were assigned as the reference mean and reference standard deviation. Then, 3DIV generated scaled bias-removed interaction frequency matrix by scaling sample mean and standard deviation to be the same as reference mean and standard deviation. The scaled Hi-C interaction frequency matrix is used only in the comparative visualization mode.

### **ChIP-seq data collection and processing**

In addition to Hi-C experiments, we processed 178 ChIP-seq experiments performed on 38 corresponding Hi-C samples [Table S2]. Raw sra files were downloaded from Short Read Archive (SRA) and converted to fastq files using sra-toolkit. ChIP-seq reads were aligned to human reference genome (hg19) using BWA-mem with -M option. Then, PCR duplicates were removed with Picard and poorly aligned reads (MAPQ<10) were discarded. The quality of ChIP-seq data was checked according to the ENCODE guideline, which includes on-redundant fraction, existence of input control, usable read depth, and read length. The ChIP-seq signal enrichment scores were measured at 100bp resolution, and then, the strongest enrichment score value within a 5kb bin was used to suggest the putative role of genomic loci involving interactions.

### **Enhancer and super-enhancer calling**

3DIV processed H3K27ac ChIP-seq data to identify enhancers and super-enhancers. H3K27ac peaks were identified using MACS2 (14), then ROSE (15) was applied to call enhancers and super-enhancers.

### **Implementation of 3DIV**

3DIV was implemented in a three-tiered architecture: data, logic, and presentation tiers (Figure S2).

The data tier integrates data from core databases holding Hi-C information and related data, such as SNPs and epigenomic data, with MySQL. Hi-C interaction data is stored as 12-column table containing the following information: chromosome, IDs of first and second bins, raw interaction frequency, coverage of first and second bins, genomic distance, expected bias, bias-removed interaction frequency, expected distance dependent background signal, distance-normalized interaction frequency, and rescaled interaction frequency. Data in 3DIV database was fully indexed and normalized to enable fast and precise search and retrieval.

The logic tier serves as an intermediary for data exchange between the presentation tier and the data tier. The logic tier was implemented using Java Spring Framework 3.1.1 with Java Development Kit (JDK) 1.6. MyBatis framework was used to connect the logic layer and its two surrounding layers.

The presentation tier occupies the front-end, and shows information related to services available on the website. The presentation tier was developed using Javascript. JQuery library enabled the user-interface design to be more interactive. The Ajax technology allows data to be smoothly loaded from 3DIV server. For all the graphics on 3DIV website, we used D3.js, a JavaScript library for producing dynamic, interactive data visualizations in web browsers.

1. Sloan, C.A., Chan, E.T., Davidson, J.M., Malladi, V.S., Strattan, J.S., Hitz, B.C., Gabdank, I., Narayanan, A.K., Ho, M., Lee, B.T. *et al.* (2016) ENCODE data at the ENCODE portal. *Nucleic Acids Res*, **44**, D726-732.
2. Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754-1760.
3. Dixon, J.R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J.E., Lee, A.Y., Ye, Z., Kim, A., Rajagopal, N., Xie, W. *et al.* (2015) Chromatin architecture reorganization during stem cell differentiation. *Nature*, **518**, 331.
4. Schmitt, A.D., Hu, M., Jung, I., Xu, Z., Qiu, Y., Tan, C.L., Li, Y., Lin, S., Lin, Y., Barr, C.L. *et al.* (2016) A compendium of chromatin contact maps reveals spatially active regions in the human genome. *Cell reports*, **17**, 2042-2059.
5. Hu, M., Deng, K., Selvaraj, S., Qin, Z., Ren, B. and Liu, J.S. (2012) HiCNorm: removing biases in Hi-C data via Poisson regression. *Bioinformatics*, **28**, 3131-3133.
6. 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.
7. Leung, D., Jung, I., Rajagopal, N., Schmitt, A., Selvaraj, S., Lee, A.Y., Yen, C.-A., Lin, S., Lin, Y., Qiu, Y. *et al.* (2015) Integrative analysis of haplotype-resolved epigenomes across human tissues. *Nature*, **518**, 350.
8. Imakaev, M., Fudenberg, G., McCord, R.P., Naumova, N., Goloborodko, A., Lajoie, B.R., Dekker, J. and Mirny, L.A. (2012) Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat Methods*, **9**, 999-1003.
9. Fraser, J., Ferrai, C., Chiariello, A.M., Schueler, M., Rito, T., Laudanno, G., Barbieri, M., Moore, B.L., Kraemer, D.C., Aitken, S. *et al.* (2015) Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation. *Mol Syst Biol*, **11**, 852.
10. Rao, S.S., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S. *et al.* (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, **159**, 1665-1680.
11. 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.
12. Lazaris, C., Kelly, S., Ntziachristos, P., Aifantis, I. and Tsirigos, A. (2017) HiC-bench: comprehensive and reproducible Hi-C data analysis designed for parameter exploration and benchmarking. *BMC Genomics*, **18**, 22.
13. Dali, R. and Blanchette, M. (2017) A critical assessment of topologically associating domain prediction tools. *Nucleic Acids Res*, **45**, 2994-3005.
14. Feng, J., Liu, T., Qin, B., Zhang, Y. and Liu, X.S. (2012) Identifying ChIP-seq enrichment using MACS. *Nature protocols*, **7**.
15. Whyte, W.A., Orlando, D.A., Hnisz, D., Abraham, B.J., Lin, C.Y., Kagey, M.H., Rahl, P.B., Lee, T.I. and Young, R.A. (2013) Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell*, **153**, 307-319.

## Supplementary Figure and Table legends for Yang D. Jang I. et al, 3DIV: A 3D-genome Interaction Viewer and Database

**Figure S1. Evaluation of negative binomial model-based implicit normalization.** (A) Reproducibility of Hi-C library using different restriction enzymes, HindIII and NcoI was measured along the distance between genomic loci. Irrelevant to the resolution of interaction frequency matrices, negative binomial model-based implicit normalization was highly reproducible, similar to other Hi-C processing pipelines such as HiCNorm and ICE. (B) Fraction of overlapping TAD boundaries called from normalized GM12878 *in-situ* Hi-C data using MboI restriction enzyme. Regardless of normalization methods, most of the TAD boundaries are detected from at least one other normalized interaction frequency matrices.

**Figure S2. Implementation diagram of 3DIV.** 3DIV consists of data, logic, and presentation tiers. MySQL-based data tier manages chromatin interaction data. Java Spring framework-based logic tier determines which subset of the database to be presented. JavaScript-based presentation tier provides the user-interface via web browser and presentation of chromatin interaction data.

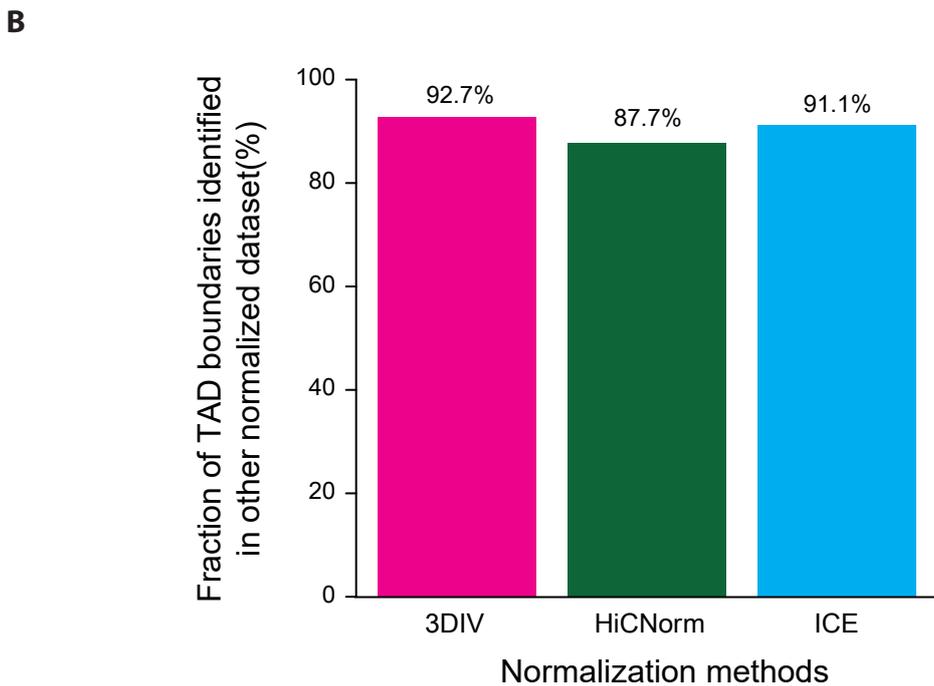
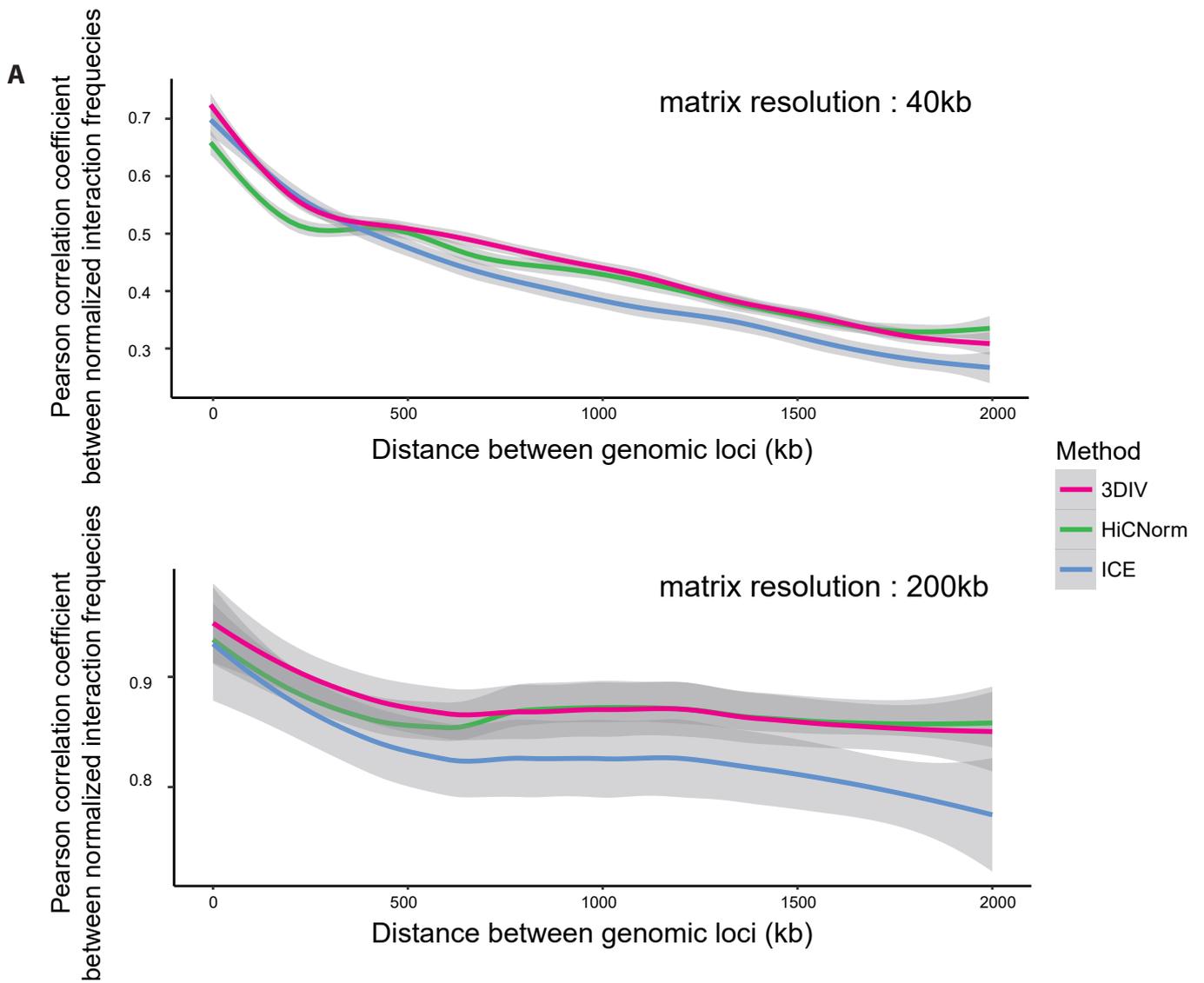
**Figure S3. 3DIV query panel and session management.** (A) The query for 3DIV is both the type of sample and the genomic locus. Experiment list loader button (red rectangle) loads the full list of samples for Hi-C experiments. To pass the query, users can click the check box (orange rectangle) for the desired sample, type genomic locus into bait box (yellow rectangle), then click “Add sample” button (green rectangle). Finally, clicking “Run” button (turquoise rectangle) passes the query to 3DIV. Users can alternate the TAD calling options by clicking the TAD calling combobox (blue rectangle). (B) Alternatively, users can upload session files to reproduce 3DIV results. Session manager button (purple rectangle) can save and load the session.

**Figure S4. 3DIV dynamic browsing system.** 3DIV allows flexible visualization by providing numerous adjustment options. (A) Each component of interaction visualization can be shown or hidden by clicking relevant buttons (red rectangle). Visualization range can be adjusted by clicking zoom in/out button (orange rectangle) or dragging the visualization range pane (orange square). Interaction heatmap color range scroll and resolution combobox (light green rectangle) adjust color range and resolution of interaction heatmap, respectively. (B) Fold change cut-off scroll (green rectangle) assigns distance-normalized interaction frequency cut-off. Dragging the white space (cyan circle) of interaction frequency graph shifts the visualization range. Dragging the red query bar reassign the query region. (C) Interaction descriptor (purple rectangle) provides descriptions for the selected chromatin interaction upon clicking the arc-represented interaction.

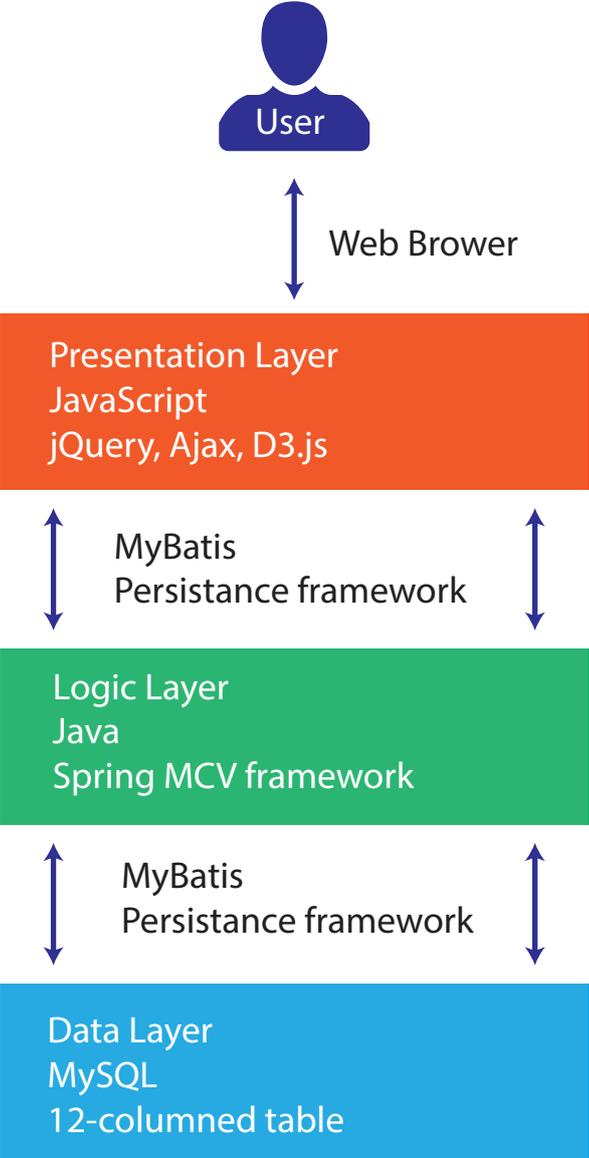
**Table S1. Statistics and quality of Hi-C chromatin interaction data.** Number of raw Hi-C read-pairs, aligned read-pairs, and valid *cis* read-pairs were measured, then the expected number of Hi-C interaction for each 40kb bin was used to represent the quality of Hi-C library.

**Table S2. Statistics and quality of histone ChIP-seq data.** Number of raw ChIP-seq reads and non-redundant aligned reads were measured. NRF (Non-Redundant Fraction), existence of input control, read depth, and read length were tested according to ENCODE guideline to represent the quality of ChIP-seq library.

Yang et al Figure S1



Yang et al Figure S2



Yang et al Figure S3

**A** **B**

Interaction table   Interaction visualization   Comparative interaction visualization

Choose Sample(s)

- Adrenal gland
- Aorta
- Bladder

▾

Bait :  (Ex. CROCCP2, chr22:27141000, rs42)

TAD :

Items selected

<input type="checkbox"/>	Sample	Bait
<input type="checkbox"/>		

Interaction range

Yang et al Figure S4

