Discovery of biased orientation of human DNA motif sequences affecting enhancer-promoter interactions and transcription of genes

Naoki Osato¹*

¹Department of Bioinformatic Engineering, Graduate School of Information Science and Technology, Osaka University, Osaka 565-0871, Japan

*Corresponding author

E-mail address: naokiosato11@gmail.com, nosato@ist.osaka-u.ac.jp
Abstract

Chromatin interactions have important roles for enhancer-promoter interactions (EPI) and regulating the transcription of genes. CTCF and cohesin proteins are located at the anchors of chromatin interactions, forming their loop structures. CTCF has insulator function limiting the activity of enhancers into the loops. DNA binding sequences of CTCF indicate their orientation bias at chromatin interaction anchors – forward-reverse (FR) orientation is frequently observed. DNA binding sequences of CTCF were found in open chromatin regions at about 40% - 80% of chromatin interaction anchors in Hi-C and in situ Hi-C experimental data. It has been reported that long range of chromatin interactions tends to include less CTCF at their anchors. It is still unclear what proteins are associated with chromatin interactions.

To find DNA binding motif sequences of transcription factors (TF) such as CTCF affecting the interaction between enhancers and promoters of transcriptional target genes and their expression, here, I predicted human transcriptional target genes of TF bound in open chromatin regions in enhancers and promoters in monocytes and other cell types using experimental data in public database. Transcriptional target genes were predicted based on enhancer-promoter association (EPA). EPA was shortened at the genomic locations of FR orientation of DNA binding motifs of TF, which were supposed to be at chromatin interaction anchors.

The expression level of the target genes predicted based on EPA was compared with target genes predicted from only promoters. Total 351 biased orientation of DNA motifs (192 forward-reverse (FR) and 179 reverse-forward (RF) orientation, the reverse
complement sequences of some DNA motifs are also registered in databases, so the total number is smaller than the number of FR and RF) affected the expression level of putative transcriptional target genes significantly in monocytes of four people in common. Moreover, EPI predicted using FR or RF orientation of some DNA motifs were overlapped with chromatin interaction data (Hi-C) more than the other EPA (Total 62 biased orientation of DNA motifs, 41 FR and 24 RF showed this result).

**Keywords:** transcriptional target genes, gene expression, transcription factors, enhancer, enhancer-promoter interactions, chromatin interactions, CTCF, cohesion, open chromatin regions

**Background**

Chromatin interactions have important roles for enhancer-promoter interactions (EPI) and regulating the transcription of genes. CTCF and cohesin proteins are located at the anchors of chromatin interactions, forming their loop structures. CTCF has insulator function limiting the activity of enhancers into the loops (Fig. 1A). DNA binding sequences of CTCF indicate their orientation bias at chromatin interaction anchors – forward-reverse (FR) orientation is frequently observed (de Wit et al. 2015; Guo et al. 2015). About 40% - 80% of chromatin interaction anchors of Hi-C and in situ Hi-C experiments include DNA binding motif sequences of CTCF. However, it has been reported that long range of chromatin interactions tends to include less CTCF at their anchors (Jeong et al. 2017). Other DNA binding proteins such as ZNF143, YY1, and
SMARCA4 (BRG1) are found to be associated with chromatin interactions and EPI (Bailey et al. 2015; Barutcu et al. 2016; Weintraub et al. 2017). CTCF, cohesin, ZNF143, YY1 and SMARCA4 have other biological functions as well as chromatin interactions and enhancer-promoter interactions. The DNA binding motif sequences of the transcription factors are found in open chromatin regions near transcriptional start sites (TSS) as well as chromatin interaction anchors.

DNA binding motif sequence of ZNF143 was enriched at both chromatin interaction anchors. ZNF143’s correlation with the CTCF-cohesin cluster relies on its weakest binding sites, found primarily at distal regulatory elements defined by the ‘CTCF-rich’ chromatin state. The strongest ZNF143-binding sites map to promoters bound by RNA polymerase II (POL2) and other promoter-associated factors, such as the TATA-binding protein (TBP) and the TBP-associated protein, together forming a ‘promoter’ cluster (Bailey et al. 2015).

DNA binding motif sequence of YY1 does not seem to be enriched at both chromatin interaction anchors (Z-score < 2), whereas DNA binding motif sequence of ZNF143 is significantly enriched (Z-score > 7; Bailey et al. 2015 Figure 2a). In the analysis of YY1, to identify a protein factor that might contribute to EPI, (Ji et al. 2015) performed chromatin immune precipitation with mass spectrometry (ChIP-MS), using antibodies directed toward histones with modifications characteristic of enhancer and promoter chromatin (H3K27ac and H3K4me3, respectively). Of 26 transcription factors that occupy both enhancers and promoters, four are essential based on a CRISPR cell-essentiality screen and two (CTCF, YY1) are expressed in >90% of tissues.
examined (Weintraub et al. 2017). These analyses started from the analysis of histone modifications of an enhancer mark rather than chromatin interactions. Other protein factors associated with chromatin interactions may be found from other researches.

As computational approaches, machine-learning analyses to predict chromatin interactions have been proposed (Schreiber et al. 2017; Zhang et al. 2017). However, they were not intended to find DNA motif sequences of transcription factors affecting chromatin interactions, EPI, and the expression level of transcriptional target genes, which were examined in this study.

DNA binding proteins involved in chromatin interactions are supposed to affect the transcription of genes in the loop formed by chromatin interactions. In my previous analyses, the expression level of human putative transcriptional target genes was affected, according to the criteria of enhancer-promoter association (EPA) (Fig. 1B; (Osato 2018)). EPI predicted based on EPA shortened at the genomic locations of forward-reverse orientation of CTCF binding sites affected the expression level of putative transcriptional target genes the most, compared with the expression level of transcriptional target genes predicted from only promoters (Fig. 1B). The expression level tended to be increased in monocytes and CD4$^+$ T cells, implying that enhancers activated the transcription of genes, and decreased in ES and iPS cells, implying that enhancers repressed the transcription of genes. These analyses suggested that enhancers affected the transcription of genes significantly, when EPI were predicted properly. Other DNA binding proteins involved in chromatin interactions, as well as CTCF, would locate at chromatin interaction anchors with a pair of biased orientation of DNA
binding motif sequences, affecting the expression level of putative transcriptional target genes in the loop formed by the chromatin interactions. As experimental issues of the analyses of chromatin interactions, chromatin interaction data are changed, according to experimental techniques, depth of DNA sequencing, and even replication sets of the same cell type. Chromatin interaction data may not be saturated enough to cover all chromatin interactions. Supposing these characters of DNA binding proteins associated with chromatin interactions and avoiding experimental issues of the analyses of chromatin interactions, here I searched for DNA motif sequences of transcription factors and DNA repeat sequences, affecting EPI and the expression level of putative transcriptional target genes in monocytes of four people and other cell types without using chromatin interaction data. Then, putative EPI were compared with chromatin interaction data.

**Results**

**Search for biased orientation of DNA motif sequences**

Transcription factor binding sites (TFBS) were predicted using open chromatin regions and DNA motif sequences of transcription factors collected from various databases and journal papers (see Methods). Transcriptional target genes were predicted using TFBS in promoters and enhancer-promoter association (EPA) shortened at the genomic locations of DNA binding motif sequences of transcription factors such as CTCF and cohesins (RAD21 and SMC3). To find DNA motif sequences of transcription factors, other than CTCF and cohesin, and DNA repeat sequences affecting the
expression level of putative transcriptional target genes, transcriptional target genes of each transcription factor were predicted based on EPA shortened at the genomic locations of DNA motif sequences, and the expression level of the putative transcriptional target genes was compared with those predicted from promoters using Mann Whitney U test ($p$-value < 0.05). The number of transcription factors showing a significant difference of expression level of putative transcriptional target genes was counted. To examine whether the orientation of DNA motif sequences shortening EPA affected the number of transcription factors showing a significant difference of expression level of putative transcriptional target genes, the number of the transcription factors was compared among forward-reverse, reverse-forward, and any orientation of DNA motif sequences shortening EPA, using chi-square test ($p$-value < 0.05). To avoid missing DNA motif sequences showing a relatively weak statistical significance by multiple testing collection, the above analyses were conducted using monocytes of four people independently, and DNA motif sequences found in monocytes of four people in common were selected. Total 351 of biased (192 forward-reverse and 179 reverse-forward) orientation of DNA binding motif sequences of transcription factors have been found in monocytes of four people in common, whereas only one any orientation (i.e. without considering orientation) of DNA binding motif sequences have been found in monocytes of four people in common (Fig. 2; Table 1; Supplemental Table S1). Forward-reverse orientation of DNA motif sequences included CTCF, cohesin (RAD21 and SMC3), ZNF143 and YY1, which are associated with chromatin interactions and enhancer-promoter interactions (EPI). SMARCA4 (BRG1) is
associated with topologically associated domain (TAD), which is higher-order chromatin organization. The DNA binding motif sequences of SMARCA4 was not registered in the databases used in this study. Forward-reverse orientation of DNA motif sequences included SMRC2, a member of the same SWI/SNF family of proteins as SMARCA4. These results suggested that the expression level of putative transcription target genes of some transcription factors would be different, depending on the genomic locations (enhancers or promoters) of DNA binding motif sequences of the transcription factors in monocytes of four people.

To examine whether the biased orientation of DNA binding motif sequences of transcription factors was observed in other cell types, the same analyses were conducted in other cell types. However, for other cell types, experimental data of one sample were available in ENCODE database, so the analyses of DNA motif sequences were performed by comparing with the result in monocytes of four people. Among the biased orientation of DNA binding motif sequences found in monocytes, 79, 76, 91, 61, and 59 DNA binding motif sequences were also observed in CD4+ T cells, H1-hESC, iPS, Huvec and MCF-7 respectively, including CTCF and cohesin (RAD21 and SMC3) (Table 2; Supplemental Table S2). The scores of DNA binding motif sequences were the highest in monocytes, and the other cell types showed lower scores. The results of the analysis of DNA motif sequences in CD20+ B cells and macrophages did not include CTCF and cohesin, because these analyses can be utilized in cells where the expression level of putative transcriptional target genes of each transcription factor show a significant difference between promoters and EPA shortened at the genomic locations of
DNA motif sequences. Some experimental data of a cell did not show a significant
difference between promoters and EPA (Osato 2018).

Instead of DNA binding motif sequences of transcription factors, DNA repeat
sequences were also examined. The expression level of transcriptional target genes
predicted based on EPA shortened at the genomic locations of DNA repeat sequences
was compared with the expression level of transcriptional target genes predicted from
promoters. Three reverse-forward orientation of DNA repeat sequences showed a
significant difference of expression level of putative transcriptional target genes in
monocytes of four people in common (Table 3). Among them, LTR16C DNA repeat
sequence was observed in iPS and H1-hESC.

**Co-location of biased orientation of DNA motif sequences**

To examine association of the 351 biased (forward-reverse and reverse-forward)
orientation of DNA binding motif sequences of transcription factors. Co-location of the
DNA binding motif sequences in open chromatin regions in monocytes was analyzed.
The number of open chromatin regions with the same pairs of DNA binding motif
sequences was counted, and when the pairs of DNA binding motif sequences were
enriched with statistical significance (chi-square test, \( p \)-value < 1.0 \( \times \) \( 10^{-10} \)), they were
listed (Table 4; Supplemental Table S3). Open chromatin regions overlapped with
histone modification of an enhancer mark (H3K27ac) (total 26,095 regions) showed a
larger number of enriched pairs of DNA motifs than all open chromatin regions (Table
4; Supplemental Table S3). H3K27ac is known to be enriched at chromatin interaction
anchors (Phanstiel et al. 2017). As already known, CTCF was found with cohesin such as RAD21 and SMC3 (Table 4). Top 30 pairs of FR and RF orientations of DNA motifs co-occupied in the same open chromatin regions were shown (Table 4). Total number of pairs of DNA motifs was 601, consisting of 155 kinds of DNA motifs, when the pair of DNA motifs were observed in more than 80% of the number of open chromatin regions with the DNA motifs.

**Comparison with chromatin interaction data**

To examine whether the biased orientation of DNA motif sequences is associated with chromatin interactions, enhancer-promoter interactions (EPI) predicted based on enhancer-promoter associations (EPA) were compared with chromatin interaction data (Hi-C). Due to the resolution of Hi-C experimental data used in this study (50kb), EPI were adjusted to 50kb resolution. EPI were predicted based on three types of EPA: (i) EPA without being shortened by DNA motif sequences, (ii) EPA shortened at the genomic locations of DNA motif sequences of transcription factors such as CTCF and cohesin (RAD21, SMC3) without considering their orientation, and (iii) EPA shortened at the genomic locations of forward-reverse or reverse-forward orientation of DNA motif sequences of transcription factors. EPA (iii) showed the highest ratio of EPI overlapped with chromatin interactions (Hi-C) using DNA binding motif sequences of CTCF and cohesin (RAD21 and SMC3), compared with the other two types of EPA (Figure 3). Total 62 biased orientation (41 FR and 24 RF) of DNA motif sequences including CTCF and cohesin showed a higher ratio of EPI overlapped with Hi-C than...
the other types of EPA in monocytes (Table 5). Chromatin interaction data were obtained using different samples from DNase-seq, open chromatin regions, so individual differences seemed to be large from the results of this analysis. Since, for some DNA motif sequences of transcription factors, the number of EPI overlapped with chromatin interactions was small, if higher resolution of chromatin interaction data (such as Hi-ChIP, in situ DNase Hi-C, and in situ Hi-C data, or a tool to improve the resolution such as HiCPlus) is available, the number of EPI overlapped with chromatin interactions would be increased and the difference of the numbers among three types of EPA might be larger and more significant than the current results (Rao et al. 2014; Ramani et al. 2016; Mumbach et al. 2017; Zhang et al. 2018).

Discussion

To find DNA motif sequences of transcription factors and DNA repeat sequences affecting human putative transcriptional target genes, DNA motif sequences were searched from open chromatin regions of monocytes of four people. Total 351 biased (192 forward-reverse and 179 reverse-forward) orientation of DNA motif sequences were found in monocytes of four people in common, whereas only one any orientation (i.e. without considering orientation) of DNA motif sequence was found to affect putative transcriptional target genes, suggesting that EPA shortened at the genomic locations of forward-reverse or reverse-forward orientation of DNA motif sequences of transcription factors is an important character for EPI and transcription of genes.

When DNA motif sequences were searched from monocytes of one person, a
larger number of biased orientation of DNA motif sequences affecting human putative
transcriptional target genes has been found. When the number of people, from which
experimental data were obtained, was increased, the number of DNA motif sequences
found in all people in common decreased and in some cases, known transcription
factors involved in chromatin interactions such as CTCF and cohesin (RAD21 and
SMC3) were not identified by statistical tests. This would be caused by individual
difference of the same cell type, low quality of experimental data, and experimental
errors. Moreover, though forward-reverse orientation of DNA binding motif sequences
of CTCF and cohesin is frequently observed at chromatin interaction anchors, the
percentage of forward-reverse orientation is not 100, and other orientations of the DNA
binding motif sequences are also observed. Though DNA binding motif sequences of
CTCF and cohesins are found in various open chromatin regions, DNA binding motif
sequences of some transcription factors would be observed less frequently in open
chromatin regions. The analyses of cells of a number of people would avoid missing
relatively weak statistical significance of DNA motif sequences in cells of each person
by multiple testing correction of thousands of statistical tests. A DNA motif sequence
was found with $p$-value $< 0.05$ in cells of one person and the DNA motif sequence
found in the same cell type of four people in common would have $p$-value $< (0.05)^4 =
0.00000625$. Actually, DNA motif sequences with $p$-value slightly less than 0.05 in
monocytes of one person were observed in monocytes of four people in common.

EPI were compared with chromatin interactions (Hi-C) in monocytes. EPA
shortened at the genomic locations of DNA binding motif sequences of CTCF and
cohesin (RAD21 and SMC3) showed a significant difference of the ratios of EPI overlapped with chromatin interactions, according to three types of EPA (FR, any, and others). Using open chromatin regions overlapped with ChIP-seq experimental data of histone modification of an enhancer mark (H3K27ac), the ratio of EPI not overlapped with Hi-C was reduced and the difference of the ratio of EPI overlapped with chromatin interactions among the three types of EPA became larger (data not shown). (Phanstiel et al. 2017) also reported that there was an especially strong enrichment for loops with H3K27 acetylation peaks at both ends (Fisher’s Exact Test, $p = 1.4 \times 10^{-27}$). However, the total number of EPI overlapped with chromatin interactions was also reduced using H3K27ac peaks, so more chromatin interaction data would be needed to obtain reliable results in this study. As an issue of experimental data, chromatin interaction data were obtained using different samples from DNase-seq, open chromatin regions, so individual differences were not reflected to chromatin interaction data. Moreover, the resolution of chromatin interaction data used in this study is about 50kb, thus the number of chromatin interactions was relatively small (72,284 at 50kb resolution with cutoff score of CHiCAGO tool > 1). EPA shortened at the genomic locations of DNA binding motif sequences of transcription factors found in various open chromatin regions such as CTCF and cohesin (RAD21 and SMC3) tend to be overlapped with a larger number of chromatin interactions than DNA binding motif sequences of transcription factors less frequently observed in open chromatin regions. Therefore, to examine the difference of the numbers of EPI overlapped with chromatin interactions, according to three types of EPA (FR, any, and others), the number of chromatin
interactions should be large enough. If the resolution of chromatin interaction data
would be increased by using other experimental data such as Hi-ChIP, in situ DNase
Hi-C, and in situ Hi-C data, or a tool to improve the resolution of chromatin interaction
data such as HiCPlus, more DNA binding motif sequences of transcription factors less
frequently observed in open chromatin regions would be examined (Rao et al. 2014;
Ramani et al. 2016; Mumbach et al. 2017; Zhang et al. 2018).

Methods

Search for biased orientation of DNA motif sequences

To examine transcriptional regulatory target genes, bed files of hg38 of Blueprint
DNase-seq data for Monocytes CD14⁺ (EGAD00001002286) were obtained from
Blueprint web site (http://dcc.blueprint-epigenome.eu/#/home), and the bed files of
hg38 were converted into those of hg19 using Batch Coordinate Conversion (liftOver)
web site (https://genome.ucsc.edu/util.html). Bed files of hg19 of ENCODE
CD4⁺_Naive_Wb11970640 (GSM1014537; UCSC Accession: wgEncodeEH003156),
H1-hESC (GSM816632; UCSC Accession: wgEncodeEH000556), iPSC (GSM816642;
UCSC Accession: wgEncodeEH001110), HUVEC (GSM1014528; UCSC Accession:
wgEncodeEH002460), and MCF-7 (GSM816627; UCSC Accession:
wgEncodeEH000579) were obtained from the ENCODE websites
(http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeUwDgf/; http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeUwDnase/).

To identify transcription factor binding sites (TFBS) from the DNase-seq data,
TRANSFAC (2013.2), JASPAR (2010), UniPROBE, BEEMI-PBM, high-throughput SELEX, Human Protein-DNA Interactome, and transcription factor binding sequences of ENCODE ChIP-seq data were used (Wingender et al. 1996; Newburger and Bulyk 2009; Portales-Casamar et al. 2010; Xie et al. 2010; Zhao and Stormo 2011; Jolma et al. 2013; Kheradpour and Kellis 2014). Position weight matrices of transcription factor binding sequences were transformed into TRANSFAC matrices and then into MEME matrices using in-house scripts and transfac2meme in MEME suite (Bailey et al. 2009). Transcription factor binding sequences of transcription factors derived from vertebrates were used for further analyses. Searches were conducted for transcription factor binding sequences from each narrow peak using FIMO with \( p \)-value threshold of \( 10^{-5} \) (Grant et al. 2011). Transcription factors corresponding to transcription factor binding sequences were searched computationally by comparing their names and gene symbols of HGNC (HUGO Gene Nomenclature Committee) -approved gene nomenclature and 31,848 UCSC known canonical transcripts (http://hgdownload.cse.ucsc.edu/goldenpath/hg19/database/knownCanonical.txt.gz), as transcription factor binding sequences were not linked to transcript IDs such as UCSC, RefSeq, and Ensembl transcripts.

Target genes of a transcription factor were assigned when its TFBS was found in DNase-seq narrow peaks in promoter or extended regions for enhancer-promoter association of genes (EPA). Promoter and extended regions were defined as follows: promoter regions were those that were within distances of ±5 kb from transcriptional start sites (TSS). Promoter and extended regions were defined as per the following
association rule, which is the same as that defined in Figure 3A of a previous study (McLean et al. 2010): the single nearest gene association rule, which extends the regulatory domain to the midpoint between the TSS of the gene and that of the nearest gene upstream and downstream without the limitation of extension length. Extended regions for EPA were shortened at the genomic locations of DNA binding sites of transcription factors that were the closest to a transcriptional start site, and transcriptional target genes were predicted from the shortened enhancer regions using TFBS. Furthermore, promoter and extended regions for EPA were shortened at the genomic locations of forward–reverse orientation of DNA binding sites of transcription factors. When forward or reverse orientation of DNA binding sites were continuously located in genome sequences several times, the most external forward–reverse orientation of DNA binding sites were selected. The genomic positions of genes were identified using ‘knownGene.txt.gz’ file in UCSC bioinformatics sites (Karolchik et al. 2014). The file ‘knownCanonical.txt.gz’ was also utilized for choosing representative transcripts among various alternate forms for assigning promoter and extended regions for EPA. From the list of transcription factor binding sequences and transcriptional target genes, redundant transcription factor binding sequences were removed by comparing the target genes of a transcription factor binding sequence and its corresponding transcription factor; if identical, one of the transcription factor binding sequences was used. When the number of transcriptional target genes predicted from a transcription factor binding sequence was less than five, the transcription factor binding sequence was omitted.
For gene expression data, RNA-seq reads mapped onto human hg19 genome sequences were obtained, including ENCODE long RNA-seq reads with poly-A of H1-hESC, iPSC, HUVEC, IMR90, and MCF-7 (GSE26284, GSM958733, GSM2344099, GSM2344100, GSM958734, and GSM765388), and UCSF-UBC human reference epigenome mapping project RNA-seq reads with poly-A of naive CD4\(^+\) T cells (GSM669617). Two replicates were present for H1-hESC, iPSC, HUVEC, and MCF-7, and a single one for CD4\(^+\) T cells. RPKMs of the RNA-seq data were calculated using RSeQC (Wang et al. 2012). For monocytes, Blueprint RNA-seq RPKM data (GSE58310, GSE58310_GeneExpression.csv.gz, Monocytes_Day0_RPMI) was used (Saeed et al. 2014). Based on RPKM, UCSC transcripts with expression levels among top 30% of all the transcripts were selected in each cell type.

The expression level of transcriptional target genes predicted based on EPA shortened at the genomic locations of DNA motif sequences of transcription factors or DNA repeat sequences was compared with the expression level of transcriptional target genes predicted from promoter. For each DNA motif sequence shortening EPA, transcriptional target genes were predicted using about 3,000 – 5,000 DNA binding motif sequences of transcription factors, and the expression level of putative transcriptional target genes of each transcription factor was compared between EPA and only promoter using Mann-Whitney test ($p$-value < 0.05). The number of transcription factors showing a significant difference of expression level of putative transcriptional target genes between EPA and promoter was compared among forward-reverse, reverse-forward, and any orientation of DNA motif sequences shortening EPA using
chi-square test ($p$-value < 0.05). When DNA motif sequences of transcription factors or DNA repeat sequences shortening EPA showed a significant difference of expression level of putative transcriptional target genes in forward-reverse, reverse-forward, or any orientation in monocytes of four people in common, the DNA motif sequences were listed.

Though forward-reverse orientation of DNA binding motif sequences of CTCF and cohesin is frequently observed at chromatin interaction anchors, the percentage of forward-reverse orientation is not 100, and other orientations of the DNA binding motif sequences are also observed. Though DNA binding motif sequences of CTCF and cohesin are found in various open chromatin regions, DNA binding motif sequences of some transcription factors would be observed less frequently in open chromatin regions. The analyses of cells of a number of people would avoid missing relatively weak statistical significance of DNA motif sequences in cells of each person by multiple testing correction of thousands of statistical tests. A DNA motif sequence was found with $p$-value < 0.05 in cells of one person and the DNA motif sequence found in the same cell type of four people in common would have $p$-value < $0.05^4 = 0.00000625$.

Co-location of biased orientation of DNA motif sequences

Co-location of biased orientation of DNA binding motif sequences of transcription factors was examined. The number of open chromatin regions with the same pairs of DNA binding motif sequences was counted, and when the pairs of DNA binding motif sequences were enriched with statistical significance (chi-square test,
For chromatin interactions, ‘PCHiC_peak_matrix_cutoff0.txt.gz’ file was downloaded from ‘Promoter Capture Hi-C in 17 human primary blood cell types’ web site (https://osf.io/u8tzw/files/), and chromatin interactions for Monocytes with scores of CHiCAGO tool > 1 were extracted from the file (Javierre et al. 2016).

EPI were predicted using three types of EPA in monocytes: (i) EPA shortened at the genomic locations of forward-reverse or reverse-forward orientation of DNA motif sequences of transcription factors, (ii) EPA shortened at the genomic locations of any orientation (i.e. without considering orientation) of DNA motif sequences of transcription factors, and (iii) EPA without being shortened by DNA motif sequences. EPI predicted using the three types of EPA in common were removed. First, EPI predicted based on EPA (i) were compared with chromatin interactions (Hi-C). The resolution of chromatin interaction data used in this study was 50kb, so EPI were adjusted to 50kb, before their comparison. The number and ratio of EPI overlapped with chromatin interactions were counted. Second, EPI were predicted based on EPA (ii), and EPI predicted based on EPA (i) were removed from the EPI. The number and ratio of EPI overlapped with chromatin interactions were counted. Third, EPI were predicted based on EPA (iii), and EPI predicted based on EPA (i) and (ii) were removed from the EPI. The number and ratio of EPI overlapped with chromatin interactions were counted.
The number and ratio of the EPI compared two times between EPA (i) and (iii), and EPA (i) and (ii) (binomial distribution, $p$-value < 0.025 for each test).

Acknowledgements

The supercomputing resource was provided by Human Genome Center of the Institute of Medical Science at the University of Tokyo. Computations were partially performed on the NIG supercomputer at ROIS National Institute of Genetics. Publication charges for this article were funded by JSPS KAKENHI Grant Number 16K00387. This research was partially supported by the Platform Project for Supporting in Drug Discovery and Life Science Research (Platform for Dynamic Approaches to Living System) from Japan Agency for Medical Research and Development (AMED). This research was partially supported by Development of Fundamental Technologies for Diagnosis and Therapy Based upon Epigenome Analysis from Japan Agency for Medical Research and Development (AMED). This work was partially supported by JST CREST Grant Number JPMJCR15G1, Japan.

Figures
Figure 1. Chromatin interactions and enhancer-promoter association. (A) Forward–reverse orientation of CTCF-binding sites is frequently found in chromatin interaction anchors. CTCF can block the interaction between enhancers and promoters limiting the activity of enhancers to certain functional domains (de Wit et al. 2015; Guo et al. 2015). (B) Computationally-defined regulatory domains for enhancer-promoter association (McLean et al. 2010). The single nearest gene association rule extends the regulatory domain to the midpoint between this gene’s TSS and the nearest gene’s TSS both upstream and downstream. Enhancer-promoter association was shortened at the genomic locations of forward-reverse orientation of DNA binding motif sequences of transcription factors (e.g. CTCF in the figure).
Figure 2. Biased-orientation of DNA motif sequences of transcription factors. Total 351 of biased orientation of DNA binding motif sequences of transcription factors have been found to affect the expression level of putative transcriptional target genes in monocytes of four people in common, whereas only one any orientation (i.e. without considering orientation) of DNA binding motif sequences have been found to affect the expression level.

Figure 3. Comparison of DNA binding motif sequences of transcription factors with chromatin interactions. Total 62 biased orientation (41 FR and 24 RF) of DNA motif sequences including CTCF and cohesin showed a higher ratio of EPI overlapped with Hi-C than the other types of EPA in monocytes. FR: forward-reverse orientation of DNA motifs. any: any orientation of DNA motifs. others: enhancer-promoter association not shortened at the genomic locations of DNA motifs. * p-value < 10^{-5}.

Tables
### Table 1. Top 90 biased orientation of DNA binding motif sequences of transcription factors in monocytes. TF: DNA binding motif sequences of transcription factors. Score: –log$_{10}$ (p-value).

#### Forward-reverse orientation

| TF       | Score | TF       | Score | TF       | Score |
|----------|-------|----------|-------|----------|-------|
| SRF      | 75.75 | YLL054C  | 34.01 | NR2F2    | 25.16 |
| IRF7     | 71.35 | ZIC1     | 33.85 | PLAG1    | 24.95 |
| MIN20    | 61.98 | SMG3     | 33.67 | REST     | 24.83 |
| CTCF     | 61.90 | ESR2     | 32.37 | GAL4P    | 24.48 |
| EIF1     | 59.63 | TP63     | 32.28 | ZNF143   | 23.97 |
| STAT1    | 59.24 | ZNF836   | 31.96 | HLH2_HLH3| 23.85 |
| ZNF317   | 56.27 | STAGA    | 31.54 | BM9      | 23.63 |
| CEBPG    | 55.76 | MYB      | 31.23 | YB1      | 23.29 |
| ZNF123   | 55.01 | AT3G16280| 31.22 | PRDM15   | 22.93 |
| ZNF93    | 54.53 | E2F8     | 30.92 | ZNF93    | 22.62 |
| HMBOX1   | 53.29 | P53      | 30.86 | CRRF1    | 22.42 |
| SMAD3_SMAD3_SMAD4 | 50.96 | STAGB    | 30.24 | ZNF214   | 22.41 |
| FOXN2    | 50.18 | STAG3    | 30.15 | PTF1A    | 22.39 |
| CLAMP    | 49.82 | AP1      | 30.05 | SMRC2    | 22.16 |
| RDR1     | 46.44 | AT3G04100| 29.77 | ZNF695   | 22.08 |
| ZNF846   | 44.25 | CEBPZ    | 29.73 | ZNF28    | 22.03 |
| TF3      | 42.76 | RAD21    | 28.98 | ZNF316   | 21.96 |
| CMYB     | 42.28 | GATA2    | 28.82 | TATA     | 21.90 |
| KR       | 41.45 | SMAD2    | 28.81 | NF1B     | 21.75 |
| STAT4    | 40.84 | KLF6     | 28.08 | PHX2B    | 21.12 |
| ZNF16    | 40.69 | HXB1     | 27.93 | ATB1     | 20.96 |
| KLF14    | 39.94 | GL1      | 27.39 | DEC2     | 20.92 |
| DREB2E   | 39.93 | GCMA     | 27.03 | ZNF709   | 20.88 |
| FOXA1    | 38.43 | ZNF676   | 26.96 | YY1      | 20.87 |
| USF      | 38.04 | E47      | 26.62 | EIF5A2   | 20.85 |
| HME1     | 37.31 | EHF      | 26.62 | ZNF148   | 20.75 |
| ZNF19    | 35.67 | TCF12    | 26.59 | PAZ5     | 20.35 |
| D        | 35.16 | RSPOC4   | 26.53 | POF1     | 20.24 |
| IKZF1    | 34.09 | HIC1     | 25.64 | FOXD4    | 19.97 |
| YBR033W  | 34.01 | EGR2     | 25.23 | EBF      | 19.56 |

#### Reverse-forward orientation
Table 2. Top 30 forward-reverse orientation of DNA binding motif sequences of transcription factors in MCF-7 and iPS. TF: DNA binding motif sequences of transcription factors. Score: $-\log_{10}(p$-value).

| TF      | Score | TF          | Score | TF      | Score |
|---------|-------|-------------|-------|---------|-------|
| RFX5    | 102.60| GABP1_GABP2 | 36.27 | WT1     | 29.37 |
| TCF1    | 75.65 | RDS2        | 35.16 | E47     | 29.16 |
| SRF     | 65.83 | TCFA2B      | 34.85 | CREM    | 28.95 |
| ZNF28   | 64.19 | TBP         | 34.77 | ZNF31   | 28.64 |
| TCF2    | 62.76 | ZNF670      | 34.63 | HAND1_E47 | 28.37 |
| ZNF525  | 62.48 | ZBRK1       | 34.60 | ZNF449  | 28.26 |
| CEBPB   | 59.99 | ZNF687      | 34.12 | ZNF233  | 27.97 |
| NFX     | 59.26 | NFKB2       | 33.35 | SOX2    | 26.77 |
| ZNF286A | 55.03 | RXF3        | 33.04 | RORA    | 26.37 |
| CTCF    | 53.46 | NRF3        | 32.99 | MYF6    | 26.15 |
| PAX5    | 52.87 | VND         | 32.75 | MAFK    | 26.12 |
| STAT1   | 52.41 | RNF96       | 32.57 | CEBPZ   | 26.03 |
| ISL1    | 50.03 | SP3         | 32.52 | PO2F2   | 25.89 |
| ETV7    | 47.38 | FLI1        | 32.45 | MADS10  | 25.62 |
| KLF18   | 46.75 | RB4K        | 32.22 | CETS2   | 25.35 |
| SP1     | 46.33 | NRF2F1      | 32.10 | HSF1    | 25.06 |
| SMAD3   | 45.57 | NKK3        | 31.85 | GCN4P   | 24.53 |
| STAT5B  | 45.02 | ZIC2        | 31.68 | CRZ1    | 24.46 |
| TEAD1   | 43.30 | RAFR1       | 31.49 | AHR     | 23.86 |
| ZP679   | 43.25 | ZNF195      | 31.36 | CRZ1P   | 23.82 |
| ZBTB2   | 42.58 | BD1P        | 31.14 | PAX6    | 23.78 |
| ZNF343  | 41.31 | SOX9        | 31.03 | TOE3    | 23.67 |
| DOBOX5  | 39.35 | SPB         | 30.79 | IRF4    | 22.77 |
| NANOG   | 39.14 | EMBP1B      | 30.62 | RAP1    | 22.54 |
| LEU3    | 38.92 | ZNF121      | 30.27 | ETV3    | 22.37 |
| NRF2    | 38.83 | ZFP202      | 30.31 | FOXP2   | 21.97 |
| OBOX2   | 38.30 | CD59        | 30.15 | TATA    | 21.93 |
| ZNF682  | 38.14 | RORB        | 29.87 | CDX2    | 21.75 |
| ZNF30   | 37.89 | STAT5A      | 29.61 | NFAC1   | 21.54 |
| MYOD1   | 36.94 | SOX11       | 29.39 | SP4     | 21.44 |

MCF-7  
iPS
Table 3. Biased orientation of DNA repeat sequences in monocytes. Score: $-\log_{10} (p$-value).

Table 4. Top 30 of co-locations of biased orientation of DNA binding motif sequences of transcription factors in monocytes. Co-locations of DNA motif sequences of CTCF with other biased orientation of DNA motif sequences were shown in the separate table.

Motif 1,2: DNA binding motif sequences of transcription factors. # overlap: the number
of open chromatin regions including both Motif 1 and Motif 2. # motif 1: the number of open chromatin regions including Motif 1. # motif 2: the number of open chromatin regions including Motif 2. # others: the number of open chromatin regions not including Motif 1 or Motif 2. Open chromatin regions overlapped with histone modification (H3K27ac) were used (Total 26,095 regions).

Table 5. Comparison of DNA binding motif sequences of transcription factors with chromatin interaction data. TF: DNA binding motif sequences of transcription factors.
Score: $-\log_{10}(p\text{-value})$.

### Forward-reverse orientation

| TF    | Score 1 | Score 2 |
|-------|---------|---------|
| AP1   | 28.64   | 3.75    |
| ATF3G04100 | 24.51 | 7.07    |
| BTEB2 | 8.64    | 2.69    |
| CTCF  | 6.09    | 6.80    |
| E2F   | 19.43   | 3.24    |
| E2F5  | 10.58   | 2.02    |
| EGR2  | 4.28    | 10.11   |
| ETS   | 7.66    | 5.74    |
| FOISB | 7.98    | 1.99    |
| FOXO4 | 7.98    | 5.84    |
| GCM1  | 2.51    | 3.47    |
| GCR   | 6.39    | 4.50    |
| IRF7  | 16.36   | 2.03    |
| KLF14 | 18.52   | 8.08    |
| KLF3  | 6.42    | 6.43    |
| MAZ   | 1.82    | 4.50    |
| MG2   | 42.64   | 6.22    |
| PARG  | 5.38    | 4.21    |
| PITX3 | 10.31   | 5.62    |
| PLAL1 | 11.50   | 5.42    |
| PO6F1 | 31.28   | 7.60    |
| PTF1A | 8.81    | 4.50    |
| PU1   | 3.86    | 1.71    |
| RAD21 | 9.28    | 5.16    |
| RPEB1 | 7.47    | 6.40    |
| RUNX2 | 2.97    | 1.68    |
| SETB1 | 1.63    | 2.08    |
| SMAD2 | 4.10    | 3.48    |
| SMC3  | 10.24   | 6.21    |
| SP1   | 20.98   | 4.64    |
| TCFAP2A | 5.10 | 1.63    |
| TCF3B | 3.21    | 5.75    |
| TFC3B | 20.26   | 4.66    |
| TFE3  | 94.90   | 3.64    |
| YY1   | 13.88   | 7.04    |
| ZBTB6 | 15.08   | 3.93    |
| ZIC1  | 5.37    | 1.63    |
| ZNF148| 14.51   | 4.47    |
| ZNF219| 15.44   | 7.99    |
| ZNF28 | 11.88   | 3.36    |
| ZNF460| 14.14   | 1.63    |

### Reverse-forward orientation

| TF    | Score 1 | Score 2 |
|-------|---------|---------|
| ASCL2 | 6.23    | 3.80    |
| ATF3  | 17.21   | 2.23    |
| BC11A | 1.65    | 4.07    |
| BDPI1 | 23.49   | 6.85    |
| CTCF  | 13.32   | 11.74   |
| E2F2  | 23.33   | 1.95    |
| EBF1  | 14.83   | 7.11    |
| GABP1-GABP2 | 6.43 | 2.80 |
| IRF3  | 2.38    | 1.98    |
| IRF8  | 4.30    | 4.62    |
| NRR3  | 17.95   | 1.78    |
| PAX5  | 48.74   | 2.74    |
| POLR3A| 41.85   | 9.98    |
| RAD21 | 1.79    | 2.26    |
| RNF96 | 7.69    | 1.93    |
| RPA4  | 21.32   | 4.48    |
| RRA   | 28.89   | 11.87   |
| SMAD4 | 41.27   | 2.44    |
| SP2   | 2.09    | 11.15   |
| SP4   | 12.74   | 1.84    |
| SPB   | 20.03   | 2.17    |
| TCFAP2A | 8.90 | 3.53    |
| TCFAP2E | 5.72 | 1.84    |
| ZFP202| 23.15   | 4.85    |

### References

Bailey SD, Zhang X, Desai K, Aid M, Corradin O, Cowper-Sal Lari R, Akhtar-Zaidi B, Scacheri PC, Haibe-Kains B, Lupien M. 2015. ZNF143 provides sequence specificity to secure chromatin interactions at gene promoters. *Nat Commun* **2**: 6186.
Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS. 2009. MEME SUITE: tools for motif discovery and searching. *Nucleic acids research* **37**: W202-208.

Barutcu AR, Lajoie BR, Fritz AJ, McCord RP, Nickerson JA, van Wijnen AJ, Lian JB, Stein JL, Dekker J, Stein GS et al. 2016. SMARCA4 regulates gene expression and higher-order chromatin structure in proliferating mammary epithelial cells. *Genome research* **26**: 1188-1201.

de Wit E, Vos ES, Holwerda SJ, Valdes-Quezada C, Verstegen MJ, Teunissen H, Splinter E, Wijchers PJ, Krijger PH, de Laat W. 2015. CTCF Binding Polarity Determines Chromatin Looping. *Molecular cell* **60**: 676-684.

Grant CE, Bailey TL, Noble WS. 2011. FIMO: scanning for occurrences of a given motif. *Bioinformatics (Oxford, England)* **27**: 1017-1018.

Guo Y, Xu Q, Canzio D, Shou J, Li J, Gorkin DU, Jung I, Wu H, Zhai Y, Tang Y et al. 2015. CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function. *Cell* **162**: 900-910.

Jeong M, Huang X, Zhang X, Su J, Shamim M, Bochkov I, Reyes J, Jung H, Heikamp E, Presser Aiden A et al. 2017. A Cell Type-Specific Class of Chromatin Loops Anchored at Large DNA Methylation Nadirs. *bioRxiv*.

Jolma A, Yan J, Whitington T, Toivonen J, Nitta KR, Rastas P, Morgunova E, Enge M, Taipale M, Wei G et al. 2013. DNA-binding specificities of human transcription factors. *Cell* **152**: 327-339.

Ji X, Dadon DB, Abraham BJ, Lee TI, Jaenisch R, Bradner JE, Young RA. 2015. Chromatin proteomic profiling reveals novel proteins associated with histone-marked genomic regions. *Proc Natl Acad Sci U S A* **112**: 3841-3846.

Jolma A, Yan J, Whittington T, Toivonen J, Nitta KR, Rastas P, Morgunova E, Enge M, Taipale M, Wei G et al. 2013. DNA-binding specificities of human transcription factors. *Cell* **152**: 327-339.

Karolchik D, Barber GP, Casper J, Clawson H, Cline MS, Diekhans M, Dreszer TR, Fujita PA, Guruvadoo L, Haeussler M et al. 2014. The UCSC Genome Browser database: 2014 update. *Nucleic acids research* **42**: D764-770.

Kheradpour P, Kellis M. 2014. Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic acids research* **42**:
McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM, Bejerano G. 2010. GREAT improves functional interpretation of cis-regulatory regions. *Nature biotechnology* **28**: 495-501.

Mumbach MR, Satpathy AT, Boyle EA, Dai C, Gowen BG, Cho SW, Nguyen ML, Rubin AJ, Granja JM, Kazane KR et al. 2017. Enhancer connectome in primary human cells identifies target genes of disease-associated DNA elements. *Nature genetics* **49**: 1602-1612.

Newburger DE, Bulyk ML. 2009. UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. *Nucleic acids research* **37**: D77-82.

Osato N. 2018. Characteristics of functional enrichment and gene expression level of human putative transcriptional target genes. *BMC Genomics* **19**: 957.

Phanstiel DH, Van Bortle K, Spacek D, Hess GT, Shamim MS, Machol I, Love MI, Aiden EL, Bassik MC, Snyder MP. 2017. Static and Dynamic DNA Loops form AP-1-Bound Activation Hubs during Macrophage Development. *Molecular cell* **67**: 1037-1048.e1036.

Portales-Casamar E, Thongjuea S, Kwon AT, Arenillas D, Zhao X, Valen E, Yusuf D, Lenhard B, Wasserman WW, Sandelin A. 2010. JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic acids research* **38**: D105-110.

Ramani V, Cusanovich DA, Hause RJ, Ma W, Qiu R, Deng X, Blau CA, Disteche CM, Noble WS, Shendure J et al. 2016. Mapping 3D genome architecture through in situ DNase Hi-C. *Nat Protoc* **11**: 2104-2121.

Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES et al. 2014. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**: 1665-1680.

Saeed S, Quintin J, Kerstens HH, Rao NA, Aghajanirefah A, Matarese F, Cheng SC, Ratter J, Berentsen K, van der Ent MA et al. 2014. Epigenetic programming of monocyte-to-macrophage differentiation and trained innate immunity. *Science (New York, NY)* **345**: 1251086.

Schreiber J, Libbrecht M, Bilmes J, Noble W. 2017. Nucleotide sequence and DNasel
sensitivity are predictive of 3D chromatin architecture. *bioRxiv*.

Wang L, Wang S, Li W. 2012. RSeQC: quality control of RNA-seq experiments. *Bioinformatics (Oxford, England)* 28: 2184-2185.

Weintraub AS, Li CH, Zamudio AV, Sigova AA, Hannett NM, Day DS, Abraham BJ, Cohen MA, Nabet B, Buckley DL et al. 2017. YY1 Is a Structural Regulator of Enhancer-Promoter Loops. *Cell* 171: 1573-1588 e1528.

Wingender E, Dietze P, Karas H, Knuppel R. 1996. TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic acids research* 24: 238-241.

Xie Z, Hu S, Blackshaw S, Zhu H, Qian J. 2010. hPDI: a database of experimental human protein-DNA interactions. *Bioinformatics (Oxford, England)* 26: 287-289.

Zhang H, Li F, Jia Y, Xu B, Zhang Y, Li X, Zhang Z. 2017. Characteristic arrangement of nucleosomes is predictive of chromatin interactions at kilobase resolution. *Nucleic acids research* 45: 12739-12751.

Zhang Y, An L, Xu J, Zhang B, Zheng WJ, Hu M, Tang J, Yue F. 2018. Enhancing Hi-C data resolution with deep convolutional neural network HiCPlus. *Nat Commun* 9: 750.

Zhao Y, Stormo GD. 2011. Quantitative analysis demonstrates most transcription factors require only simple models of specificity. *Nature biotechnology* 29: 480-483.