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

Naoki Osato[1*]

[1]Department 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. Though the number of chromatin interactions was about seventy thousand in Hi-C at 50kb resolution, about twenty millions of chromatin interactions were recently identified by HiChIP at 5kb resolution. 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, and repeat DNA sequences affecting the interaction between enhancers and promoters of genes and their expression, first I predicted TF bound in enhancers and promoters using DNA motif sequences of TF and experimental data of open chromatin regions in monocytes and other cell types, which were obtained from public and commercial databases. Second, transcriptional target genes of each TF were predicted based on enhancer-promoter association (EPA). EPA was shortened at the genomic locations of FR or reverse-forward (RF) orientation of DNA motif sequence of a TF, which were supposed to be at chromatin interaction anchors and acted as insulator sites like CTCF. Then, the expression levels of the transcriptional target genes predicted based on the EPA were compared with those predicted from only promoters.

Total 369 biased orientation of DNA motifs (232 FR and 178 RF orientation, the reverse complement sequences of some DNA motifs were also registered in databases, so the total number was smaller than the number of FR and RF) affected the expression level of putative transcriptional target genes significantly in CD14$^+$ monocytes of four people in common. The same analysis was conducted in CD4$^+$ T cells of four people. DNA motif sequences of CTCF, cohesin and other transcription factors involved in chromatin interactions were found to be a biased orientation. Transposon sequences, which are known to be involved in insulators and enhancers, showed a biased orientation. The biased orientation of DNA motif sequences tended to be co-localized in the same open chromatin regions. Moreover, for 36 – 95% of FR and RF orientations of DNA motif sequences, EPI predicted from EPA that were shortened at the genomic locations of the biased orientation of DNA motif sequence were overlapped with chromatin interaction data (Hi-C and HiChIP) significantly more than other types of EPAs.

Keywords: transcriptional target genes, gene expression, transcription factors, enhancer, enhancer-promoter interactions, chromatin interactions, CTCF, cohesin, open chromatin regions, co-location of transcription factors, homodimer, heterodimer, complex

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. Though the number of chromatin interactions was about seventy thousand in Hi-C at 50kb resolution (Javierre et al. 2016), about twenty millions of chromatin interactions were recently identified by HiChIP at 5kb resolution (Mumbach et al. 2017). 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 EPI. The DNA binding motif sequences of the transcription factors (TF) 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 enhancer and promoter marks 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 TF 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 loops formed by chromatin interactions. In my previous analysis, 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 were predicted based on EPA shortened at the genomic locations of FR orientation of CTCF binding sites, and transcriptional target genes of each TF bound in enhancers and promoters were predicted based on the EPI. The EPA affected the expression levels of putative transcriptional target genes the most among three types of EPAs, compared with the expression levels of transcriptional target genes predicted from only promoters (Fig. 2). The expression levels 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, may 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 loops 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 TF and repeat DNA sequences, affecting EPI and the expression level of putative transcriptional target genes in CD14$^+$ monocytes and CD4$^+$ T cells 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 (TF) 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 sequence of a TF acting as insulator such as CTCF and cohesin (RAD21 and SMC3) (Fig. 1B). To find DNA motif sequences of TF acting as insulator, other than CTCF and cohesin, and repeat DNA sequences affecting the expression level of genes, EPI were predicted based on EPA shortened at the genomic locations of the DNA motif sequence of a TF or a repeat DNA sequence, and transcriptional target genes of each TF bound in enhancers and promoters were predicted based on the EPI. The expression levels of the putative transcriptional target genes were compared with those predicted from promoters using Mann Whiteney U test, two-sided (*p*-value < 0.05) (Fig. 2A). The number of TF showing a significant difference of expression level of their putative transcriptional target genes was counted. To examine whether the orientation of a DNA motif sequence, which is supposed to act as insulator sites and shorten EPA, affected the number of TF showing a significant difference of expression level of their putative transcriptional target genes, the number of the TF was compared among forward-reverse (FR), reverse-forward (RF), and any orientation (i.e. without considering orientation) of a DNA motif sequence 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 369 of biased (232 FR and 178 RF) orientation of DNA binding motif sequences of TF were found in monocytes of four people in common, whereas only seven any orientation of DNA binding motif sequence was found in monocytes of four people in common (Fig. 2B; Table 1; Supplemental Table S1). FR orientation of DNA motif sequences included CTCF, cohesin (RAD21 and SMC3), ZNF143 and YY1, which are associated with chromatin interactions and EPI. SMARCA4 (BRG1) is associated with topologically associated domain (TAD), which is higher-order chromatin organization. The DNA binding motif sequence of SMARCA4 was not registered in the databases used in this study. FR orientation of DNA motif sequences included SMARCC2, a member of the same SWI/SNF family of proteins as SMARCA4.

The same analysis was conducted using DNase-seq data of CD4$^+$ T cells of four people. Total 376 of biased (203 FR and 213 RF) orientation of DNA binding motif sequences of TF were found in T cells of four people in common, whereas only seven any orientation (i.e. without considering orientation) of DNA binding motif sequences were found in T cells of four people in common (Supplemental Fig. S1 and Supplemental Table S2). Biased orientation of DNA motif sequences in T cells included CTCF, cohesin (RAD21 and SMC3), and SMARC. Among 369, 73 of biased orientation of DNA binding motif sequences of TF were found in both monocytes and T cells in common (Supplemental Table S5). For each orientation, 46 FR and 34 RF orientation of DNA binding motif sequences of TF were found in both monocytes and T cells in common. Without considering the difference of orientation of DNA binding motif sequences, 113 of biased orientation of DNA binding motif sequences of TF were found in both monocytes and T cells. As a reason for the increase of the number (113) from 73, a TF or an isoform of the same TF may bind to a different DNA binding motif sequence according to cell types and/or in the same cell type. About 50% or more of alternative splicing isoforms are differently expressed among tissues, indicating that most alternative splicing is subject to tissue-specific regulation (Wang et al. 2008) (Chen and Manley 2009) (Das et al. 2007). The same TF has several DNA binding motif sequences and in some cases one of the motif sequences is almost the same as the reverse complement sequence of another motif sequence of the same TF. For example, RAD21 had both FR and RF orientation of DNA motif sequences, but the number of the FR orientation of DNA motif sequence was relatively small in the genome, and the RF orientation of DNA motif sequence was frequently observed and co-localized with

CTCF. I previously found that a complex of TF would bind to a slightly different DNA binding motif sequence from the combination of DNA binding motif sequences of TF composing the complex in *C. elegans* (Tabuchi et al. 2011). From another viewpoint of this study, the expression level of putative transcription target genes of some TF would be different, depending on the genomic locations (enhancers or promoters) of DNA binding motif sequences of the TF in monocytes and T cells of four people.

Moreover, using open chromatin regions overlapped with H3K27ac histone modification marks known as enhancer and promoter marks, the same analyses were performed in monocytes and T cells. H3K27ac histone modification marks were used in the analysis of EPI, but were not used in the analysis of TF as insulator like CTCF and cohesin in this study, since new biased orientation of DNA motif sequences were found in this criterion. When H3K27ac histone modification marks were used in the analysis of TF as insulator like CTCF and cohesin, the number of biased orientation of DNA motif sequences was decreased. Total 233 of biased (179 FR and 70 RF) orientation of DNA binding motif sequences of TF were found in monocytes of four people in common, whereas only two any orientation of DNA binding motif sequence was found (Supplemental Table S3). Though the number of biased orientation of DNA motif sequences was reduced, CTCF, RAD21, SMC3, ZNF143, and YY1 were found. For T cells using H3K27ac histone modification marks, total 291 of biased (173 FR and 143 RF) orientation of DNA binding motif sequences of TF were found in T cells of four people in common, whereas only 10 any orientation of DNA binding motif sequences were found (Supplemental Table S4). Though the number of biased orientation of DNA motif sequences was reduced, CTCF, RAD21, SMC3, and YY1 were found. Scores of CTCF, RAD21, and SMC3 were increased compared with the result of T cells without using H3K27ac histone modification marks, and they were ranked in the top four. Biased orientation of DNA motif sequences included JUNDM2 (JDP2), which is involved in histone-chaperone activity, promoting nucleosome, and inhibition of histone acetylation (Jin et al. 2006). JDP2 forms a homodimer or heterodimer with various TF (https://en.wikipedia.org/wiki/Jun_dimerization_protein). As summary of the results with and without H3K27ac histone modification marks, total 433 of biased (306 FR and 178 RF) orientation of DNA motif sequences were found in monocytes of four people in common. Total 499 of biased (285 FR and 278 RF) orientation of DNA motif sequences were found in T cells of four people in common. Total number of these results in monocytes and T cells was 773 biased (513 FR and 413 RF) orientation of DNA motif sequences. Biased orientation of DNA motif sequences found in both monocytes and T cells were listed in Supplemental Table S5.

To examine whether the biased orientation of DNA binding motif sequences of TF were 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, 61, 135, 95, and 108 DNA binding motif sequences were also observed in H1-hESC, iPS, Huvec and MCF-7 respectively, including CTCF and cohesin (RAD21 and SMC3) (Table 2; Supplemental Table S6).

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 TF show a significant difference between promoters and EPA shortened at the genomic locations of a DNA motif sequence acting as insulator sites. Some experimental data of a cell did not show a significant difference between promoters and the EPA (Osato 2018).

Instead of DNA binding motif sequences of TF, repeat DNA sequences were also examined. The expression levels of transcriptional target genes of each TF predicted based on EPA that were shortened at the genomic locations of a repeat DNA sequence were compared with those predicted from promoters. Three RF orientation of repeat DNA 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 repeat DNA sequence was observed in iPS and H1-hESC with enough statistical significance considering multiple tests (p-value < 10^{-7}). The same as CD14$^+$ monocytes, biased orientation of repeat DNA sequences were examined in CD4$^+$ T cells. Three FR and two RF orientation of repeat DNA sequences showed a significant difference of expression level of putative transcriptional target genes in T cells of four people in common (Supplemental Table S7). MIRb and MIR3 were also found in the analysis using open chromatin regions overlapped with H3K27ac histone modification marks, which are enhancer and promoter marks. MIR and other transposon sequences are known to act as insulators and enhancers (Bejerano et al. 2006; Rebollo et al. 2012; de Souza et al. 2013; Jjingo et al. 2014; Wang et al. 2015).

Co-location of biased orientation of DNA motif sequences

To examine the association of 369 biased (FR and RF) orientation of DNA binding motif sequences of TF, co-location of the DNA binding motif sequences in open chromatin regions was analyzed in monocytes. 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 < 1.0×10^{-10}), they were listed (Table 4; Supplemental Table S8). Open chromatin regions overlapped with histone modification of enhancer and promoter marks (H3K27ac) (total 26,095 regions) showed a larger number of enriched pairs of DNA motifs than all open chromatin regions (Table 4; Supplemental Table S8). 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 428, consisting of 120 unique 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. Biased orientation of DNA binding motif sequences of TF tended to be co-localized in the same open chromatin regions.

To examine the association of 376 biased orientation of DNA binding motif sequences of TF in CD4$^+$ T cells, co-location of the DNA binding motif sequences in open chromatin regions was analyzed. Top 30 pairs of FR and RF orientations of DNA motifs co-occupied in the same open chromatin regions were shown (Supplemental Table S9). Total number of pairs of DNA motifs was 99, consisting of 72 unique 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 (chi-square test, p-value < 1.0×10^{-10}). Among them, 11 pairs of DNA motif sequences including a pair of CTCF, SMC3, and RAD21 in T cells were found in monocytes in common (Supplemental Table S9).

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 shortened at the genomic locations of FR or RF orientation of DNA motif sequence of a TF, (ii) EPA shortened at the genomic locations of DNA motif sequence of a TF acting as insulator sites such as CTCF and cohesin (RAD21, SMC3) without considering their orientation, and (iii) EPA without being shortened by the genomic locations of a DNA motif sequence. EPA (i) showed a significantly higher ratio of EPI overlapped with chromatin interactions (Hi-C) using DNA binding motif sequences of CTCF and cohesin (RAD21 and SMC3) than the other two types of EPAs (n = 4, binomial test, two-sided) (Supplemental Fig. S2). Total 58 biased orientation (38 FR and 22 RF) of DNA motif sequences including CTCF, cohesin, and YY1 showed a significantly higher ratio of EPI overlapped with Hi-C chromatin interactions (a cutoff score of CHiCAGO tool > 1) than the other types of EPAs in monocytes (Supplemental Table S10). When comparing EPI predicted based on only EPA (i) and (iii) with chromatin interactions, total 215 biased orientation (130 FR and 102 RF) of DNA motif sequences showed a significantly higher ratio of EPI predicted based on EPA (i) overlapped with the chromatin interactions than EPI predicted based on EPA (iii) (Supplemental material 2). The difference between EPI predicted based on EPA (i) and (ii) seemed to be difficult to distinguish using the chromatin interaction data and the statistical test in some cases. However, as for the difference between EPI predicted based on EPA (i) and (iii), a larger number of biased orientation of DNA motif sequences was found to be correlated with chromatin interaction data. Chromatin interaction data were obtained from 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 HiChIP, 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 would be larger and more significant (Rao et al. 2014; Ramani et al. 2016; Mumbach et al. 2017; Zhang et al. 2018).

After the analysis of CD14$^+$ monocytes, to utilize HiChIP chromatin interaction data in CD4$^+$ T cells, the same analysis for CD14$^+$ monocytes was performed using DNase-seq data of four donors in CD4$^+$ T cells (Mumbach et al. 2017). EPI predicted based on EPA were compared with three replications (B2T1, B2T2, and B3T1) of HiChIP chromatin interaction data in CD4$^+$ T cells respectively. The resolutions of HiChIP chromatin interaction data and EPI were adjusted to 5kb. EPI were predicted based on the three types of EPA in the same way as CD14$^+$ monocytes using top 60% expression level of all transcripts (genes) excluding transcripts not expressed in T cells. The criteria of the analysis were determined to include known DNA motif sequences involved in chromatin interactions such as CTCF and cohesin in the result, and the result was consistent with that using Hi-C chromatin interaction data. EPA (iii) showed the highest ratio of EPI overlapped with chromatin interactions (HiChIP) using DNA binding motif sequences of CTCF and cohesin (RAD21 and SMC3), compared with the other two types of EPA (i) and (ii) (n = 4, binomial test, two-sided, 95% confidence interval) (Fig. 3). Total 136 biased orientation (70 FR and 73 RF) of DNA motif sequences, which included CTCF, cohesin (RAD21 and SMC3), and SMARC in three replications (B2T1, B2T2, and B3T1) and ZNF143 in two replications (B2T2 and B3T1), showed a significantly higher ratio of EPI overlapped with HiChIP chromatin interactions (more than 1,000 counts for each interaction) than the other types of EPAs in T cells (Table 5). When comparing EPI predicted based on only EPA (i) and (iii) with the chromatin interactions, total 356 biased orientation (194 FR and 200 RF) of DNA motif sequences showed a significantly higher ratio of EPI predicted based on EPA (iii) overlapped with the chromatin interactions than EPI predicted based on EPA (i) (Table 5; Supplemental material 2). As expected, the number of EPI overlapped with chromatin interactions (HiChIP) was increased, compared with Hi-C chromatin interactions. Most of biased orientation of DNA motif sequences (95%) were found to be correlated with chromatin interactions, when comparing EPI predicted based on EPA (i) and (iii) with HiChIP chromatin interactions.

Moreover, to examine the enhancer activity of EPI, the distribution of expression level of putative target genes of EPI was compared between EPI overlapped with HiChIP chromatin interactions and EPI not overlapped with them. Though the target genes of EPI were selected from top 60% expression level of all transcripts (genes) excluding transcripts not expressed in T cells, target genes of EPI overlapped with chromatin interactions showed a significantly higher expression level than EPI not overlapped with them, suggesting that EPI overlapped with chromatin interactions activated the expression of target genes in T cells. Almost all (99.9%) FR and RF orientations of DNA motifs showed a significantly higher expression level of putative target genes of EPI overlapped with chromatin interactions than EPI not overlapped. When a biased orientation of DNA motif showed a significantly higher expression level of putative target genes of EPI overlapped with chromatin interactions than EPI not overlapped, ‘1’ was marked with in the tables of the comparison between EPI and HiChIP chromatin interactions in Supplemental material 2. When a DNA motif showed a significantly lower expression level, ‘-1’ was marked with, however, it was not observed in this analysis. When there was not significant difference of expression level, ‘0’ was marked with.

If biased orientation of DNA motif sequences of TF found in both monocytes and T cells are biologically meaningful, these may match the result of the analysis of HiChIP data. Among 376 FR and RF orientations of biased orientation of DNA motifs of TF in T cells, 136 (36%) were biased orientation of DNA motifs in the analysis of HiChIP data for three types of EPA (i) (ii) and (iii). Among 73 FR and RF orientations of DNA motifs of TF found in both monocytes and T cells, 31 (42%) were biased orientation of DNA motifs in the analysis of HiChIP data, which was significantly higher ratio than all 376 biased orientation of DNA motifs in T cell, and included CTCF, RAD21 and SMC3 (*p*-value < 0.015, binomial test, two-sided, 95% confidence interval) (Supplemental Table S5 and Table 5). However, this may not imply that all 376 biased orientation of DNA motifs included false-positive predictions, and may be due to the limitation of resolution of the HiChIP data (5kb) or the small number of DNA binding sites of a TF in genome sequences. Then, among 113 FR and RF orientations of DNA motifs of TF found in both monocytes and T cells without considering the difference of orientation (FR or RF) of DNA binding motifs, 42 (37%) were biased orientation of DNA motifs in the analysis of HiChIP data, which was not significantly higher ratio. This implied that the difference of orientation of DNA motifs was important to predict EPI in comparison with HiChIP data.

Though the ratios of EPI overlapped with chromatin interactions were increased by using many chromatin interaction data including lower score and count of chromatin interactions (a cutoff score of CHiCAGO tool > 1 for Hi-C and more than 1,000 counts for HiChIP), the ratios of EPI overlapped with chromatin interactions showed the same tendency among the three types of EPAs. The ratio of EPI overlapped with Hi-C chromatin interactions was increased using H3K27ac marks in both monocytes and T cells. The ratio of EPI overlapped with HiChIP chromatin interactions was also increased using H3K27ac marks. Chromatin interaction data were obtained from different samples from DNase-seq, open chromatin regions in CD4$^+$ T cells, so individual differences seemed to be large from the results of this analysis, and (Mumbach et al. 2017) suggested that individual differences of chromatin interactions were larger than those of open chromatin regions. ATAC-seq data, open chromatin regions were available in CD4$^+$ T cells in the paper, however, when using ATAC-seq data, the result of the analysis of biased orientation of DNA motif sequences was different from DNase-seq data, and not included a part of CTCF and cohesin. Thus, DNase-seq data collected from ENCODE and Blueprint projects were employed in this study.

Discussion

To find DNA motif sequences of transcription factors (TF) and repeat DNA sequences affecting the expression level of human putative transcriptional target genes, the DNA motif sequences were searched from open chromatin regions of monocytes of four people. Total 369 biased [232 forward-reverse (FR) and 178 reverse-forward (RF)] orientation of DNA motif sequences of TF were found in monocytes of four people in common, whereas only seven any orientation (i.e. without considering orientation) of

DNA motif sequence of TF was found to affect the expression level of putative transcriptional target genes, suggesting that enhancer-promoter association (EPA) shortened at the genomic locations of FR or RF orientation of the DNA motif sequence of a TF or a repeat DNA sequence is an important character for the prediction of enhancer-promoter interactions (EPI) and the transcriptional regulation of genes.

When DNA motif sequences were searched from monocytes of one person, a larger number of biased orientation of DNA motif sequences affecting the expression level of human putative transcriptional target genes were found. When the number of donors, 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 FR orientation of DNA binding motif sequences of CTCF and cohesin is frequently observed at chromatin interaction anchors, the percentage of FR 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 TF would be observed less frequently in open chromatin regions. The analyses of experimental data of a number of people would avoid missing relatively weak statistical significance of DNA motif sequences of TF in experimental data of each person by multiple testing correction of thousands of statistical tests. A DNA motif sequence was found with *p*-value < 0.05 in experimental data of one person and the DNA motif sequence found in the same cell type of four people in common would have *p*-value < $0.05^4 = 6.25 \times 10^{-6}$. 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. EPAs 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 EPAs (see Methods). 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. (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 analysis. As an issue of experimental data, data for chromatin interactions and open chromatin regions were came from different samples and donors, so individual differences would exist in the data. Moreover, the resolution of chromatin interaction data used in monocytes was about 50kb, thus the number of chromatin interactions was relatively small (72,284 at 50kb resolution with a cutoff score of CHiCAGO tool > 1 and 16,501 with a cutoff score of CHiCAGO tool > 5). EPI predicted based on EPA shortened at the genomic locations of

DNA binding motif sequence of TF that were found in various open chromatin regions such as CTCF and cohesin (RAD21 and SMC3) tended to be overlapped with a larger number of chromatin interactions than TF less frequently observed in open chromatin regions. Therefore, to examine the difference of the numbers of EPI overlapped with chromatin interactions, according to the three types of EPAs, the number of chromatin interactions should be large enough.

As HiChIP chromatin interaction data were available in CD4$^+$ T cells, biased orientation of DNA motif sequences of TF were examined in T cells using DNase-seq data of four people. The resolutions of chromatin interactions and EPI were adjusted to 5kb by fragmentation of genome sequences. In monocytes, the resolution of Hi-C chromatin interaction data was converted by extending anchor regions of chromatin interactions to 50kb length and merging the chromatin interactions overlapped with each other. Fragmentation of genome sequences may affect the classification of chromatin interactions of which anchors are located near the border of a fragment, but the number of chromatin interactions would not be decreased, compared with merging chromatin interactions. The number of HiChIP chromatin interactions was 19,926,360 at 5kb resolution, 666,149 at 5kb resolution with chromatin interactions (more than 1,000 counts for each interaction), and 78,209 at 5kb resolution with chromatin interactions (more than 6,000 counts for each interaction). As expected, the number of EPI overlapped with chromatin interactions was increased, and 36 – 95% of biased orientation of DNA motif sequences of TF showed a statistical significance in EPI predicted based on EPA shortened at the genomic locations of the DNA motif sequence, compared with the other types of EPAs or EPA not shortened. False positive predictions of EPI would be decreased by using H3K27ac marks and other features. The ratio of EPI overlapped with Hi-C chromatin interactions was increased using H3K27ac marks in both monocytes and T cells. The ratio of EPI overlapped with HiChIP chromatin interactions was also increased using H3K27ac marks. However, the number of biased orientation of DNA motif sequences showing a higher ratio of EPI overlapped with HiChIP chromatin interactions than the other types of EPAs was decreased using H3K27ac marks (Supplemental material 2).

When forming a homodimer or heterodimer with another TF, TF may bind to genome DNA with a specific orientation of their DNA binding sequences (Fig. 4). From the analysis of biased orientation of DNA motif sequences of TF, TF forming heterodimer would also be found. If the DNA binding motif sequence of only the mate to a pair of TF was found in EPA, EPA was shortened at one side, which is the genomic location of the DNA binding motif sequence of the mate to the pair, and transcriptional target genes were predicted using the EPA shortened at the side. In this analysis, the mate to both heterodimer and homodimer of TF can be used to examine the effect on the expression level of transcriptional target genes predicted based on the EPA shortened at one side. For heterodimer, biased orientation of DNA motif sequences may also be found in forward-forward or reverse-reverse orientation.

Some DNA binding sites of TF predicted using DNA binding motif sequences of TF were changed according to the parameters of FIMO tool, particularly background frequencies of ATGC nucleotides in genome sequences and *p*-value threshold. Repeat

DNA sequences also affected the result of the prediction. Without repeat masking, the number of any orientation of DNA motifs of TF was increased. However, the *p*-value of the DNA motifs was relatively high and close to the threshold, so these DNA motifs seemed to be false positives. To decrease false-positive and false-negative predictions of DNA binding sites of TF, improve the prediction of biased orientation of DNA motifs, and obtain a robust result of the analysis, there may be a room to explore more suitable parameters and methods, such as stricter *p*-value threshold, using genomic regions conserved among species, masking some exons encoding mRNA, removing DNA motifs highly affected by parameter changes (low information content), changing the parameter of nucleotide frequencies according to genomic regions, considering epigenetic modifications (DNA methylation and histone) and so on. For the analysis of EPI, instead of using all DNA motif sequences of TF in databases, selecting DNA motif sequences of TF indicating enhancer activity in a cell type using my method would reduce the effect of TF not acting as enhancer (Osato 2018).

It has been reported that CTCF and cohesin-binding sites are frequently mutated in cancer (Katainen et al. 2015). Some biased orientation of DNA motif sequences would be associated with chromatin interactions and might be associated with diseases including cancer.

The analyses in this study revealed novel characters of DNA binding motif sequences of TF and repeat DNA sequences to analyze TF involved in chromatin interactions, insulator function and forming a homodimer, heterodimer or complex with other TF, affecting the transcriptional regulation of genes.

Methods

Search for biased orientation of DNA motif sequences

To examine transcriptional regulatory target genes of transcription factors (TF), bed files of hg38 of Blueprint DNase-seq data for CD14$^+$ monocytes of four donors (EGAD00001002286; Donor ID: C0010K, C0011I, C001UY, C005PS) were obtained from Blueprint project 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 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/).

As high resolution of chromatin interaction data using HiChIP became available in CD4$^+$ T cells, to promote the same analysis of CD14$^+$ monocytes in CD4$^+$ T cells, DNase-seq data of four donors were obtained from a public database and Blueprint projects web sites. DNase-seq data of only one donor was available in Blueprint project for CD4$^+$ T cells (Donor ID: S008H1), and three other DNase-seq data were obtained from NCBI Gene Expression Omnibus (GEO) database. Though the peak calling of DNase-seq data available in GEO database was different from other DNase-seq data in

ENCODE and Blueprint projects where 150bp length of peaks were usually predicted using HotSpot (John et al. 2011), FASTQ files of DNase-seq data were downloaded from NCBI Sequence Read Archive (SRA) database (SRR097566, SRR097618, and SRR171574). Read sequences of the FASTQ files were aligned to the hg19 version of the human genome reference using BWA (Li and Durbin 2009), and the BAM files generated by BWA were converted into SAM files, sorted, and indexed using Samtools (Li et al. 2009). Peaks of the DNase-seq data were predicted using HotSpot-4.1.1.

To identify transcription factor binding sites (TFBS) from the DNase-seq data, TRANSFAC (2019.1), JASPAR (2018), UniPROBE (2018), BEEML-PBM, high-throughput SELEX, Human Protein-DNA Interactome, transcription factor binding sequences of ENCODE ChIP-seq data, and HOCOMOCO version 9 and 11 were used to predict insulator sites (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) (Kulakovskiy et al. 2018). TRANSFAC (2011.1), JASPAR (2012), UniPROBE (2012), BEEML-PBM, high-throughput SELEX, and Human Protein-DNA Interactome were used to analyze enhancer-promoter interactions, since these data were sufficient to identify biased orientation of DNA motif sequences of TF with less computational time, reducing the number of any orientation of DNA motif sequences of TF. 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 TF derived from vertebrates were used for further analyses. Transcription factor binding sequences were searched from each narrow peak of DNase-seq data in repeat-masked hg19 genome sequences using FIMO with *p*-value threshold of 10^{-5} and background frequencies of ATGC nucleotides in repeat-masked hg19 genome sequences (Grant et al. 2011). Repeat-masked hg19 genome sequences were downloaded from UCSC genome browser (http://genome.ucsc.edu/, http://hgdownload.soe.ucsc.edu/goldenPath/hg19/bigZips/hg19.fa.masked.gz). TF 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.soe.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 TF 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 distance 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 a TF that was 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 (FR) orientation of DNA binding sites of a TF. 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 TF; 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.

Repeat DNA sequences were searched from the hg19 version of the human reference genome using RepeatMasker (Smit, AFA & Green, P RepeatMasker at http://www.repeatmasker.org) and RepBase RepeatMasker Edition (http://www.girinst.org).

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, and MCF-7 (GSM26284, 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. FPKMs of the RNA-seq data were calculated using RSeQC (Wang et al. 2012). For monocytes, Blueprint RNA-seq FPKM data ('C0010KB1.transcript_quantification.rsem_grape2_crg.GRCh38.20150622.results', 'C0011IB1.transcript_quantification.rsem_grape2_crg.GRCh38.20150622.results', 'C001UYB4.transcript_quantification.rsem_grape2_crg.GRCh38.20150622.results', and 'C005PS12.transcript_quantification.rsem_grape2_crg.GRCh38.20150622.results') were downloaded from Blueprint DCC portal (http://dcc.blueprint-epigenome.eu/#/files). Based on FPKM, UCSC transcripts with top 50% expression level of all the transcripts excluding transcripts not expressed 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 sequence of a TF or a repeat DNA sequence 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 TF, and the distribution of expression level of putative transcriptional target genes of each TF was compared between EPA and only promoter using Mann-Whitney test, two-sided (p-value < 0.05). The number of TF showing a significant difference of expression level of putative transcriptional target genes between EPA and promoter was compared among forward-reverse (FR), reverse-forward (RF), and any orientation (i.e. without considering orientation) of a DNA motif sequence shortening EPA using chi-square test (p-value < 0.05). When a DNA motif sequence of a TF or a repeat DNA sequence shortening EPA showed a significant difference of expression level of putative transcriptional target genes among FR, RF, or any orientation in monocytes of four people in common, the DNA motif sequence was listed.

Though forward-reverse orientation of DNA binding motif sequences of CTCF and cohesin are 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 experimental data of a number of people would avoid missing relatively weak statistical significance of DNA motif sequences in experimental data of each person by multiple testing correction of thousands of statistical tests. A DNA motif sequence was found with p-value < 0.05 in experimental data of one person and the DNA motif sequence found in the same cell type of four people in common would have $p\text{-value} < 0.05^4 = 6.25 \times 10^{-6}$.

Co-location of biased orientation of DNA motif sequences

Co-location of biased orientation of DNA binding motif sequences of TF was examined. The number of open chromatin regions with the same pair of DNA binding motif sequences was counted, and when the pair of DNA binding motif sequences were enriched with statistical significance (chi-square test, p-value < 1.0×10^{-10}), they were listed. For histone modification of an enhancer mark (H3K27ac), bed files of hg38 of Blueprint ChIP-seq data for CD14^{+} monocytes (EGAD00001001179) and CD4^{+} T cells (Donor ID: S000RD) 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). Networks of co-locations of biased orientation of DNA motif sequences were plotted using Cytoscape v3.71 with yFiles Layout Algorithms for Cytoscape (Shannon et al. 2003).

Comparison with chromatin interaction data

For comparison of EPA in monocytes with 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/u8tzp/files/), and chromatin interactions for Monocytes with scores of CHiCAGO tool > 1 and CHiCAGO tool > 5 were extracted from the file (Javierre et al. 2016). In the same way as monocytes, Hi-C chromatin interaction data of CD4^{+} T cells (Naive CD4^{+} T cells, nCD4) were obtained.

Enhancer-promoter interactions (EPI) were predicted using three types of EPAs in monocytes: (i) EPA shortened at the genomic locations of FR or RF orientation of DNA motif sequence of a TF, (ii) EPA shortened at the genomic locations of any orientation (i.e. without considering orientation) of DNA motif sequence of a TF, and (iii) EPA without being shortened by a DNA motif sequence. EPI predicted using the three types of EPAs 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 were compared two times between EPA (i) and (iii), and EPA (i) and (ii) (binomial distribution, *p*-value < 0.025 for each test, two-sided, 95% confidence interval).

For comparison of EPA with chromatin interactions (HiChIP) in CD4$^+$ T cells, 'GSM2705049_Naive_HiChIP_H3K27ac_B2T1_allValidPairs.txt', 'GSM2705050_Naive_HiChIP_H3K27ac_B2T2_allValidPairs.txt', and 'GSM2705051_Naive_HiChIP_H3K27ac_B3T1_allValidPairs.txt' files were downloaded from Gene Expression Omnibus (GEO) database (GSM2705049, GSM2705050 and GSM2705051). The resolutions of chromatin interaction data and EPI were adjusted to 5kb before their comparison. Chromatin interactions with more than 6,000 and 1,000 counts for each interaction were used in this study.

Putative target genes for the analysis of EPI were selected from top 50% expression level of all transcripts excluding transcripts not expressed in monocytes and top 60% expression level of transcripts in CD4$^+$ T cells. The expression level of putative target genes of EPI overlapped with HiChIP chromatin interactions was compared with EPI not overlapped with them. For each FR or RF orientation of DNA motif, EPI were predicted based on EPA and the overlap of EPI with chromatin interactions was examined. When a putative transcriptional target gene of a TF in an enhancer was found in both EPI overlapped with a chromatin interaction and EPI not overlapped with, the target gene was removed. The distribution of expression level of putative target genes was compared using Mann-Whitney test, two-sided (p-value < 0.05).

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

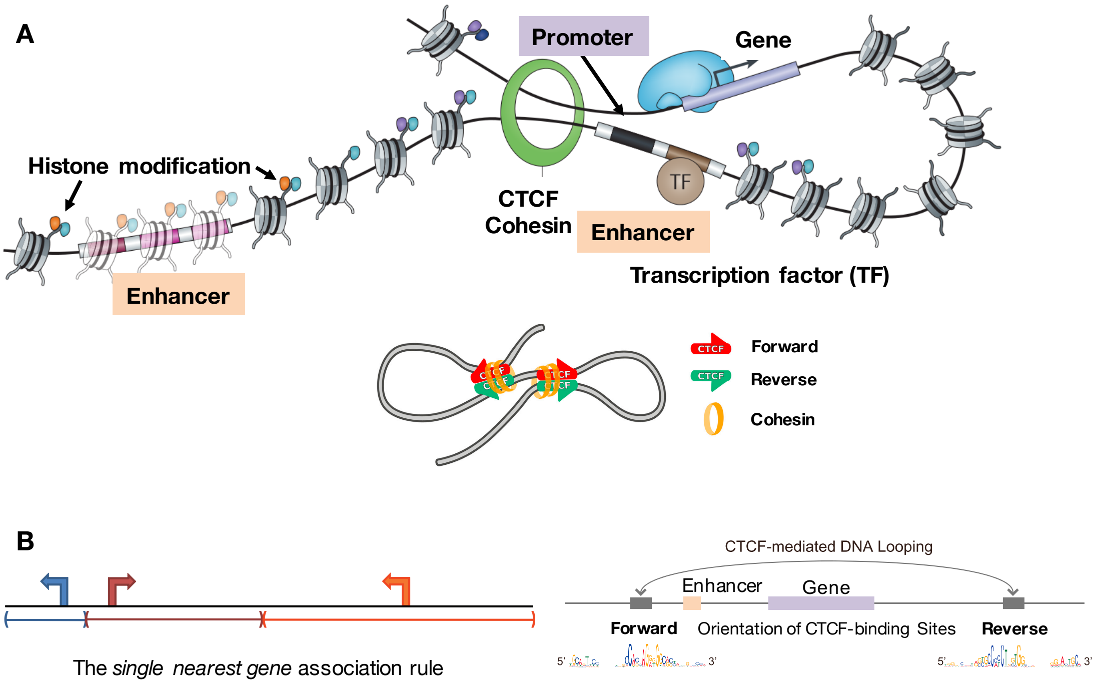

Figure 1. Chromatin interactions and enhancer-promoter association. (A) Forward–reverse orientation of CTCF-binding sites are frequently found in chromatin interaction anchors. CTCF can block the interaction between enhancers and promoters limiting the activity of enhancers to certain functional domains (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 sequence of a transcription factor (e.g. CTCF in the figure).

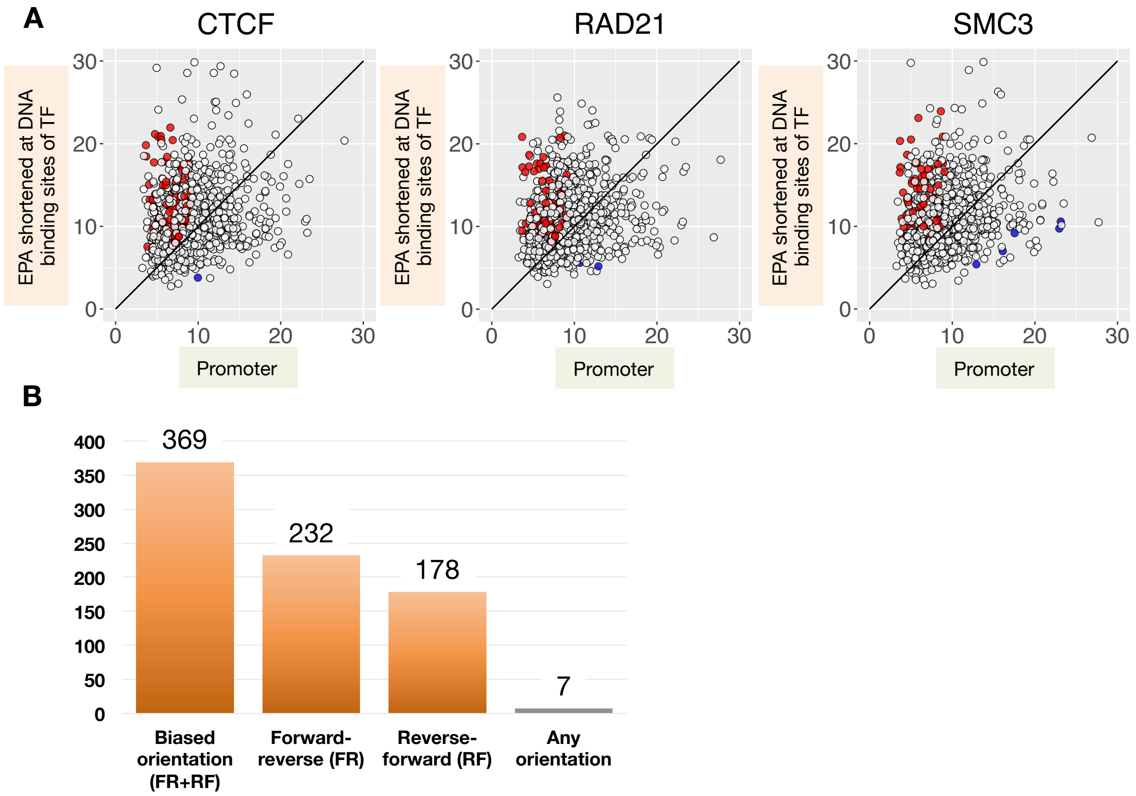

Figure 2. Biased orientation of DNA motif sequences of transcription factors affecting the transcription of genes. (A) Comparison of expression level of putative transcriptional target genes. The median expression levels of target genes of the same transcription factor binding sequences were compared between promoters and enhancer-promoter association (EPA) shortened at the genomic locations of forward-reverse orientation of DNA motif sequence of CTCF and cohesin (RAD21 and SMC3) respectively in monocytes. Red and blue dots show statistically significant difference (Mann-Whitney test) of the distribution of expression levels of target genes between promoters and the EPA. Red dots show the median expression level of target genes was higher based on the EPA than promoters, and blue dots show the median expression level of target genes was lower based on the EPA than promoters. The

expression level of target genes predicted based on the EPA tended to be higher than promoters, implying that transcription factors acted as activators of target genes. (B) Biased orientation of DNA motif sequences of transcription factors in monocytes. Total 369 of biased orientation of DNA binding motif sequences of transcription factors were 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 sequence was found to affect the expression level.

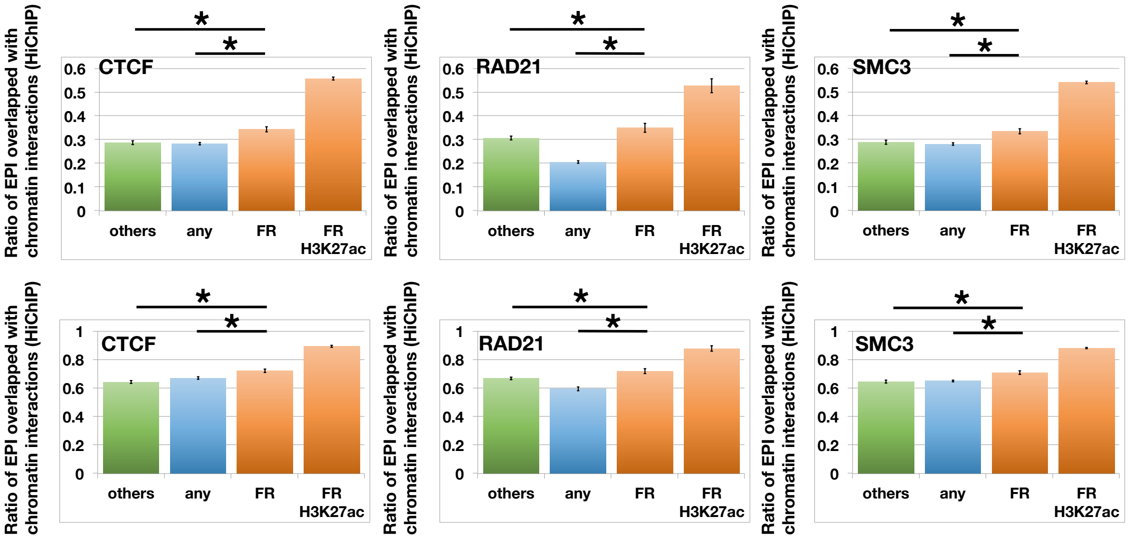

Figure 3. Comparison of enhancer-promoter interactions (EPI) with chromatin interactions in T cells. EPI were predicted based on enhancer-promoter association (EPA) shortened at the genomic locations of biased orientation of DNA binding motif sequence of a transcription factor. Total 136 biased orientation (70 FR and 73 RF) of DNA motif sequences including CTCF and cohesin showed a significantly higher ratio of EPI overlapped with three replications of HiChIP chromatin interactions respectively than the other types of EPAs in T cells. The upper part of the figure is a comparison of EPI with HiChIP chromatin interactions with more than 6,000 counts for each interaction, and the lower part of the figure is a comparison of EPI with HiChIP chromatin interactions with more than 1,000 counts for each interaction. FR: forward-reverse orientation of DNA motif. any: any orientation (i.e. without considering orientation) of DNA motif. others: enhancer-promoter association not shortened at the genomic locations of DNA motif. FR H3K27ac: EPI were predicted using DNA motif in open chromatin regions overlapped with H3K27ac histone modification marks. *

p-value < 10^{-9}.

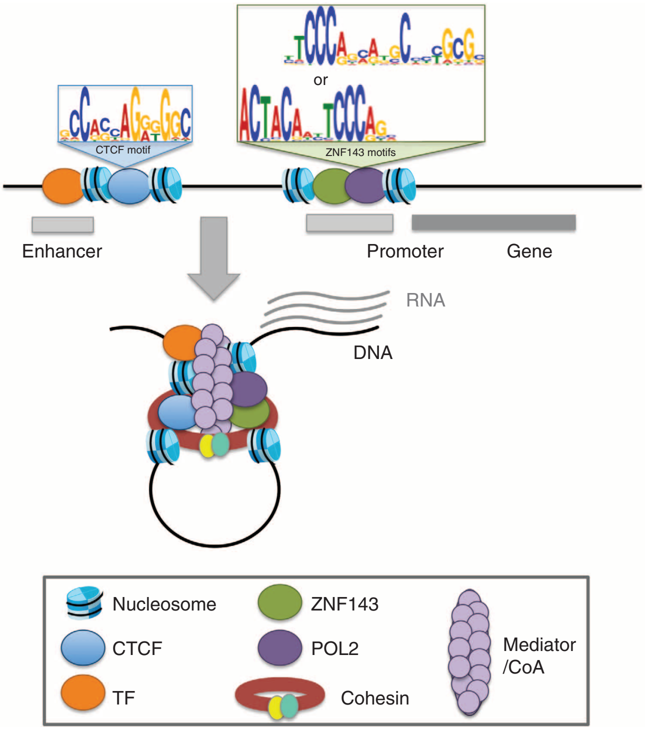

Figure 4. Schematic representation of chromatin interactions involving gene promoters. ZNF143 contributes the formation of chromatin interactions by directly binding the promoter of genes establishing looping with distal element bound by CTCF (Bailey et al. 2015).

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
SRF	74.05
YY1	71.30
ZNF658	70.76
IRF7	65.05
PHYPADRAFT	63.98
ESE3	60.18
GCM1:CEBPB	58.35
STAT1	55.66
STA5A	54.84
RAD21	54.11
TFE2	53.78
ZNF623	53.59
ZNF317	52.78
ZNF93	52.13
CTCFL	50.49
HMBOX1	49.33
CTCF	47.24
ZKSC3	46.61
STAT4	45.64
E2F	45.63
SMAD2:SMAD3:SMAD4	43.91
ABR1	43.81
P53	40.09
BPTF	39.06
RFX5	38.00
HME1	37.88
NR1I2	37.70
KLF1	37.61
AT4G28140	37.47
ZKSCAN1	37.45
EVI1	37.37
ELK1:PAX1	36.54
RFX2	36.34
SP2	36.15
ELF2	35.92
NFIC	34.36
HSF2	34.27
KLF14	33.83
ZF5	33.79
FXR	33.36
EPAS1	33.08
USF	33.07
ZNF646	32.47
ERR1	32.30
NEUROD1	32.08
E2F8	31.95
ZNF660	31.55
ZN219	31.45
TCFAP2C	31.08
MYB	30.93
HOXD12:ELK1	30.47
THAP1	30.20
GATA1	30.17
STAT3	29.85
MXI1	29.73
SOX10	29.57
HAT5	29.56
PURA	29.03
SMC3	28.59
NFYB	28.50
PAX5	28.42
RUNX2	27.95
AP1	27.62
ZNF836	27.60
IRF	27.56
NR2F2	27.44
NR2C2	27.02
IKZF1	26.87
ESR2	26.64
HIC1	26.58
RREB1	26.56
TCF12	26.29
FOS:JUN	26.07
CMYB	25.86
NKX61	25.41
TFAP2A	25.40
NR4A3	25.37
ZNF709	25.15
GTF3C2	25.04
SMAD5	24.40
WT1	24.19
HIC2	24.16
ZNF695	24.07
ZSC31	23.98
GLIS1	23.79
TCF3	23.55
ZNF284	23.53
PGAM2	23.44
ZFP82	23.35
ZNF214	23.32

Reverse-forward orientation

TF	Score	TF	Score	TF	Score
HNF1B	77.23	FOXI1	37.39	MSX2	26.00
STAT4	73.89	RARA	36.47	HOXC8	25.97
ZNF28	71.72	ZNF195	35.51	RXRA	25.45
TF7L1	67.52	ZNF449	34.50	MYOG	25.21
ZNF225	61.87	DOBOX5	34.48	SMAD4	25.15
CTCF	53.54	LEF1	33.87	PO3F2	24.99
STAT3	53.06	CETS2	33.69	IRF4	24.96
STAT5A	52.61	E2F3	33.22	ZNF233	24.53
PAX5	52.10	NKX3	33.07	RFX3	24.45
SRF	51.83	RBPJ	32.95	LDSPOLYA	23.76
ZNF670	50.69	SP1	32.79	ZNF682	23.56
ETV7	50.48	ZNF524	32.31	RXRB	23.34
TEL1	49.47	CDX2	32.10	EVI1	23.14
BCL6	49.19	TFAP4:ETV1	31.94	ETS2	22.98
ETV6	48.78	SOX2	30.80	OBOX2	22.95
NFY	47.73	SOX9	30.72	ZNF770	22.85
ZNF286B	46.85	PRDM1	30.66	ERG	22.75
ZFP82	46.84	CD59	29.85	TCF3	22.67
ZNF343	45.78	SOX11	29.34	IRF1	22.44
ZNF681	45.36	ZN320	28.86	TAACC	22.31
STAT5B	44.72	MYC	28.58	ZNF687	22.26
KLF16	44.33	ZNF316	28.34	IRF3	22.13
HNF4A	42.24	NR2F1	28.34	NRSE	21.97
P73	41.49	ZNF121	28.26	P53	21.92
EOMES	40.45	HOXB2:NHLH1	28.18	TCFAP2B	21.78
WT1	40.11	E2F2	27.94	AHR	21.40
BCL3	39.66	BDP1	27.63	ZFP641	20.09
SP8	38.96	HSF1	27.03	NRF3	19.94
TEAD1	38.73	ZBTB2	26.79	ZNF76	19.81
E2F4	37.67	SPI1	26.15	BRF2	19.77

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

HUVEC

TF	Score
PAX5	19.30
TBX5	14.60
NANOG	12.98
BATF	11.52
HNF1A	10.94
THAP1	10.45
NR2C2	9.67
SULT1A2	9.67
CTCF	9.50
RAD21	9.01
COT2	6.84
MYB	6.18
EKLF	6.17
PIF7	5.95
ZNF695	5.90
SMAD5	5.70
TCFAP2A	5.52
YY1	5.50
YBX1	5.37
CMYB	5.26
HME1	5.07
TCF12	5.06
TFAP2A	4.83
HNF4A	4.36
SMC3	4.20
RXRA:VDR	3.64
P53	3.60
STA5A	3.40
STAT4	3.36
ZNF71	2.93

MCF-7

TF	Score
E2F6	57.73
EGR1	48.41
ABR1	34.79
CACBINDINGPROTEIN	28.34
KLF16	28.16
RREB1	25.42
PO4F1	24.64
ZN219	23.13
SP1	22.41
SP2	21.88
KLF15	21.33
ZNF521	21.05
CTCF	20.02
ESE3	19.92
ZNF143	17.47
FOXP1	16.99
GTF2I	16.32
ERF9	15.59
TF3A	15.51
TCF3	14.03
HIC2	13.94
GKLF	13.80
CXXC1	13.75
SMC3	12.91
CTCFL	12.67
RAD21	12.00
STA5B	11.94
STAT1	11.78
KLF12	11.57
SP4	11.20

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

Forward-reverse (FR) orientation (using H3K27ac histone modification marks)

Repeat DNA seq.	Score
MLT1F1	6.75

Reverse-forward (RF) orientation

Repeat DNA seq.	Score
LTR16C	69.39
L1ME4b	23.90
AluSg	23.61

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 sequence of CTCF with another biased orientation of DNA motif sequence were shown in a separate table. Motif 1,2: DNA binding motif sequences of transcription factors. # both: 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 and Motif 2. Open chromatin regions overlapped with histone modification marks (H3K27ac) were used (Total 26,095 regions).

Motif 1	Motif 2	# both	# motif 1	# motif 2	# others	*p*-value
MAZ	RREB1	7506	809	1770	16010	0
KLF5	MAZ	7181	1151	2095	15668	0
KLF13	MAZ	7128	881	2148	15938	0
MAZ	ZN148	7079	1027	2197	15792	0
MAZ	PATZ1	6931	1088	2345	15731	0
CDA7L	MAZ	6837	1050	2439	15769	0
KLF13	RREB1	6802	1513	1207	16573	0
MAZ	SP1	6698	811	2578	16008	0
CDA7L	KLF5	6668	1219	1664	16544	0
KLF5	ZN148	6644	1688	1462	16301	0
KLF11	MAZ	6548	802	2728	16017	0
PATZ1	RREB1	6538	1777	1481	16299	0
RREB1	ZN148	6521	1794	1585	16195	0
KLF11	KLF5	6446	904	1886	16859	0
MAZ	WT1	6444	1005	2832	15814	0
KLF5	SP1	6337	1995	1172	16591	0
CDA7L	ZN148	6313	1574	1793	16415	0
MAZ	SMAD2	6297	1168	2979	15651	0
SP1	ZN148	6292	1217	1814	16772	0
KLF5	SMAD2	6281	2051	1184	16579	0
CDA7L	KLF11	6232	1655	1118	17090	0
RREB1	SP1	6219	2096	1290	16490	0
WT1	ZN148	6205	1901	1244	16745	0
PATZ1	WT1	6145	1874	1304	16772	0
RREB1	WT1	6103	2212	1346	16434	0
KLF13	SP1	6084	1425	1925	16661	0
CDA7L	SP1	6051	1836	1458	16750	0
CDA7L	SMAD2	6034	1853	1431	16777	0
KLF11	RREB1	6034	1316	2231	16514	0
KLF11	ZN148	6030	1320	2076	16669	0

Motif 1	Motif 2	# both	# motif 1	# motif 2	# others	*p*-value
CTCF	MAZ	3371	671	5905	16148	0
CTCF	KLF5	3344	698	4988	17065	0
CTCF	RREB1	3126	741	5189	17039	0
CTCF	RAD21	2456	1586	166	21887	0
CTCF	CTCFL	1519	137	1078	23361	0
CTCF	RXRA	1459	336	1641	22659	0
CTCF	SMC3	1036	170	1561	23328	0
CTCFL	MAZ	2462	469	6814	16350	0
CTCFL	KLF5	2351	5981	580	17183	0
CTCFL	RAD21	1371	285	433	24006	0
CDA7L	CTCF	3238	4649	804	17404	0

Table 5. Comparison of enhancer-promoter interactions (EPI) with chromatin interactions in T cells. The upper tables show the number of biased orientation of DNA motifs, where a significantly higher ratio of EPI, which were predicted based on enhancer-promoter association (EPA) (iii), overlapped with HiChIP chromatin interaction data than the other types of EPA (i) and (ii). The middle tables show the 70 FR and 73 RF orientations of DNA motifs found in common among B2T1, B2T2 and B3T1 replications in the upper table. The lower table shows that among 43 biased orientation of DNA motifs found in both monocytes and T cell, 31 were matched with the analysis of HiChIP for three types of EPA. TF: DNA binding motif sequence of a transcription factor. Score: –$\log_{10}$ (*p*-value). Inf: *p*-value = 0. Score 1: Comparison of EPA shortened at the genomic locations of FR or RF orientation of DNA motif sequence [EPA (iii)] with EPA not shortened [EPA (i)]. Score 2: Comparison of EPA shortened at the genomic locations of FR or RF orientation of DNA motif sequence [EPA (iii)] with EPA shortened at the genomic locations of any orientation of DNA motif sequence [EPA (ii)].

1. Comparison of EPI with HiChIP data among three EPA (i), (ii) and (iii)

Replication of HiChIP data	Total no. (FR + RF) of DNA motifs	No. of FR DNA motifs	No. of RF DNA motifs
B2T1	165	86	90
B2T2	168	85	92
B3T1	189	96	105
DNA motifs found in common among replications			
B2T1 and B2T2	148	75	81
B2T1 and B3T1	148	48	81
B2T2 and B3T1	147	76	79
B2T1, B2T2 and B3T1	136	70	73

2. Comparison of EPI with HiChIP data between two EPA (i) and (iii)

Replication of HiChIP data	Total no. (FR + RF) of DNA motifs	No. of FR DNA motifs	No. of RF DNA motifs
B2T1	358	195	201
B2T2	360	196	203
B3T1	365	195	209
DNA motifs found in common among replications			
B2T1 and B2T2	357	195	200
B2T1 and B3T1	357	194	201
B2T2 and B3T1	358	194	203
B2T1, B2T2 and B3T1	356	194	200

Forward-reverse (FR) orientation

TF	Score 1	Score 2
ABL1	120.20	4.48
AP2GAMMA	64.98	6.76
BATF	104.50	2.89
BC11A	114.77	4.72
CDC5L	46.20	12.32
CEBPG	113.24	8.73
CHOP	323.31	2.83
CTCF	109.95	16.53
CTCFL	89.06	16.88
CXXC1	98.89	10.97
DEC1	105.31	6.77
EGR1	129.00	12.41
EGR2	97.63	3.69
EGR3	105.43	4.92
EKLF	98.41	10.61
ELF5	114.76	3.90
ER71	246.20	2.48
ERF:EOMES	323.31	15.05
ETV4	206.76	7.18
FOXA2	323.31	3.48
FOXO3	323.31	3.71
GABPBETA	187.48	4.43
GATA5	176.92	1.89
GFI1	323.31	10.70
GFI1B	252.86	12.43
HIC1	99.40	1.96
HMGA2	322.83	7.48
INSM2	323.31	2.54
JUN	184.60	16.41
JUNB	158.94	2.57
KLF1	188.04	1.92
LHX8	323.31	3.39
MAF	136.76	6.56
MAFK	42.30	2.43
MYBB	253.52	8.74
NFAC2	102.90	10.64
NFAC3	117.02	2.13
NFAC4	129.06	11.87
NFAT5	169.03	10.79
P50:P50	147.45	6.59
PRDM6	243.21	13.67
RAD21	98.81	10.93
REL	100.28	8.06
RFX5	138.49	2.53
RREB	98.91	4.13
RXRA	99.72	3.24
SIN3A	39.49	6.17
SMC3	78.49	20.74
SMRC2	106.88	8.79
SNF5	69.86	2.27
SP5	58.99	5.59
SRP9	147.33	2.92
STAF	201.83	9.43
SUZ12	102.37	7.25
TAL1	109.90	13.57
TF65	118.80	4.02
VEZF1	43.90	3.26
YBOX1	91.34	4.41
YBX1	165.05	4.58
ZBP89	82.45	6.26
ZBRK1	323.31	4.03
ZBT16	157.35	20.95
ZN263	63.43	8.19
ZNF227	323.31	3.42
ZNF253	Inf	18.36
ZNF331	120.46	2.13
ZNF585A	179.25	8.49
ZNF681	Inf	35.47
ZNF721	148.23	3.75
ZNF846	Inf	11.20

Reverse-forward (RF) orientation

TF	Score 1	Score 2
AML3	94.46	2.21
AP1	71.45	6.80
ARI5B	57.76	2.72
CFOS	121.76	12.76
CFOS:CJUN	194.70	21.29
CHD2	76.76	2.22
CREB3	166.87	18.45
CREM	323.31	6.89
CTCF	74.93	18.84
CTCFL	60.12	12.79
DBP	263.45	5.42
E2F1	122.59	4.07
EFC	323.31	4.05
EGR1	263.41	5.42
EGR4	102.64	2.88
ELF2	95.04	3.38
ELF5	279.29	1.66
EP300	82.68	3.97
ETV2:PAX5	323.31	6.03
FOSB	146.18	14.44
FOSL1	137.95	13.98
FOXA	323.31	7.27
FOXA1	323.31	9.93
FOXO3A	323.31	6.53
GATA	209.11	8.22
GCM2:EOMES	56.33	4.61
GFI1	123.54	4.62
GFI1B	323.31	5.94
HSF1	44.39	3.11
HSF2	144.04	5.18
HXB1	323.31	3.91
HXB6	323.31	5.57
IKZF1	65.38	6.13
JUN:FOSB	77.62	2.75
JUN:JUNB	126.45	3.73
LYL1	74.51	2.65
MTF1	303.27	1.86
MZF1	100.14	3.15
NKX2-1	323.31	7.34
NKX2-4	Inf	5.82
NKX2-5	323.31	4.23
P50	185.48	10.77
P53	71.49	6.47
P73	172.99	8.20
PKNX1	111.56	2.61
POU2F1:ELK1	323.31	4.66
RELA	306.11	4.11
RP58	210.42	3.73
RREB1	128.85	2.05
RXRB	57.75	2.29
SALL4	100.01	13.24
SIX6	96.78	11.15
TCF3	220.95	6.37
TCF4	64.91	2.34
TCFE2A	212.46	6.00
TF7L2	120.84	3.14
XBP1	105.73	2.57
ZBTB4	323.31	5.23
ZFP2	Inf	5.32
ZNF134	53.47	1.70
ZNF16	243.81	3.54
ZNF260	98.46	2.69
ZNF28	323.31	4.61
ZNF282	126.15	5.08
ZNF341	109.16	4.93
ZNF468	105.25	2.85
ZNF484	Inf	12.66
ZNF542P	244.54	5.52
ZNF580	78.41	2.95
ZNF625	Inf	6.14
ZNF714	39.80	3.57
ZNF721	111.46	4.37
ZNF837	87.54	1.74

TF	Score 1	Score 2
BATF	104.50	2.89
CTCF	109.95	16.53
CTCFL	89.06	16.88
CXXC1	98.89	10.97
EGR1	129.00	12.41
EKLF	98.41	10.61
ELF5	114.76	3.90
EP300	82.68	3.97
ETV4	206.76	7.18
FOSL1	137.95	13.98
GFI1	323.31	10.70
HIC1	99.40	1.96
HSF1	44.39	3.11
HSF2	144.04	5.18
KLF1	188.04	1.92
MAFK	42.30	2.43
MYBB	253.52	8.74
P53	71.49	6.47
P73	172.99	8.20
RAD21	98.81	10.93
RFX5	138.49	2.53
RREB1	98.91	4.13
RXRA	99.72	3.24
RXRB	57.75	2.29
SMC3	78.49	20.74
TAL1	109.90	13.57
TCF3	220.95	6.37
TF65	118.80	4.02
YBX1	165.05	4.58
ZNF28	323.31	4.61
ZNF721	111.46	4.37

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.

Bejerano G, Lowe CB, Ahituv N, King B, Siepel A, Salama SR, Rubin EM, Kent WJ, Haussler D. 2006. A distal enhancer and an ultraconserved exon are derived from a novel retroposon. *Nature* **441**: 87-90.

Chen M, Manley JL. 2009. Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. *Nat Rev Mol Cell Biol* **10**: 741-754.

Das D, Clark TA, Schweitzer A, Yamamoto M, Marr H, Arribere J, Minovitsky S, Poliakov A, Dubchak I, Blume JE et al. 2007. A correlation with exon expression approach to identify cis-regulatory elements for tissue-specific alternative splicing. *Nucleic acids research* **35**: 4845-4857.

de Souza FS, Franchini LF, Rubinstein M. 2013. Exaptation of transposable elements into novel cis-regulatory elements: is the evidence always strong? *Mol Biol Evol* **30**: 1239-1251.

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.

Javierre BM, Burren OS, Wilder SP, Kreuzhuber R, Hill SM, Sewitz S, Cairns J, Wingett SW, Varnai C, Thiecke MJ et al. 2016. Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell* **167**: 1369-1384 e1319.

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*.

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.

Jin C, Kato K, Chimura T, Yamasaki T, Nakade K, Murata T, Li H, Pan J, Zhao M, Sun K et al. 2006. Regulation of histone acetylation and nucleosome assembly by transcription factor JDP2. *Nat Struct Mol Biol* **13**: 331-338.

Jjingo D, Conley AB, Wang J, Marino-Ramirez L, Lunyak VV, Jordan IK. 2014. Mammalian-wide interspersed repeat (MIR)-derived enhancers and the regulation of human gene expression. *Mob DNA* **5**: 14.

John S, Sabo PJ, Thurman RE, Sung MH, Biddie SC, Johnson TA, Hager GL, Stamatoyannopoulos JA. 2011. Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nature genetics* **43**: 264-268.

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.

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.

Katainen R, Dave K, Pitkanen E, Palin K, Kivioja T, Valimaki N, Gylfe AE, Ristolainen H, Hanninen UA, Cajuso T et al. 2015. CTCF/cohesin-binding sites are frequently mutated in cancer. *Nature genetics* **47**: 818-821.

Kheradpour P, Kellis M. 2014. Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic acids research* **42**: 2976-2987.

Kulakovskiy IV, Vorontsov IE, Yevshin IS, Sharipov RN, Fedorova AD, Rumynskiy EI, Medvedeva YA, Magana-Mora A, Bajic VB, Papatsenko DA et al. 2018. HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic acids research* **46**: D252-D259.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)* **25**: 1754-1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)* **25**: 2078-2079.

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.

Rebollo R, Romanish MT, Mager DL. 2012. Transposable elements: an abundant and natural source of regulatory sequences for host genes. *Annu Rev Genet* **46**: 21-42.

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

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research* **13**: 2498-2504.

Tabuchi TM, Deplancke B, Osato N, Zhu LJ, Barrasa MI, Harrison MM, Horvitz HR, Walhout AJ, Hagstrom KA. 2011. Chromosome-biased binding and gene regulation by the Caenorhabditis elegans DRM complex. *PLoS Genet* **7**: e1002074.

Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. 2008. Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**: 470-476.

Wang J, Vicente-Garcia C, Seruggia D, Molto E, Fernandez-Minan A, Neto A, Lee E, Gomez-Skarmeta JL, Montoliu L, Lunyak VV et al. 2015. MIR retrotransposon sequences provide insulators to the human genome. *Proc Natl Acad Sci U S A* **112**: E4428-4437.

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.