Discovery of directional chromatin-associated regulatory motifs affecting human gene transcription

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

Abstract

Background: Chromatin interactions are essential in enhancer-promoter interactions (EPIs) and transcriptional regulation. CTCF and cohesin proteins located at chromatin interaction anchors and other DNA-binding proteins such as YY1, ZNF143, and SMARCA4 are involved in chromatin interactions. However, there is still no good overall understanding of proteins associated with chromatin interactions and insulator functions.

Results: Here, I describe a systematic and comprehensive approach for discovering DNA-binding motifs of transcription factors (TFs) that affect EPIs and gene expression. This analysis identified 96 biased orientations [64 forward-reverse (FR) and 52 reverse-forward (RF)] of motifs that significantly affected the expression level of putative transcriptional target genes in monocytes, T cells, HMEC, and NPC and included CTCF, cohesin (RAD21 and SMC3), YY1, and ZNF143; some TFs have more than one motif in databases; thus, the total number is smaller than the sum of FRs and RFs. KLF4, ERG, RFX, RFX2, HIF1, SP1, STAT3, and AP1 were associated with chromatin interactions. Many other TFs were also known to have chromatin-associated functions. The predicted biased orientations of motifs were compared with chromatin interaction data. Correlations in expression level of nearby genes separated by the motif sites were then examined among 53 tissues.

Conclusion: One hundred FR and RF orientations associated with chromatin

interactions and functions were discovered. Most TFs showed weak directional biases at chromatin interaction anchors and were difficult to identify using enrichment analysis of motifs. These findings contribute to the understanding of chromatin-associated motifs involved in transcriptional regulation, chromatin interactions/regulation, and histone modifications.

Keywords: transcriptional regulation, transcription factors, enhancer-promoter interactions, chromatin interactions, CTCF, open chromatin regions, histone modifications

Background

Chromatin interactions play essential roles in enhancer-promoter interactions (EPIs) and transcriptional regulation. CTCF and cohesin proteins are located at the chromatin interaction anchors and facilitate the formation of loop structures. CTCF functions as an insulator by limiting the activity of enhancers in the loops (Fig. 1a) [1, 2]. The DNA-binding sequence of CTCF reveals its orientation bias at chromatin interaction anchors in terms of the forward-reverse (FR) orientation [1, 3, 4]. Other DNA-binding proteins such as ZNF143, YY1, and SMARCA4 (BRG1) are involved in chromatin interactions and EPIs [5-7]. ZNF143 is known to bind more often to promoters than distal regions, thereby establishing looping with the distal element bound by CTCF [5]. The DNA-binding motif of ZNF143 is enriched at chromatin

interaction anchors (Z-score > 7), whereas the DNA-binding motif of YY1 is not enriched (Z-score < 2; [5] Figure 2a). CTCF, cohesin, ZNF143, YY1, and SMARCA4 are involved in various biological functions other than chromatin interaction. These transcription factors (TFs) bind to open chromatin regions near transcriptional start sites (TSS) as well, which may reduce the enrichment of the motifs at chromatin interaction anchors.

Computational methods to predict chromatin interactions have been developed [8, 9]. However, they were not intended to find DNA motifs of TFs involved in chromatin interactions. Xie et al. [10] proposed that the next challenge will be to develop systematic methods to discern the specific functions of these motifs in a genome-wide fashion.

DNA-binding proteins involved in chromatin interactions are considered to affect the transcription of genes within the loops formed by chromatin interactions. To confirm this hypothesis, I previously found that the expression level of human putative transcriptional target genes of a TF was significantly changed according to enhancer-promoter association (EPA) rules [11]. A regulatory domain for EPA was limited within the FR orientation of CTCF-binding sites, and the transcriptional target genes of each TF bound to enhancers and promoters were predicted using the EPA domain (Fig. 1b and c) [12]. The expression levels predicted using the EPA domain differed significantly from the expression levels predicted using promoters alone (or closed chromatin regions, where enhancers would be silenced, and nearby genes would express at low level [13]). The expression levels tended to increase in monocytes and

$CD4^+$ T cells, implying that TFs serve as activators of target genes. The expression levels tended to decrease in embryonic stem (ES) and induced pluripotent stem (iPS) cells, implying that TFs serve as repressors of target genes. The expression level of putative transcriptional target genes of a TF that binds enhancers is changed significantly when EPIs are predicted properly using the insulator sites of CTCF in different cell types. This observation suggests that the difference in the expression level reflects the accuracy of prediction of transcriptional target genes using a given EPA domain. Using this approach, I planned to search for other DNA-binding proteins involved in chromatin interactions, which affect the expression level of putative transcriptional target genes through their insulator function.

Experimental analyses of chromatin interactions in molecular biology have technical limitations. Chromatin interaction data are altered according to experimental techniques, the number of DNA sequencing reads, and even replication sets of the same cell type. Additionally, chromatin interaction data may not be abundant enough to contain all chromatin interactions and may include experimental noise such as ligation errors. To compensate the technical limitations of experimental analyses, a computational search was conducted for DNA motifs of TFs that affect EPIs and the expression level of putative transcriptional target genes in monocytes, T cells, human mammary epithelial cells (HMEC), and neural progenitor cells (NPC) without using chromatin interaction data. Then, I compared putative EPIs and transcription factor binding sites (TFBSs) with chromatin interaction and gene expression data.

Results

Search for biased orientations of DNA motifs

TFBSs were predicted using open chromatin regions and DNA motifs of TFs collected from various databases and journals (see Methods). Transcriptional target genes were predicted based on TFBSs in promoters and enhancers. An EPA domain was limited within the TFBSs with insulator function, such as CTCF and cohesin (RAD21 and SMC3). To find motifs of TFs with insulator function other than CTCF and cohesin, an EPA domain was limited within the TFBSs of a putative insulator TF (Fig. 1c). Transcriptional target genes of each general TF bound in (i) distal open chromatin regions, which include enhancers, and (ii) closed chromatin regions in promoters were predicted using the EPA domain in order to estimate the enhancer activity of each general TF. The ratio of median expression level of target genes of each general TF between (i) and (ii) was calculated. The distribution of the ratios of all general TFs was compared among forward-reverse (FR), reverse-forward (RF) orientations, and non-oriented (i.e., without considering orientation) groups of TFBSs using a two-sided Mann-Whitney test (p-value $< 10^{-7}$). Putative insulator TFs with significant differences in the distributions of the ratios of expression level among FR, RF, and non-oriented were enumerated (Fig. 2a). TFBSs overlapping H3K27ac ChIP-seq peaks showed a significant difference in the expression level of their putative target genes between (i) and (ii), suggesting that the TFBSs were associated with enhancer (and promoter) signals of H3K27ac (Fig. 2a).

To confirm the predictions of TFBSs, the distributions of TFBSs were examined

121 in the human genome. TFBSs involved in chromatin interactions were identified near
122 TSS (Fig. 2b). ZNF143 is known to bind to the promoter of genes, thereby establishing
123 looping with the distal element bound by CTCF [5]. YY1 is a multi-functional
124 transcriptional regulator and is involved in transcriptional activation, repression, and
125 initiation [14, 15]. The distributions of TFBSs seemed to be consistent with the features
126 of the TFs. Most TFBSs were identified in intergenic and intron regions of the human
127 genome (Fig. 2c).

128 Several hundreds of FR and RF orientations were found in monocytes, T cells,
129 HMEC, and NPC, which included DNA motifs with weak biased orientations (Fig. 3).
130 The number of TFs was likely to be changed, according to the number of DNase-seq
131 reads, since the number of the reads in monocytes was larger than those of the other
132 three cell types. When DNase-seq reads increase, more TFBSs may be predicted to
133 some extent. Notably, 96 biased (64 FR and 52 RF) orientations of DNA-binding motifs
134 of TFs were discovered and found to be common between the four cell types, whereas
135 non-oriented DNA-binding motifs of TFs were not detected (Fig. 3; Table 1; Additional
136 file 1: Table S1). The FR orientation of DNA motifs included CTCF, cohesin (RAD21
137 and SMC3), YY1, and ZNF143, which are involved in chromatin interactions and EPIs.
138 Some TFs have both FR and RF orientations of DNA motifs; thus, the total number 96
139 is smaller than the sum of FRs and RFs. Palindromic DNA motifs of TFs do not have
140 FR and RF distinction; thus, they do not show a directional bias.

141 Among the 96 TFs, KLF4 [16], ERG [17, 18], RFX [19], RFX2 [20], HIF1 [21],
142 SP1 [22], STAT3 [23], and AP1 [24] were found to be associated with chromatin

interactions. An important example is KLF4, which is associated with three-dimensional enhancer rewiring and transcriptional changes during the reprogramming of mouse embryonic fibroblasts to pluripotent stem cells [16]. ATF3 [25], ZBTB7A [26], NRF1 [6, 27], TFAP2 [28], Myc [29, 30], the ETS family, ELK3 [31], ELK4 [32], RUNX3 [33], USF1 [34, 35], AHR/ARNT [36], SP2 [37], EGR2 [38-40], ZBTB33 [41], NF-Y [42, 43], MITF [44, 45], NONO [46], FLI1 [47], E2F [48-50], BCL [51], CXXC1 [52], E2F1 [53], EGR1 [54], REST [55, 56], RFX [57], SIN3A [58-60], AP1 [61], and STAT3 and STAT5 [62] were associated with chromatin remodeling, chromatin accessibility, pioneer factor, and/or histone modification. TFBSs of the TFs other than those involved in chromatin interactions would affect gene expression by chromatin-associated functions. Huang et al. [34] have mentioned that the insulator element at the 5' end of the chicken beta-globin locus acts as a barrier, thus protecting transgenes against silencing effects of adjacent heterochromatin. The transcription factor USF1 binds within the insulator at a site important in adjacent nucleosomes for generating histone modifications associated with active chromatin and, by inference, with barrier function.

The biased orientations of TFBSs might be located near those of known TFs involved in chromatin interactions such as CTCF, RAD21, SMC3, YY1, and ZNF143; thus, they showed significant differences in the expression level of putative transcriptional target genes among FR, RF, and non-orientation. However, among 447 FR and 496 RF orientations of DNA motifs in monocytes, the TFBSs of only 16 (4%) FR and 28 (6%) RF orientations were located within 1 kb of ≥ 20% of the TFBSs of either of the known TFs and their associated TFs such as CTCFL and YY2 (Additional

file 1: Table S2).

To examine whether the biased orientations of TFBSs were associated with CTCF-binding sites, an EPA domain was limited within the TFBSs of biased orientations of a TF in the 96 TFs found to be common among the four cell types. Then, transcriptional target genes of each general TF were predicted from (i) open chromatin regions in the EPA domain that did not include CTCF-binding sites upstream and downstream of a gene, and (ii) closed chromatin regions in promoters. The ratio of expression level between (i) and (ii) was calculated. The biased orientations of CTCF consist of 19 DNA-binding motifs in monocytes. When upstream and downstream of genes with the TFBSs of each DNA-binding motif sequence of CTCF were eliminated, 93, 92, 93, and 84 TFs out of 95, except for CTCF, showed a significant difference in the distribution of the ratios of expression level (p-value $< 10^{-7}$) in monocytes, $CD4^+$ T cells, HMEC, and NPC, respectively (Additional file 1: Table S3). This suggests that the biased orientations of TFBSs are not significantly associated with CTCF binding-sites.

Comparison with chromatin interaction data

To examine whether the biased orientations of DNA motifs were associated with chromatin interactions, EPIs predicted using an EPA domain were compared with chromatin interaction data, using three replications (B2T1, B2T2, and B3T1) of H3K27ac HiChIP chromatin interaction data in $CD4^+$ T cells separately [63]. The resolutions of HiChIP chromatin interaction data and EPIs were adjusted to 5 kilobases (kb). The number of unique HiChIP cis (i.e., in the same chromatin) chromatin

interactions was 19,926,360 at 5-kb resolution in B2T1 replication of T cells; 666,149 at 5-kb resolution with ≥1,000 counts for each interaction; and 78,209 at 5-kb resolution with ≥6,000 counts for each interaction (Additional file 1: Table S4). EPIs were predicted using three types of EPA domains: (i) EPA domain limited within the FR or RF orientation of TFBSs of a TF with putative insulator function such as CTCF, (ii) EPA domain limited within the TFBSs of a TF without considering their orientation, and (iii) EPA domain which was not limited within TFBSs. Despite the limited resolution of the chromatin interaction data, a total of 201 biased orientations [133 of 273 (49%) FR and 90 of 211 (43%) RF; some TFs have both FR and RF motifs; thus, the total number is smaller than the sum of FRs and RFs] of DNA motifs in T cells included CTCF, cohesin (RAD21 and SMC3), ZNF143, and YY1 in three replications (B2T1, B2T2, and B3T1). For the DNA motifs, a significantly higher ratio of EPIs overlapped with HiChIP chromatin interactions with ≥1,000 counts for each interaction in EPA domain (i) than in EPA domain (ii) and (iii) in T cells (Table 2). For a total of 390 biased orientations [261 of 273 (96%) FR and 199 of 211 (94%) RF] of DNA motifs in T cells, a significantly higher ratio of EPIs overlapped with the chromatin interactions in EPA domain (i) than EPA domain (iii) (Table 2; Additional file 1: Table S4 and S5). The difference between EPIs predicted using EPA domain (i) and (ii) seemed to be difficult to distinguish using the chromatin interaction data and a statistical test in some cases. However, for the difference between EPIs predicted using EPA domain (i) and (iii), a larger number of biased orientations of DNA motifs were correlated with chromatin interaction data. Most (95%) of the biased orientations of DNA motifs were correlated

with chromatin interactions when comparing EPIs predicted using EPA domain (i) and (iii) with HiChIP chromatin interactions. Chromatin interaction data were obtained from different samples of DNase-seq open chromatin regions; thus, individual differences may exist. Individual differences of chromatin interactions were larger than those of open chromatin regions [63]. Additionally, the analyses for Th17, Treg, and GM12878 were shown in Additional file 2.

In situ Hi-C data in HMEC, which are available in a 4D nucleome data portal, were compared with putative EPIs. The number of unique in situ Hi-C cis chromatin interactions was 121,873,642 at 3-kb resolution. For a total of 222 biased orientations [155 of 310 (50%) FR and 108 of 269 (40%) RF] of DNA motifs, which included CTCF, cohesin (RAD21 and SMC3), ZNF143, and YY1, a significantly higher ratio of EPIs overlapped with in situ Hi-C chromatin interactions (≥1 count for each interaction, i.e., all interactions) in EPA domain (i) than in EPA domain (ii) and (iii) (Additional file 1: Table S7). When comparing EPIs predicted using only EPA domain (i) and (iii) with the chromatin interactions, for a total of 406 biased orientations [268 of 310 (86%) FR and 230 of 269 (86%) RF] of DNA motifs, a significantly higher ratio of EPIs overlapped with the chromatin interactions in EPA domain (i) than EPA domain (iii) (Additional file 1: Table S7). There were six replications of in situ Hi-C experimental data of HMEC, but other replications contained much smaller numbers of chromatin interactions. The HiChIP chromatin interaction data in T cells were expected to be enriched with EPIs, since they were produced with H3K27ac antibody to capture chromatin interactions. In situ Hi-C data consist of all types of chromatin interactions and thus contain a smaller

number of EPIs than HiChIP data. Chromatin interactions with a small number of counts may be artifacts caused by ligation errors during experiments [64, 65]. Therefore, other replications of in situ Hi-C data with a smaller number of chromatin interactions will not be available for this analysis.

Promoter capture Hi-C data of NPC were downloaded and analyzed by the same procedure as that for T cells and HMEC [66]. The number of unique promoter capture Hi-C cis chromatin interactions was 9,436,689 at 5-kb resolution. For a total of 109 biased orientations [61 of 205 (28%) FR and 60 of 241 (25%) RF] of DNA motifs, which included CTCF, cohesin (RAD21 and SMC3), ZNF143, and YY1, a significantly higher ratio of EPIs overlapped with promoter capture Hi-C chromatin interactions (≥1 count for each interaction, i.e. all interactions) in EPA domain (i) than in EPA domain (ii) and (iii) (Additional file 1: Table S8). When comparing EPIs predicted using only EPA domain (i) and (iii) with the chromatin interactions, for a total of 186 biased orientations [99 of 205 (48%) FR and 103 of 241 (43%) RF] of DNA motifs, a significantly higher ratio of EPIs overlapped with the chromatin interactions in EPA domain (i) than EPA domain (iii) (Additional file 1: Table S8).

All three types of chromatin interaction data (HiChIP, in situ Hi-C, and promoter capture Hi-C) indicated that EPIs predicted using EPA domain (i) that was limited within TFBSs of known TFs involved in chromatin interactions (CTCF, RAD21, SMC3, ZNF143 and YY1) overlapped with more chromatin interactions than those using EPA domain (ii) and (iii) in six cell types (T cells, Th17, Treg, GM12878, HMEC, and NPC) (the analyses for Th17, Treg, and GM12878 were shown in Additional file 2). In

monocytes, Hi-C data was available, but it was not analyzed due to the low resolution of chromatin interaction data (50 kb).

The biased orientations of TFBSs of known TFs involved in chromatin interactions such as CTCF, RAD21, SMC3, YY1, and ZNF143 might overlap with the biased orientations of TFBSs of other TFs; thus, they showed a significantly higher ratio of EPIs overlapped with chromatin interactions in EPA domain (i) than EPA domain (ii) and (iii). However, among 273 FR and 211 RF orientations of DNA motifs of TFs not including the known TFs in T cells, the TFBSs of only 15 (6%) FR and 9 (4%) RF orientations of DNA motifs of TFs overlapped with $\geq$ 60% of the TFBSs of biased orientations of DNA motifs of either of the known TFs at 5-kb resolution.

Moreover, to investigate the enhancer activity of EPIs, the distribution of expression level of putative target genes of EPIs was compared between EPIs that overlapped with HiChIP chromatin interactions and EPIs that did not overlap with them. Though the target genes of EPIs were selected from top 4,000 transcripts (genes) in terms of the expression level of all transcripts (genes) excluding transcripts not expressed in T, Th17, Treg, and GM12878 cells separately, target genes of EPIs that overlapped with chromatin interactions showed a significantly higher expression level than non-overlapped EPIs, suggesting that EPIs that overlapped with chromatin interactions activated the expression of target genes in the four cell types. For almost all (99%-100%) FR and RF orientations of DNA motifs, a significantly higher expression level of putative target genes was observed for EPIs that overlapped with chromatin interactions than for EPIs that did not overlap in the four cell types (Additional file 1:

Table S5 and S6). DNA motifs showing a significantly lower expression level were not observed in this analysis. HiChIP data were produced using H3K27ac antibody; thus, chromatin interactions acting as repressors would not be identified. However, for in situ Hi-C and promoter capture Hi-C data in HMEC and NPC, the significant difference in the expression level of putative target genes of EPIs was observed in a small percentage (less than 1%) of biased orientations of DNA motifs.

Recently, high resolution chromatin interaction data became publicly available [67]. The biased orientations of TFBSs of TFs were examined using Micro-C chromatin interaction data in human fibroblasts. The resolution of the data was adjusted to 200 bp. TFBSs of CTCF, RAD21, SMC3, and SMRC2 showed a significant directional bias at the pairs of chromatin interaction anchors (p-value $< 10^{-50}$). SMARCA4 is known to be involved in higher-order chromatin structures such as the topologically associating domain (TAD) [7]. DNA motifs of SMARCA4 did not exist in the databases of DNA motifs of TFs used in this analysis. SMRC2 was not identified in this analysis because it may be involved in higher-order chromatin structures that do not affect EPI and gene expression directly by its insulator function. ATF3 and ZBTB7A in Table 1 were found among the top 32 biased orientations of TFBSs of TFs at the pairs of chromatin interaction anchors with p-value $< 10^{-50}$ (Additional file 1: Table S9). Deletion of an ATF3-binding site affected the expression of a target gene, which was far apart from the site [25]. ZBTB7A is involved in the changes in chromatin accessibility and the regulation of gene expression [26]. TFBSs of YY1 and ZNF143 also indicated a weak directional bias at the chromatin interaction anchors (Additional file 1: Table S9). In this

297 analysis, biased orientations of TFBSs of TFs were investigated based on the difference
298 of distribution of the expression level of putative transcriptional target genes. YY1 and
299 ZNF143 were found to have significant directional biases, thereby suggesting that the
300 analysis in this study could find TFs that did not have a strong directional bias based on
301 the number of TFBSs at chromatin interaction anchors. The other TFs, except for CTCF,
302 RAD21, SMC3, ATF3, and ZBTB7A in Table 1, showed weak directional biases at the
303 chromatin interaction anchors (Additional file 1: Table S9).
304
305 **Correlation of expression level of gene pairs separated by TFBS**
306 To investigate whether the biased orientations of TFBSs affected gene expression,
307 the expression level of protein-coding genes, that are most closely located upstream and
308 downstream of any biased orientations of TFBSs in the human genome, was compared
309 in monocytes. The expression levels of closely located genes including divergent gene
310 pairs tend to be correlated among tissues [68, 69]. When a biased orientation of TFBSs
311 serves as an insulator (e.g. CTCF), the correlation of gene expression level would be
312 reduced. Among 96 biased (64 FR and 52 RF) orientations of TFBSs found in common
313 in monocytes, T cells, HMEC, and NPC, TFs with ≥50 genomic locations of the
314 TFBSs were selected after the elimination of ubiquitously expressed genes among 53
315 tissues and TFBSs near the genes with a coefficient of variance (CV) < 90. After this
316 filtering, 43 of 60 (72%) FR and 40 of 50 (80%) RF orientations showed significantly
317 lower correlations of expression levels of gene pairs separated by the TFBSs of a TF
318 than the correlations of all pairs of neighboring genes with intergene distance <1

319 megabase (Mb) (Mann-Whitney test, *p*-value < 0.05) (Table 3). The FR and RF
320 orientations of TFs included known TFs involved in chromatin interactions such as
321 CTCF, RAD21, SMC3, YY1, and ZNF143. The top 20 FR and RF orientations of DNA
322 motifs were selected in ascending order of the median correlation coefficient of
323 expression levels of gene pairs separated by the TFBSs of a TF. The 20 TFs included
324 known TFs involved in chromatin interactions such as CTCF, RAD21, and SMC3 (Fig.
325 4) (Mann-Whitney test, *p*-value < 0.01). Instead of using all TFBSs of a TF, when
326 TFBSs of a TF were limited to sites used in the analysis of biased orientations of DNA
327 motifs in an EPA domain, 33 of 55 (60%) FR and 36 of 47 (77%) RF orientations
328 showed significantly lower correlations of expression levels of gene pairs separated by
329 the TFBSs (Additional file 1: Table S10). With intergene distance < 20 kb, 21 of 62
330 (34%) FR and 8 of 52 (15%) RF orientations showed significantly lower correlations of
331 expression levels of gene pairs separated by the TFBSs of a TF than the correlations of
332 all pairs of neighboring genes (Additional file 1: Table S10).

333 A large number of TFBSs of a TF tend to show significantly lower correlations of
334 gene expression (median numbers of TFBSs of a TF showing significantly lower
335 correlations were 656 and 479 for FR and RF, respectively, and median numbers of
336 TFBSs of a TF not showing significantly lower correlations were 122 and 115 for FR
337 and RF, respectively. The Mann-Whitney test *p*-value = 0 and < 10^{-10} for FR and RF,
338 respectively. This suggests that if the criteria to examine the correlations of expression
339 levels of gene pairs separated by TFBSs are more stringent, and the number of TFBSs is
340 reduced, it would be more difficult to use a statistical test to compare the correlations of

expression levels of gene pairs separated by the reduced TFBSs, particularly for TFs with a small number of TFBSs.

Among 91 biased orientations of DNA motifs of TFs found in common among the four cell types, after excluding the five known TFs involved in chromatin interactions such as CTCF, RAD21, SMC3, YY1, and ZNF143, TFBSs of 88 TFs were located at ≥1 genomic regions between gene pairs not including TFBSs of the known TFs and at ≥1 genomic regions between low correlation (< 0.13) of expression level of gene pairs (Table 3). TFBSs of 52 TFs were placed at ≥10 genomic regions, though some proportion of TFBSs of the known TFs might not be associated with insulator function.

Co-location of motifs with biased orientations

To examine the association of directional DNA-binding motifs, co-location of motifs within 200 bp (approximately the minimum length of an open chromatin region of 146 bp) of each other was analyzed in monocytes, T cells, HMEC, and NPC. The same pairs of motifs upstream and downstream of genes in EPA domain were enumerated, and the pairs of motifs found in ≥100 genomic regions were listed (Table 4; Additional file 1: Table S11). A majority of the pairs of motifs overlapped with more than 1 base, but the same tendency was found in the analysis of motifs of all TFs (Additional file 1: Table S11). CTCF partially overlapped with cohesin RAD21 and SMC3 (Table 4). The top 20 pairs of FR and RF orientations of motifs overlapped or co-localized have been shown (Table 4). Unique pairs of overlapping or co-localized

motifs were 404 FR and 662 RF, consisting of 117 FR and 218 RF of unique DNA motifs in monocytes. Only 10 overlapping or co-localized motifs were found to be common among the four cell types, including pairs of CTCF, RAD21, SMC3, and YY1, thereby implying that most overlapping and co-localized motifs serve in a cell-specific manner (Additional file 1: Table S11). Paired DNA-binding motifs of TFs with biased orientations tend to overlap, and a relatively small number of motifs are co-localized upstream and downstream of genes. Overlapping motifs act as TFBSs for several TFs either simultaneously or competitively [70-77]. CTCF and cohesin function cooperatively with another TF and a protein such as KLF4 and BRD2 [78, 79]. KLF4 recruits cohesin to the *Oct4* enhancer region. BRD2 supports boundary (insulator) activity of CTCF. The analysis of 457 ChIP-seq data sets from 119 human TFs showed that secondary motifs were detected in addition to the canonical motifs of the TFs, indicating tethered binding and co-binding between multiple TFs [80]. Wang et al. [80] observed significant position and orientation preferences between many co-binding TFs. Overlapping motifs found in this study may be biologically functional rather than artifact. Though repeat sequences in the human genome were masked, more than a hundred pairs of motifs overlapped with hundreds of upstream and downstream gene regions.

Discussion

A systematic method to find DNA-binding motifs of TFs that affect EPIs and gene expression without using chromatin interaction data is presented in this study.

385 DNA motifs of TFs such as CTCF and cohesin (RAD21 and SMC3) were searched in
386 the open chromatin regions of monocytes, T cells, HMEC, and NPC. A total of 96
387 biased (64 FR and 52 RF) orientations of DNA motifs were discovered and found to be
388 common among the four cell types. Remarkably, non-oriented (i.e., without considering
389 orientation) motifs was not found to affect the expression level of putative
390 transcriptional target genes, suggesting that an EPA domain limited within the FR or RF
391 orientation of TFBSs was a more accurate prediction of EPIs. This was supported by the
392 comparisons with chromatin interaction and gene expression data. FF and RR
393 orientations of motifs were not examined in this analysis since CTCF has FR or RF
394 orientation of their TFBSs [3]. Among the 96 biased orientations of DNA motifs studied,
395 14 were reported to be associated with chromatin interactions, and many other TFs have
396 chromatin-associated functions. The enrichment analysis of biased orientations of
397 TFBSs of the 96 TFs for pairs of chromatin interaction anchors indicated a weak
398 statistical significance, except for CTCF, RAD21, SMC3, ATB3, and ZBTB7A,
399 suggesting that the TFs were difficult to find using DNA motifs of TFs alone [5].

400 Without considering the difference of orientations among cell types, 175 biased
401 orientations were found in common among the four cell types. These numbers (96 and
402 175) are the numbers of unique TFs with biased orientations of DNA motifs and not the
403 numbers of unique DNA motifs. Possible causes of the discrepancy are as follows: (1) A
404 TF or an isoform of the same TF may bind to different DNA-binding motifs, according
405 to cell types and/or in the same cell type. About 50% or more alternative splicing
406 isoforms are differently expressed among tissues, thereby indicating that most

407 alternative splicing is subject to tissue-specific regulation [81-83]; (2) The same TF has
408 several DNA-binding motifs, and in some cases, one of the motifs is almost the same as
409 the reverse complement motif of another motif of the same TF in a database; (3) I
410 previously found that a complex of TFs would bind to a slightly different DNA-binding
411 motif by combining DNA-binding motifs from the complex in *C. elegans* [84]. In other
412 reports, the difference of binding specificity of TF pairs was investigated [85].
413 Cofactors also alter the binding specificity of TFs [70-72]. These increase the variety
414 and number of (biased orientations of) DNA-binding motifs, which are potentially
415 involved in different biological functions among cell types.

416 To predict TFBSs in genomic DNA, several tools and methods were explored and
417 then protein interaction quantitation (PIQ) was used [27]. PIQ predicts TFBSs by
418 reducing noise in DNase-seq data, and the comparison of the prediction with ChIP-seq
419 data has revealed that PIQ shows a better performance than other tools and digital
420 genome footprinting (DNase-DGF). The prediction of TFBSs using PIQ in this study
421 increased the number of biased orientations of motifs, thereby reducing the number of
422 non-oriented motifs. At first, the number of predicted motifs was expected to increase
423 relative to using consensus DNA-binding sequences of TFs and open chromatin regions
424 available in the ENCODE database, which seemed to be difficult to find all TFBSs in
425 the genome. However, the number of TFBSs was almost the same or slightly decreased
426 compared with the analysis of consensus sequences, suggesting that PIQ increased
427 true-positive predictions of TFBSs and reduced false-positive predictions.

428 When forming a homodimer or heterodimer with another TF, a TF may bind to

genomic DNA with a specific orientation of its TFBSs. The approach in this study would be able to identify TFs forming a heterodimer. If the TFBS of only the mate to a pair of TFs was found in an EPA domain, the EPA domain was limited within one side of the TFBS, and transcriptional target genes were predicted using the EPA domain limited within the side. In this analysis, the mate to both heterodimers and homodimers of TFs can be tested to examine the effect on the expression level of transcriptional target genes predicted using the EPA domain limited within one side. Biased orientations of DNA-binding motifs may also be found in forward-forward or reverse-reverse orientation.

TFs that are multi-functional transcriptional regulators such as YY1 may not indicate a strong directional bias of their TFBSs. Though some proportion of TFBSs of a TF involved in chromatin interactions and insulator functions is found in a biased orientation in the genome, the other TFBSs of the TF may not form a dimer with another TF and may not be present in a specific orientation in the genome. Thus, the overall proportion of biased orientations of TFBSs of a TF may be reduced such that a statistical test cannot detect it.

The approach presented in this study was tested to search for repeat sequences that serve as insulator sites and affect the expression level of putative transcriptional target genes of TFs. A small number of repeat sequences were identified and included MIR, which act as insulators [86].

CTCF and cohesin-binding sites are frequently mutated in cancer [87]. Some biased orientations of DNA motifs involved in chromatin interactions may be associated

with diseases including cancer.

The analyses conducted here suggest that biased orientations of DNA motifs of discovered TFs affect EPIs and the expression level of putative transcriptional target genes of general TFs bound in enhancers and promoters. This gives insight into the nature of TFs and contributes to the understanding of functions and molecular mechanisms of chromatin interactions, regulation of chromatin, histone modifications, transcriptional regulation, and transcriptional cascades.

Methods

Search for biased orientations of DNA motifs

To identify transcription factor binding sites (TFBSs) from open and closed chromatin regions, TRANSFAC (2019.2), JASPAR (2018), UniPROBE (2018), high-throughput SELEX, transcription factor binding sequences of ENCODE ChIP-seq data, and HOCOMOCO version 9 and 11 were used to predict insulator sites [88-94]. TRANSFAC (2019.2) was used to analyze enhancer-promoter interactions because these data were sufficient to identify biased orientations of DNA motifs of insulator TFs with less computational time, thereby reducing the number of non-oriented DNA motifs of TFs. To reduce false-positive predictions of TFBSs in open chromatin regions, the DNA motif search was performed by using the protein interaction quantitation (PIQ) tool with 12,249 position frequency matrices converted from DNA motifs of vertebrate TFs in the above-mentioned databases and DNase-seq data from GEO and ENCODE database (GSM1024791 $CD14^+$ monocytes; GSM665812 $CD4^+$ T cell; GSM736634

Human mammary epithelial cell, HMEC; GSM878615 Neural progenitor cell, NPC; ENCFF263GMV, ENCFF886XEV Th17; GSM1024741 Treg; GSM736620 GM12878) [27]. To find open chromatin regions associated with enhancer and promoter activity, narrow peaks of ChIP-seq experimental data of H3K27ac were predicted using MACS2 callpeak (GEO: GSM773004) [95]. To predict TFBSs of TFs from closed chromatin regions, another method was employed. Position weight matrices of vertebrate transcription factor binding sequences were converted into DNA consensus sequences, which consist of an alphabet of 15 characters (the four bases A, C, G, T, the six two-fold degenerate IUPAC codes R=[AG], Y=[CT], K=[GT], M=[AC], S=[GC], W=[AT], the four three-fold degenerate IUPAC codes B=[CGT], D=[AGT], H=[ACT], V=[ACG], and the four-fold degenerate character N=[ATGC]) using a convert matrix in the RSAT website [96]. Transcription factor binding sequences were searched from genomic regions other than narrow peaks of DNase-seq data in repeat-masked hg19 genome sequences using the Match tool in TRANSFAC with a similarity score cutoff of 1 after transforming the DNA consensus sequences of TFs into position frequency matrixes for the Match search [97]. Repeat-masked hg19 genome sequences were downloaded from the UCSC genome browser (http://genome.ucsc.edu/, http://hgdownload.soe.ucsc.edu/goldenPath/hg19/bigZips/hg19.fa.masked.gz). TFs corresponding to transcription factor binding sequences were searched computationally and manually by comparing their names and gene symbols of HUGO Gene Nomenclature Committee-approved gene nomenclature and 31,848 UCSC canonical transcripts

(http://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/knownCanonical.txt.gz) because 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 the promoter or extended regions of a gene's EPA. Promoter and extended regions were defined as follows: promoter regions were those that were within a 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 [12]: 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. The limitation of extension length was <1 Mb in this study. Extended regions for an EPA were limited within the TFBSs of a putative insulator TF closest to a TSS, and transcriptional target genes were predicted using the limited EPA domain and TFBSs of a general TF in open chromatin regions. Promoter and extended regions of an EPA were limited within the forward–reverse (FR) orientation of TFBSs of a putative insulator TF. When the forward or reverse orientation of TFBSs was located in the genome several times, the most internal (i.e., closest to a TSS) forward–reverse orientation of the TFBSs was selected. The genomic positions of genes were identified using the 'knownGene.txt.gz' file in UCSC bioinformatics sites [98]. The file 'knownCanonical.txt.gz' was also utilized to select representative transcripts among various alternate forms for assigning

516 promoter and extended regions of an EPA. From the list of TFBSs and transcriptional
517 target genes, redundant DNA motifs of TFs were removed by comparing the putative
518 target genes of a TFBS and its corresponding TF; if identical, one of the DNA motifs of
519 TF was chosen. When the number of transcriptional target genes predicted based on a
520 TFBS was <5, the DNA motif of TF was omitted.

521 For gene expression data, RNA-seq reads mapped onto the human hg19 genome
522 sequence were obtained from UCSF-UBC human reference epigenome mapping project
523 RNA-seq reads with poly-A of naive $CD4^+$ T cells (GEO: GSM669617). For monocytes,
524 Blueprint RNA-seq FPKM data
525 ('C0010KB1.transcript_quantification.rsem_grape2_crg.GRCh38.20150622.results')
526 were downloaded from Blueprint DCC portal
527 (http://dcc.blueprint-epigenome.eu/#/files). RNA-seq reads with the poly-A of human
528 mammary epithelial cells (HMEC) in the Encyclopedia of DNA Elements at UCSC
529 (http://genome.ucsc.edu/encode/,
530 'wgEncodeCshlLongRnaSeqHmecCellPapAlnRep1.bam' file), H1-derived NPC (GEO:
531 GSM915326, ENCODE: ENCFF529SIO), Th17 (GSM2859479, NCBI SRA:
532 SRR6298326), Treg (GSM2859476, SRR6298323), and GM12878 (GSE78551,
533 ENCFF297QCE) were analyzed. RNA-seq reads were aligned in repeat-masked hg19
534 genome sequences using Burrows-Wheeler Aligner with default parameters [99].
535 FPKMs of the RNA-seq data were calculated by RSeQC [100]. Based on
536 log2-transformed FPKM, UCSC transcripts were arranged in a descending order of
537 expression level, and the top 4,000 transcripts were selected in each cell type.

The expression level of transcriptional target genes predicted using an EPA domain limited within the TFBSs of a TF was compared with the expression level of transcriptional target genes predicted using closed chromatin regions in promoters. For each DNA motif limiting the EPA domain, transcriptional target genes were predicted based on DNA-binding motifs of vertebrate TFs in the TRANSFAC database, and the ratio of expression level of putative transcriptional target genes of each TF was calculated between the EPA domain and closed chromatin regions in the promoters. The distribution of the ratios of all TFs was compared among forward-reverse (FR), reverse-forward (RF), and non-orientation (i.e., without considering orientation) of TFBSs limiting an EPA domain using a two-sided Mann-Whitney test, (p-value $< 10^{-7}$). Other parameters for the analysis were set as follows: the number of transcriptional target genes of a TF was $\geq$50, the number of general TFs (not insulator TFs) to predict transcriptional target genes was $\geq$50, and the number of TFBSs of a TF (insulator TF) shortening an EPA was $\geq$100.

Comparison with chromatin interaction data

For comparison of EPIs with chromatin interactions (HiChIP) in $CD4^+$ T, Th17, Treg, and GM12878 cells, ‘GSM2705049_Naive_HiChIP_H3K27ac_B2T1_allValidPairs.txt’, ‘GSM2705050_Naive_HiChIP_H3K27ac_B2T2_allValidPairs.txt’, ‘GSM2705051_Naive_HiChIP_H3K27ac_B3T1_allValidPairs.txt’, ‘GSM2705054_Th17_HiChIP_H3K27ac_B1T2_allValidPairs.txt’,

‘GSM2705055_Th17_HiChIP_H3K27ac_B2T1_allValidPairs.txt’, ‘GSM2705056_Th17_HiChIP_H3K27ac_B3T1_allValidPairs.txt’, ‘GSM2705059_Treg_HiChIP_H3K27ac_B3T1_allValidPairs.txt’, ‘GSM2705041_GM_HiChIP_H3K27ac_B1_allValidPairs.txt’, and ‘GSM2705042_GM_HiChIP_H3K27ac_B2_allValidPairs.txt’ files were downloaded from the Gene Expression Omnibus (GEO) database. The resolutions of chromatin interactions and EPIs were adjusted to 5 kb before comparison. Chromatin interactions with ≥ 6,000 and ≥ 1,000 counts for each interaction were analyzed in the comparison in $CD4^+$ T cells. According to the unique numbers of chromatin interactions in replications in Th17, Treg, and GM12878 cells, chromatin interactions with ≥1,000 (B1T2 in Th17), ≥2,000 (B2T1 and B3T1 in Th17), ≥ 6,000 (B3T1 in Treg), and ≥4,000 (B1 and B2 in GM12878) counts for each interaction were analyzed in the comparison.

EPIs were predicted on the basis of the following three types of EPA domains in monocytes: (i) EPA domain limited within the FR or RF orientation of TFBSs of a TF, (ii) EPA domain limited within the non-oriented (i.e., without considering orientation) TFBSs of a TF, and (iii) EPA domain without being limited within a TFBS. EPIs predicted using the three types of EPA domains in common were removed. EPIs predicted using EPA domain (i) were removed from EPIs predicted using EPA domain (ii). EPIs predicted using EPA domain (i) and (ii) were removed from EPIs predicted using EPA domain (iii). The resolution of HiChIP chromatin interaction data was 1-5 kb; thus, EPIs were adjusted to 5 kb before comparison. The number and ratio of EPIs

582 overlapped with chromatin interactions were compared two times between EPI (i) and
583 (iii), and EPI (i) and (ii) (binomial distribution, *p*-value < 0.025 for each test, two-sided,
584 95% confidence interval).

585 For comparison of EPIs with chromatin interactions (in situ Hi-C) in HMEC, the
586 genomic positions of chromatin interactions in the 'HMEC_4DNFI97O9SAZ.pairs.gz'
587 file in 4D nucleome data portal (http://data.4dnucleome.org) were converted into the
588 hg19 version of human genome using the liftOver tool
589 (http://hgdownload.soe.ucsc.edu/downloads.html#liftover). The resolutions of
590 chromatin interaction data and EPIs were adjusted to 3 kb before comparison. All
591 chromatin interactions were used in the comparison. For comparison of EPIs with
592 chromatin interactions (promoter capture Hi-C) in NPC, the resolutions of chromatin
593 interaction data in the 'GSE86189_npc.po.all.txt.bz2' file (GEO: GSE86189) and EPIs
594 were adjusted to 5 kb before comparison. All chromatin interactions were used in the
595 comparison.

596 Putative target genes for the comparison of EPIs and chromatin interactions were
597 selected from top 4,000 transcripts in terms of the expression level. The expression level
598 of putative target genes of EPIs overlapped with chromatin interactions was compared
599 with EPIs not overlapped with them. When the same transcriptional target gene of a TF
600 in an enhancer was predicted using both an EPI overlapping a chromatin interaction and
601 a non-overlapping EPI, the target gene was eliminated. The distribution of expression
602 level of putative target genes was compared using a two-sided Mann-Whitney test
603 (*p*-value < 0.05).

Micro-C chromatin interaction data in human fibroblasts (4D nucleome data portal ID: 4DNFICOEXGPJ) was converted into the hg19 version of the human genome using the liftOver tool and was adjusted to 200-bp resolution. The same genomic locations of paired chromatin interaction anchors were merged.

Correlation of expression level of gene pairs separated by TFBS

To examine the correlation of expression level of genes pairs separated by TFBSs of a TF, gene expression data among 53 tissues were obtained from the 'GTEx_Analysis_2016-01-15_v7_RNASeQCv1.1.8_gene_median_tpm.gct.gz' file in the GTEx database (https://gtexportal.org/home/). The correlation of log2-transformed gene expression levels was analyzed based on RefSeq protein-coding transcripts (RefSeq ID with 'NM_'). The same genomic locations of RefSeq transcripts were removed, and TFBSs overlapped with RefSeq transcripts were also eliminated. Genes closest to a TFBS and another gene were analyzed using BEDOPS closest-features [101]. The Mann-Whitney test was performed using R.

Abbreviations

TF: Transcription factors; TFBS: Transcription factor binding sites; TSS: Transcriptional start sites; EPA domain: Promoter and extended regions for enhancer-promoter association; EPI: Enhancer-promoter interaction; FF: Forward-forward; FR: Forward-reverse; RF: Reverse-forward; RR: Reverse-reverse; ChIP-seq: ChIP-sequencing, chromatin immunoprecipitation followed by massively

parallel DNA sequencing; DNase-seq: DNase I hypersensitive sites sequencing; RNA-seq: RNA sequencing; ENCODE: Encyclopedia of DNA elements;

Acknowledgements

Supercomputing resources were provided by the Human Genome Center of the Institute of Medical Science at the University of Tokyo. This work was partially supported by JSPS KAKENHI (Grant Number 16K00387) and JST CREST (Grant Number JPMJCR15G1), Japan.

Figures

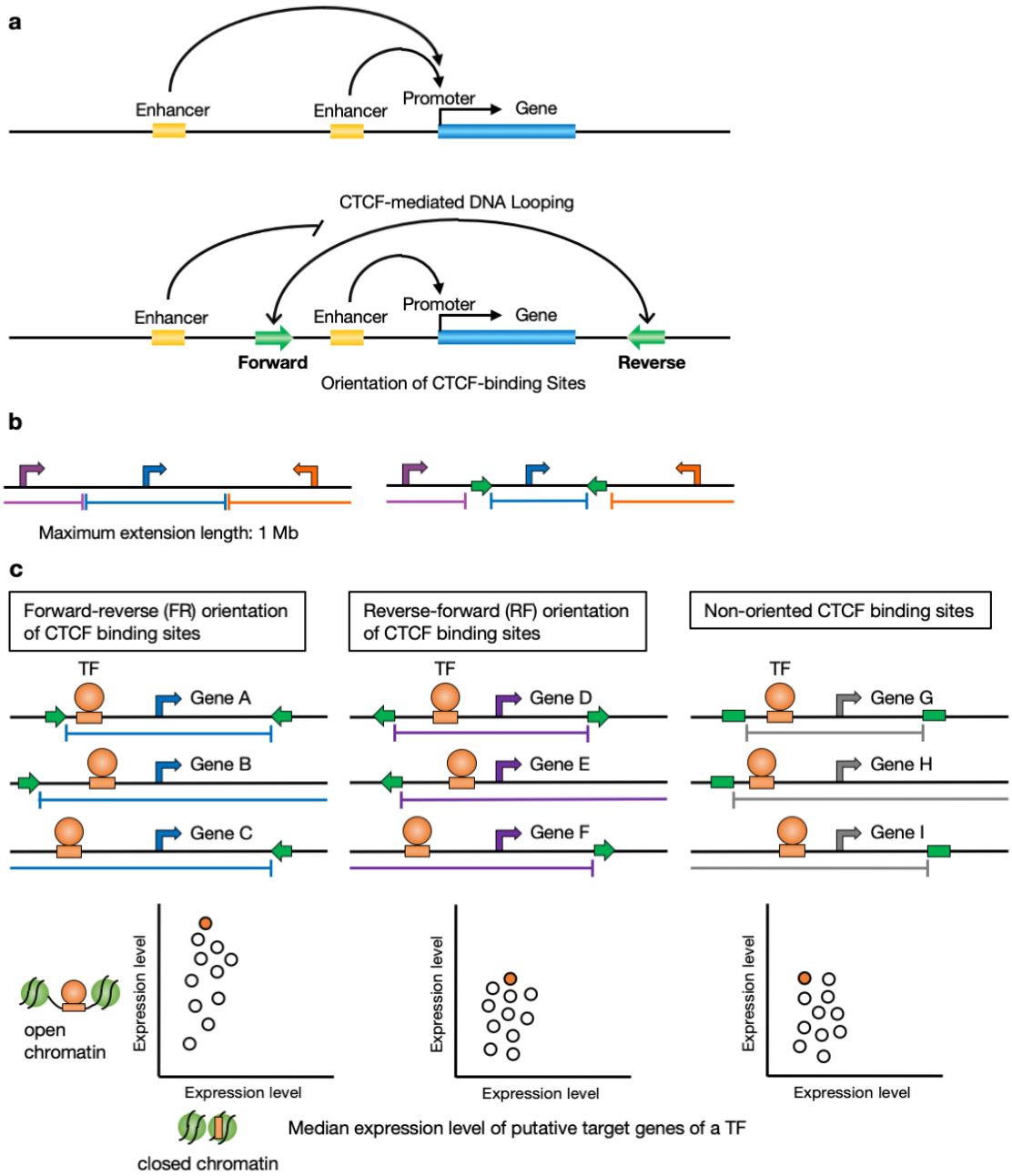
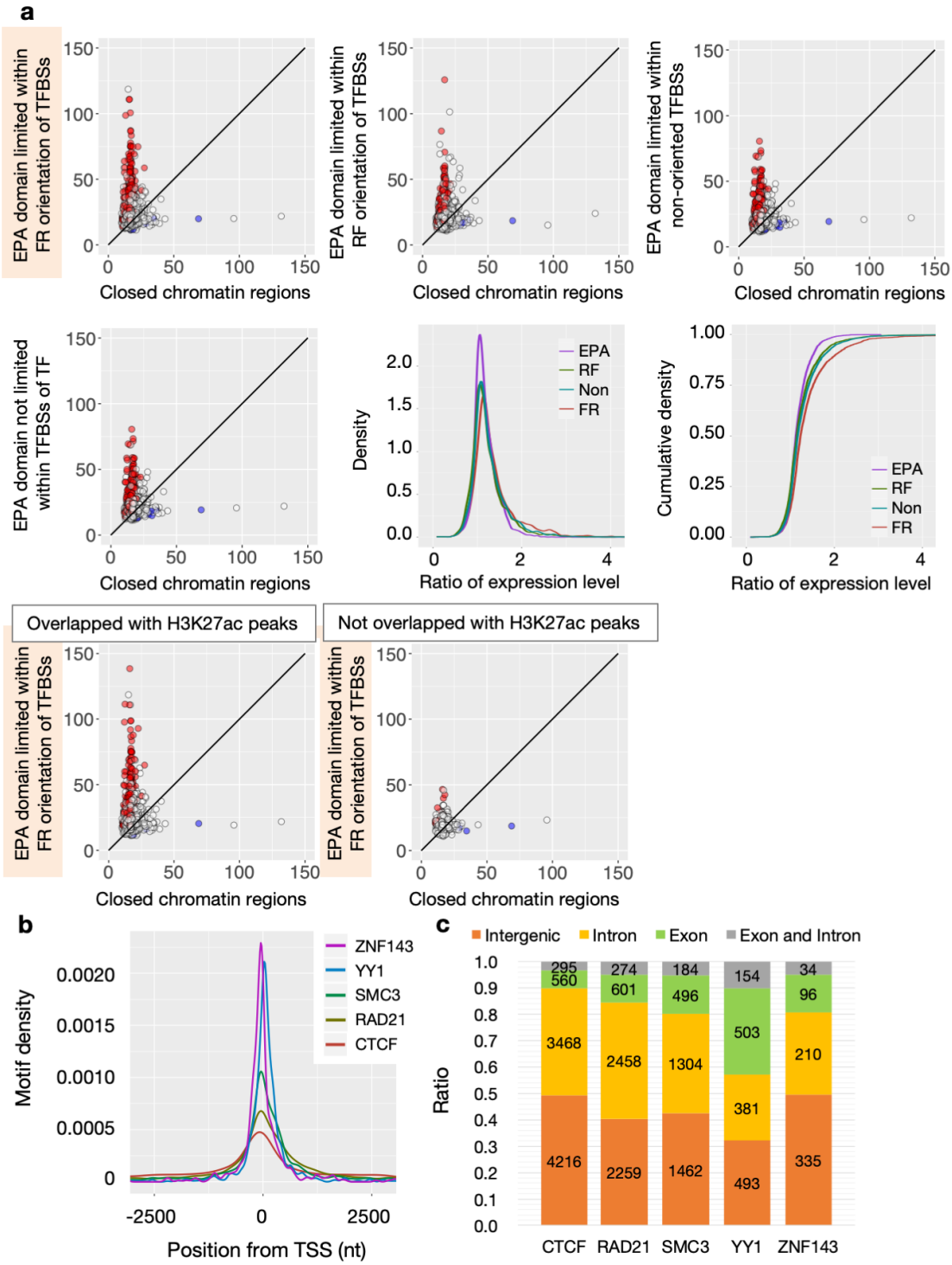


Fig. 1 Biased orientations of DNA motifs of TFs affecting enhancer-promoter interactions. **a** Forward–reverse orientation of CTCF-binding sites is frequently found in chromatin interaction anchors. CTCF can block the interaction between enhancers and promoters limiting the activity of enhancers to certain functional domains [1-4]. **b** Computationally-defined regulatory domains for enhancer-promoter association (EPA)

[12]. 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. The EPA domain was limited within the forward-reverse orientation of TFBSs of a TF in this study (e.g. CTCF in the figure). **c** Transcriptional target genes of each general TF bound in (i) distal open chromatin regions, which include enhancers, and (ii) closed chromatin regions in promoters were predicted using an EPA domain. An EPA domain was limited according to forward-reverse (FR), reverse-forward (RF) orientations, and non-oriented TFBSs of a putative insulator TF. Each dot in scatter plots represents the median expression level (FPKM) of target genes of a TF. The ratio of median expression level between (i) (y-axis) and (ii) (x-axis) was calculated to compare the distribution of the ratios of all general TFs among the three types of EPA domains.



653 **Fig. 2** Identification of TFBSs in open chromatin regions and properties of TFBSs. **a**
654 Comparison of expression level of putative transcriptional target genes. Each dot
655 represents the median expression level (FPKM) of target genes predicted based on

TFBSs of a TF in closed chromatin regions in promoters (x-axis) and an EPA domain limited within the FR orientation of CTCF-binding sites (y-axis) in T cells (upper left graph). The distribution of expression level was changed according to four criteria of EPA domains (four graphs) [11]. Red dots indicate the median expression level of target genes is significantly higher in an EPA domain than that in closed chromatin regions, and blue dots indicate the median expression level of target genes is significantly lower in an EPA domain than that in closed chromatin regions (Mann-Whitney test). The expression level of target genes predicted using an EPA domain tend to be higher than those predicted using closed chromatin regions, thereby implying that TFs serve as activators of target genes. The ratio of median expression level of target genes between an EPA domain and closed chromatin regions according to the four types of EPA domains is shown in terms of probability density and cumulative distribution (middle right two graphs). TFBSs overlapping H3K27ac ChIP-seq peaks were associated with a significant difference of expression level of target genes, and TFBSs not overlapping H3K27ac peaks were not associated with the difference (lower two graphs). This suggests that the significant difference of expression level is associated with the enhancer (and promoter) signals of H3K27ac. **b** Distribution of TFBSs of TFs near TSS. ZNF143 and YY1 tended to be located near TSS, and CTCF tended to be observed more in distant regions than those around TSS in monocytes. **c** Distribution of TFBSs of TFs in the human genome in monocytes. Most of TFBSs were found in intergenic and intron regions. YY1 is known to be at the TSS in the edge of exons of genes and is involved in transcriptional activation, repression, and initiation.

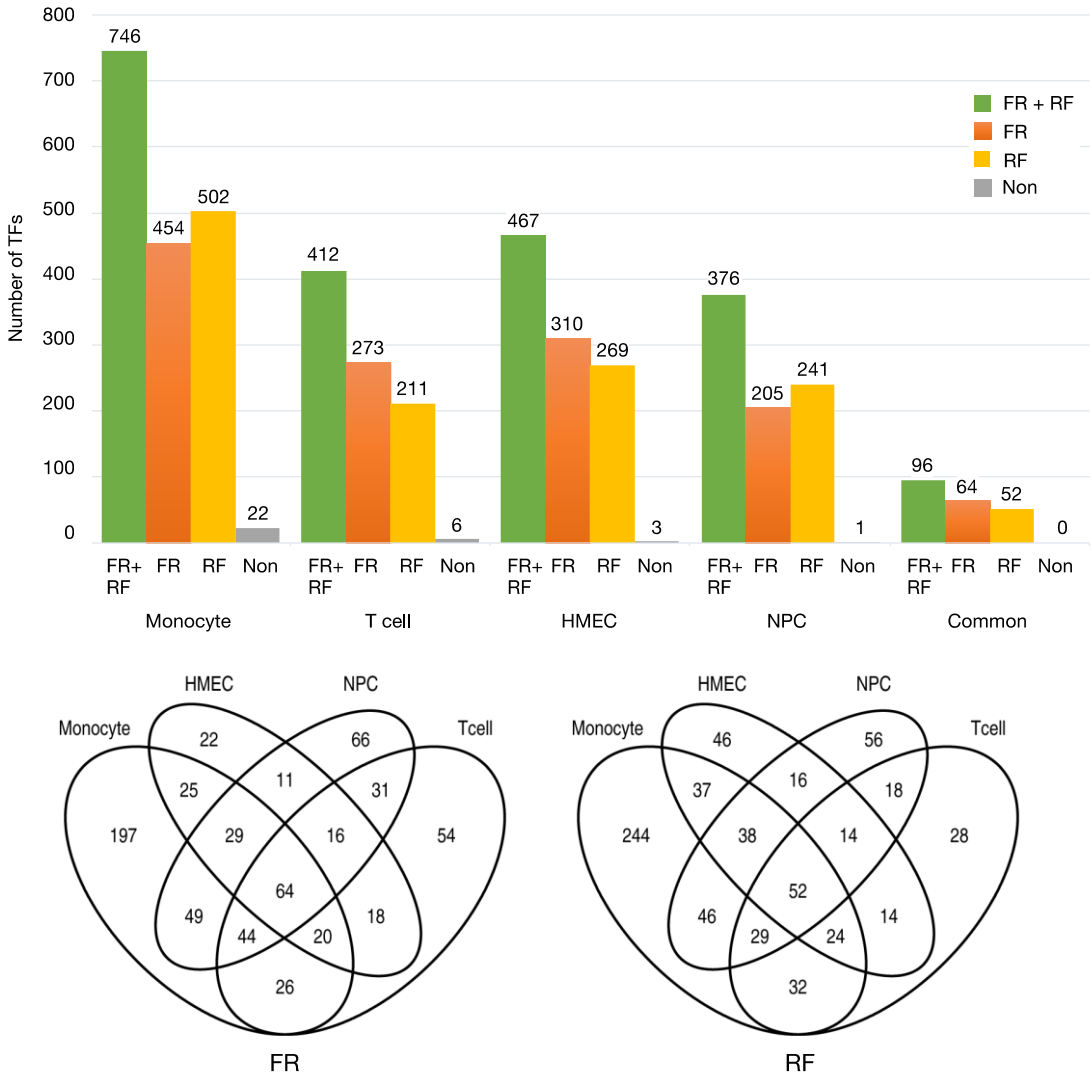


Fig. 3 Biased orientations of TFBSs of TFs affect the transcription of genes by the putative insulator function of the TFs. Ninety-six (sixty-four FR and fifty-two RF) biased orientations of TFBSs of TFs showed significant differences in the distributions of the ratios of expression level among FR, RF, and non-orientation in monocytes, T cells, HMEC, and NPC, whereas non-oriented TFBSs of TFs did not show significant differences in expression level.

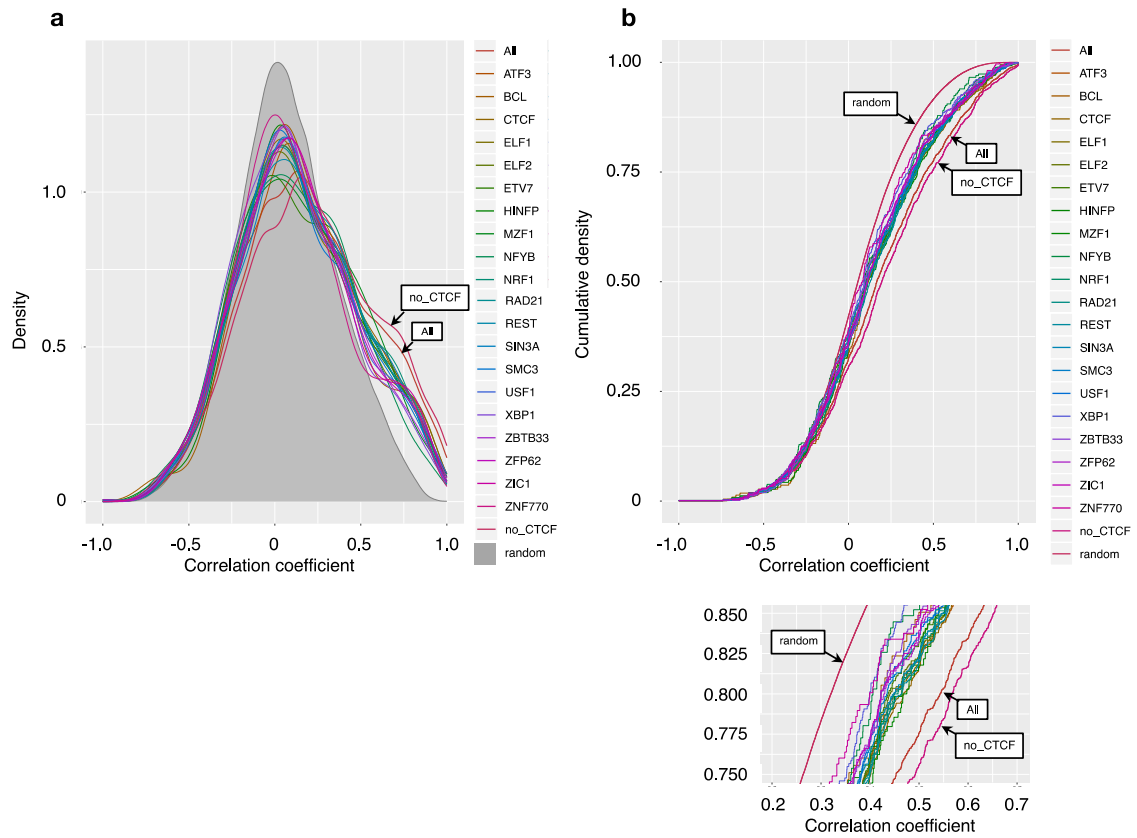


Fig. 4 Genes separated by predicted TFBSs of biased orientations of TFs are less correlated in gene expression. The top 20 TFs are selected according to ascending order of the median correlation coefficient. The correlation coefficient between neighboring gene pairs is shown in terms of probability density **a** and cumulative distribution **b**. The brown line (All) represents the correlation between all neighboring genes; gray shading in **a** and red line (random) in **b** represents correlation between randomly chosen gene pairs; and the purple line (no_CTCF) represents the correlation between gene pairs not separated by predicted CTCF-binding sites.

Tables

Table 1 Biased orientations of DNA-binding motifs of TFs in common between monocytes, T cells, HMEC, and NPC. TF: DNA-binding motifs of transcription factors.

Forward-reverse orientation

TF	P-value
AHR	0
ARNT	0
ARNTL	7.96E-08
ATF2	3.09E-14
ATF3	2.22E-15
BCL	0
BHLHE41	5.55E-14
CREB3L1	0
CTCF	0
CXXC1	0
E2F	0
E2F1	0
EGR1	0
EGR3	0
EGR4	0
ELF1	0
ELF2	5.74E-13
ELK3	3.33E-15
ELK4	6.19E-11
ERG	0
ETS	0
ETS1	0
ETV7	2.31E-08
HEY2	0
HIC1	0
HINFP	6.46E-14
INSM1	2.22E-16
KLF4	0
KLF7	9.07E-09
MAX	0

TF	P-value
MYC	0
MYCN	0
MZF1	0
NFY	1.72E-11
NFYB	0
NONO	0
NRF1	1.63E-11
OVOL2	0
PLAG1	0
PLAGL1	0
RAD21	0
REST	0
RFX1	0
SIN3A	0
SIX5	0
SMC3	0
SP2	1.04E-11
TFAP2A	0
TFAP2B	2.22E-16
TFAP2C	0
TP73	0
USF1	0
XBP1	0
YY1	0
YY2	0
ZBTB33	0
ZBTB7A	0
ZFP62	0
ZFX	0
ZIC1	0

TF	P-value
ZNF143	0
ZNF740	0
ZNF770	6.51E-08
ZNF84	0

Reverse-forward orientation

TF	P-value
AP1	0
ARNT	1.26E-11
ATF3	0
BHLHE40	8.88E-16
CREB1	0
CTCF	0
E2F	2.32E-13
E2F3	0
EFNA2	6.61E-11
EGR1	0
EGR2	0
ETS1	0
FLI1	1.05E-13
FLI1:FIGLA	0
FOS	0
GMEB2	0
HIF1A	0
KLF4	0
MITF	0
MYB	0
MYC	0
MYCN	0
MZF1	0
NFKB1	0
NFYA	2.31E-13
NR3C1	0
NRF1	0
OSR2	0
PBX3	0
PLAG1	2.78E-14

TF	P-value
RFX2	0
RFX5	0
RUNX3	0
SP1	0
SP2	0
SP4	0
SPDEF	0
SREBF2	0
STAT	0
TATA	0
TFAP2B	2.00E-15
TFAP2C	0
TFCP2L1	0
USF1	7.55E-10
YBX1	0
YY1	0
YY2	0
ZBTB33	0
ZBTB7B	0
ZNF100	6.22E-14
ZNF143	0
ZNF432	0

Table 2 Comparison of enhancer-promoter interactions (EPIs) with chromatin interactions in T cells. The upper tables show the number of biased orientations of DNA motifs, where a significantly higher ratio of EPIs that were predicted using an EPA domain (i) overlapped with HiChIP chromatin interactions than EPA domain (ii) and (iii) in T cells. The lower tables show that among 64 FR and 52 RF orientations of DNA

motifs found in common in monocytes, T cells, HMEC, and NPC, 44 (69%) FR and 28 (54%) RF were matched with the analysis of HiChIP chromatin interactions according to three types of EPA domains. TF: DNA-binding motifs of a transcription factor. Score: $-\log_{10}$ (p-value). Score 1: Comparison of EPA domain (i) (an EPA domain limited within the FR or RF orientation of TFBSs) with EPA domain (iii) (an EPA domain not limited). Score 2: Comparison of EPA domain (i) with EPA domain (ii) (an EPA domain limited within the non-oriented TFBSs).

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	226	153	101
B2T2	221	149	98
B3T1	237	160	109
DNA motifs found in common among replications			
B2T1 and B2T2	212	141	95
B2T1 and B3T1	212	143	93
B2T2 and B3T1	208	138	93
B2T1, B2T2, and B3T1	201	133	90

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	395	262	205
B2T2	398	264	204
B3T1	394	263	203
DNA motifs found in common among replications			
B2T1 and B2T2	394	262	202
B2T1 and B3T1	391	261	202
B2T2 and B3T1	392	262	200
B2T1, B2T2, and B3T1	390	261	199

FR and RF orientations of DNA motifs found to be common in the four cell types.

Forward-reverse (FR)

TF	Score 1	Score 2
AHR	38.81	13.41
ARNTL	155.78	1.96
ATF3	323.31	35.22
BCL	153.72	3.52
CREB3L1	323.31	27.96
CTCF	323.31	3.50
CXXC1	69.92	3.30
E2F	122.12	21.87
E2F1	43.22	2.85
EGR1	79.09	3.79
EGR3	85.25	34.18
ELF1	203.38	3.86
ELF2	227.62	7.83
ELK4	117.83	4.81
ERG	199.81	25.70
ETS	196.80	1.63
ETS1	308.00	3.92
HIC1	223.55	4.56
INSM1	193.85	15.87
KLF4	209.71	9.38
KLF7	229.95	11.96
MAX	323.31	8.35
MYC	84.79	24.93
MYCN	165.16	25.87
MZF1	81.83	2.29
NFY	323.31	3.42
NONO	323.31	9.24
NRF1	45.13	30.75
OVOL2	323.31	31.61
PLAG1	177.90	11.02

TF	Score 1	Score 2
RAD21	258.21	4.81
RFX1	159.78	17.05
SMC3	222.09	1.76
SP2	94.11	9.54
TFAP2A	25.86	20.26
TFAP2B	90.03	6.10
TFAP2C	169.60	8.55
TP73	198.45	7.75
USF1	194.06	3.56
XBP1	323.31	44.58
YY1	62.59	5.30
ZBTB7A	52.89	30.58
ZFX	13.33	14.55
ZNF143	323.31	35.38

Reverse-forward (RF)

TF	Score 1	Score 2
ARNT	323.31	14.12
ATF3	323.31	14.64
BHLHE40	323.31	12.42
CTCF	240.61	4.56
E2F	78.16	7.07
E2F3	65.67	8.85
EFNA2	255.01	10.95
GMEB2	323.31	7.81
KLF4	213.35	9.49
MYB	323.31	17.68
MYC	323.31	12.23
MZF1	200.49	22.73
NFYA	323.31	11.75
NR3C1	43.16	8.31
OSR2	111.77	14.65
SP2	294.86	2.59
SP4	110.94	19.63
SPDEF	185.85	14.35
SREBF2	297.27	1.62
TATA	2.16	4.02
TFAP2B	29.46	12.62
TFAP2C	112.10	51.21
YY1	323.31	9.92
YY2	47.49	2.61
ZBTB33	90.29	7.78
ZBTB7B	43.90	23.59
ZNF100	27.67	4.75
ZNF143	63.03	9.69

Table 3 Correlation of the expression level of gene pairs separated by TFBSs of a TF. Pairs of genes separated by 64 FR and 52 RF orientations of TFBSs of TFs found to be common between monocytes, T cells, HMEC, and NPC were analyzed to find significantly lower correlation of their expression level compared to the correlation of all pairs of nearby genes among 53 tissues. Score: $-\log_{10}$ (p-value) of the Mann-Whitney test. Median correlation: median correlation coefficient of expression level of nearby gene pairs separated by TFBSs. The median correlation coefficient of expression level of all pairs of nearby genes was 0.17. # TFBS: number of TFBSs of a

TF separating gene pairs with expression data. All TFBSs of a TF were used, except for TFBSs separating pairs of ubiquitously expressed genes with coefficient of variance < 90.

Forward-reverse (FR)

TF	Score	Median correlation	# TFBS
AHR	1.80	0.12	520
ARNT	1.35	0.09	141
ATF2	1.40	0.11	166
ATF3	3.61	0.11	795
BHLHE41	2.02	0.07	118
CREB3L1	2.88	0.05	127
CTCF	6.02	0.09	837
E2F	3.73	0.10	444
E2F1	3.79	0.11	868
EGR1	2.97	0.10	391
ELF1	4.86	0.10	714
ELF2	3.88	0.11	781
ELK3	1.86	0.12	547
ELK4	2.44	0.11	663
ETS	3.34	0.11	523
ETS1	4.27	0.11	687
ETV7	3.68	0.11	662
HIC1	1.65	0.11	380
HINFP	1.33	0.09	258
KLF4	2.00	0.13	718
MYC	3.86	0.11	762
MYCN	2.36	0.11	583
MZF1	1.74	0.10	261
NFYB	2.97	0.10	264
NRF1	2.21	0.11	640
PLAG1	2.93	0.10	503
PLAGL1	2.55	0.10	414
RAD21	6.08	0.09	852
REST	4.35	0.11	920
RFX1	2.38	0.08	81
SIN3A	3.31	0.11	813
SMC3	6.18	0.09	625
SP2	2.30	0.12	738
TFAP2A	4.17	0.11	767
TFAP2C	2.13	0.11	470
USF1	4.24	0.10	628
XBP1	6.65	0.07	379
ZBTB33	4.90	0.10	701
ZBTB7A	2.82	0.10	485
ZFP62	2.95	0.11	565
ZFX	2.69	0.11	581
ZIC1	4.47	0.11	829
ZNF770	3.38	0.08	271

Reverse-forward (RF)

TF	Score	Median correlation	# TFBS
AP1	3.87	0.11	744
ARNT	2.95	0.10	453
ATF3	4.34	0.10	665
BHLHE40	2.49	0.07	155
CREB1	1.79	0.11	303
CTCF	5.96	0.09	721
E2F	3.95	0.10	663
E2F3	4.05	0.09	616
EFNA2	3.51	0.10	543
EGR1	2.58	0.10	428
EGR2	1.39	0.13	747
ETS1	3.75	0.11	667
FLI1:FIGLA	5.41	0.08	428
FOS	3.51	0.11	777
GMEB2	1.36	0.09	310
HIF1A	2.34	0.11	448
KLF4	2.21	0.12	744
MITF	3.14	0.10	419
MYC	4.85	0.10	704
MYCN	2.79	0.10	507
NFKB1	1.59	0.11	342
NFYA	2.15	0.10	229
NRF1	2.01	0.12	615
OSR2	1.99	0.12	301
PBX3	2.20	0.11	253
RFX2	2.12	0.09	168
RFX5	3.03	0.05	75
RUNX3	2.01	0.11	241
SP1	2.73	0.11	725
SP2	2.15	0.12	1090
SP4	2.33	0.11	479
SPDEF	5.29	0.07	449
TFAP2B	3.42	0.11	855
TFAP2C	2.05	0.11	576
TFCP2L1	1.45	0.13	362
YBX1	1.34	0.13	160
YY1	2.23	0.01	50
ZBTB33	3.44	0.11	886
ZBTB7B	2.27	0.13	892
ZNF143	2.26	0.07	81

Table 4 Top 20 of the co-localized and overlapping biased orientations of DNA-binding motifs of TFs in monocytes. Co-localized and overlapping DNA motifs of CTCF with

another biased orientation of DNA motifs are shown in separate tables. Motif 1 and Motif 2: DNA-binding motifs of TFs. No.: the number of genomic locations with both Motif 1 and Motif 2 upstream and downstream of genes in an EPA domain (< 1 Mb distance from TSS of genes). The total number of genes was 30,523.

Forward-reverse (FR)

Motif 1	Motif 2	No.	Mean of distances bet. DNA sites (nt)	Median of distances bet. DNA sites (nt)	Mean of distances from TSS (nt)	Median of distances from TSS (nt)	Interquartile range of distance from TSS (nt)
FOS	JUN	3332	0.00	0	43202.95	18085	5524 - 48583
CTCF	**SMC3**	2876	0.01	0	34606.92	13716	3073 - 39471
FOSL2	JUN	2807	0.00	0	40389.84	17365	5424 - 45251
FOSB	FOSL2	2644	0.00	0	44302.93	17867	5633 - 47209
EGR1	EGR2	2397	0.00	0	14578.21	1251	247 - 11473
CTCF	**RAD21**	2352	0.00	0	33492.68	12913	2824 - 37672
RAD21	**SMC3**	2343	0.02	0	33572.80	14302	3856 - 37159
ATF3	FOSL2	2322	0.00	0	42784.04	17841	5366 - 48236
FOSL2	JUNB	2305	0.02	0	42884.36	17867	5706 - 46515
ETS2	ETSLIKE	2205	0.00	0	37660.66	15765	4146 - 41462
FOS	JUNB	2151	0.00	0	38403.20	16367	4849 - 44736
IRF	IRF4	1962	0.00	0	40794.41	15872	3969 - 45946
FOSL2	SMARCC1	1945	0.00	0	42429.69	17369	5094 - 46586
EGR1	WT1	1941	0.16	0	12485.23	427	93 - 5394
FOSB	JUN	1808	0.02	0	36648.74	16705	5156 - 40309
ATF3	JUN	1713	0.02	0	38101.24	17119	5294 - 43559
JUN	TFAP2A	1628	0.00	0	42262.97	18017	5089 - 47094
FOSL2	TFAP2A	1628	0.00	0	36606.38	16201	4949 - 41948
BACH2	SMARCC1	1462	0.16	0	38673.02	14778	4060 - 41262
ELF1	ETS1	1459	0.00	0	37056.72	14346	3370 - 40339

Reverse-forward (RF)

Motif 1	Motif 2	No.	Mean of distances bet. DNA sites (nt)	Median of distances bet. DNA sites (nt)	Mean of distances from TSS (nt)	Median of distances from TSS (nt)	Interquartile range of distance from TSS (nt)
EGR1	EGR3	3012	0.00	0	17005.00	709	149 - 10393
NR1I2	NR1I3	2932	0.00	0	49622.84	20548	5924 - 58436
ZIC1	ZIC3	2871	0.07	0	21511.64	3092	434 - 19128
EGR	WT1	2612	0.04	0	9949.38	396	102 - 4266
FOS	FOS:JUN	2512	0.01	0	37200.80	15466	4438 - 42373
EGR	EGR3	2293	0.17	0	11634.92	532	135 - 6044
MAFG	NFE2L1:MAFG	2272	0.03	0	34641.56	16917	5284 - 42097
KLF10	SP1	2233	0.00	0	9778.94	297	73 - 3077
EGR	EGR2	2160	0.16	0	10502.76	490	129 - 5227
KLF3	SP3	2036	0.00	0	6974.78	246	63 - 1447
KLF1	SP3	1771	0.00	0	7748.42	264	66 - 1829
KLF10	SP3	1734	0.00	0	7275.48	222	58 - 1320
KLF7	SP4	1639	0.00	0	12521.62	309	58 - 4894
KLF1	SP1	1574	0.05	0	6780.46	245	61 - 1438
AP1	JUNB	1572	0.05	0	34158.07	15451	4746 - 40315
EGR1	WT1	1543	0.31	0	12742.28	444	111 - 5553
EGR1	SP1:SP3	1393	0.08	0	7789.54	307	91 - 1886
FOS:JUN	JUNB	1386	0.14	0	33012.80	14245	3900 - 37988
SP1	SP3	1372	0.00	0	8845.31	293	77 - 1816
EHF	ELF1	1366	0.02	0	41217.39	13357	2842 - 43178

Co-localized and overlapping DNA motifs of CTCF with other DNA motifs.

FR orientation of CTCF

Motif 1	Motif 2	No.	Mean of distances bet. DNA sites (nt)	Median of distances bet. DNA sites (nt)	Mean of distances from TSS (nt)	Median of distances from TSS (nt)	Interquartile range of distance from TSS (nt)
CTCF	SMC3	2876	0.01	0	34606.92	13716	3073 - 39471
CTCF	RAD21	2352	0.00	0	33492.68	12913	2824 - 37672
CTCF	CTCFL	1151	0.00	0	30252.04	11606	2229 - 34539
CTCF	ZBTB7A	470	0.27	0	23227.96	7065	745 - 23496
CTCF	E2F	123	0.81	0	7913.36	222	61 - 1997
CTCF	WT1	120	0.00	0	3543.72	114	44 - 375
CTCF	MYC	110	2.30	0	31289.45	11948	2887 - 33972

RF orientation of CTCF

Motif 1	Motif 2	No.	Mean of distances bet. DNA sites (nt)	Median of distances bet. DNA sites (nt)	Mean of distances from TSS (nt)	Median of distances from TSS (nt)	Interquartile range of distance from TSS (nt)
CTCF	RXRA	1362	0.00	0	29981.33	10937	2100 - 31407
CTCF	CTCFL	429	0.00	0	25171.22	9834	1859 - 24741
CTCF	RAD21	337	0.00	0	38829.91	17973	6117 - 45749
CTCF	ZIC1	126	1.40	0	35480.65	10286	2316 - 29380
CTCF	ZIC3	108	0.00	0	30219.32	10904	4463 - 31037

References

1. Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, Aiden EL: **A 3D map of the**

human genome at kilobase resolution reveals principles of chromatin looping. *Cell* 2014, **159:**1665-1680.

2. Shlyueva D, Stampfel G, Stark A: **Transcriptional enhancers: from properties to genome-wide predictions.** *Nat Rev Genet* 2014, **15:**272-286.
3. Guo Y, Xu Q, Canzio D, Shou J, Li J, Gorkin DU, Jung I, Wu H, Zhai Y, Tang Y, et al: **CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function.** *Cell* 2015, **162:**900-910.
4. de Wit E, Vos ES, Holwerda SJ, Valdes-Quezada C, Verstegen MJ, Teunissen H, Splinter E, Wijchers PJ, Krijger PH, de Laat W: **CTCF Binding Polarity Determines Chromatin Looping.** *Mol Cell* 2015, **60:**676-684.
5. Bailey SD, Zhang X, Desai K, Aid M, Corradin O, Cowper-Sal Lari R, Akhtar-Zaidi B, Scacheri PC, Haibe-Kains B, Lupien M: **ZNF143 provides sequence specificity to secure chromatin interactions at gene promoters.** *Nat Commun* 2015, **2:**6186.
6. Weintraub AS, Li CH, Zamudio AV, Sigova AA, Hannett NM, Day DS, Abraham BJ, Cohen MA, Nabet B, Buckley DL, et al: **YY1 Is a Structural Regulator of Enhancer-Promoter Loops.** *Cell* 2017, **171:**1573-1588 e1528.
7. Barutcu AR, Lajoie BR, Fritz AJ, McCord RP, Nickerson JA, van Wijnen AJ, Lian JB, Stein JL, Dekker J, Stein GS, Imbalzano AN: **SMARCA4 regulates gene expression and higher-order chromatin structure in proliferating mammary epithelial cells.** *Genome Res* 2016, **26:**1188-1201.
8. Schreiber J, Libbrecht M, Bilmes J, Noble W: **Nucleotide sequence and DNaseI sensitivity are predictive of 3D chromatin architecture.** *bioRxiv* 2017.
9. Zhang H, Li F, Jia Y, Xu B, Zhang Y, Li X, Zhang Z: **Characteristic arrangement of nucleosomes is predictive of chromatin interactions at kilobase resolution.** *Nucleic Acids Res* 2017, **45:**12739-12751.
10. Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, Lander ES, Kellis M: **Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals.** *Nature* 2005, **434:**338-345.
11. Osato N: **Characteristics of functional enrichment and gene expression level of human putative transcriptional target genes.** *BMC Genomics* 2018, **19:**957.

799 12. McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM,
800 Bejerano G: **GREAT improves functional interpretation of cis-regulatory**
801 **regions.** *Nat Biotechnol* 2010, **28:**495-501.
802 13. Arnold CD, Gerlach D, Stelzer C, Boryn LM, Rath M, Stark A: **Genome-wide**
803 **quantitative enhancer activity maps identified by STARR-seq.** *Science* 2013,
804 **339:**1074-1077.
805 14. Xi H, Yu Y, Fu Y, Foley J, Halees A, Weng Z: **Analysis of overrepresented**
806 **motifs in human core promoters reveals dual regulatory roles of YY1.**
807 *Genome Res* 2007, **17:**798-806.
808 15. Shrivastava A, Calame K: **An analysis of genes regulated by the**
809 **multi-functional transcriptional regulator Yin Yang-1.** *Nucleic Acids Res*
810 1994, **22:**5151-5155.
811 16. Di Giammartino DC, Kloetgen A, Polyzos A, Liu Y, Kim D, Murphy D,
812 Abuhashem A, Cavaliere P, Aronson B, Shah V, et al: **KLF4 is involved in the**
813 **organization and regulation of pluripotency-associated three-dimensional**
814 **enhancer networks.** *Nat Cell Biol* 2019, **21:**1179-1190.
815 17. Li F, Yuan Q, Di W, Xia X, Liu Z, Mao N, Li L, Li C, He J, Li Y, et al: **ERG**
816 **orchestrates chromatin interactions to drive prostate cell fate**
817 **reprogramming.** *J Clin Invest* 2020.
818 18. Zhang Z, Chng KR, Lingadahalli S, Chen Z, Liu MH, Do HH, Cai S, Rinaldi N,
819 Poh HM, Li G, et al: **An AR-ERG transcriptional signature defined by**
820 **long-range chromatin interactomes in prostate cancer cells.** *Genome Res*
821 2019, **29:**223-235.
822 19. Majumder P, Gomez JA, Chadwick BP, Boss JM: **The insulator factor CTCF**
823 **controls MHC class II gene expression and is required for the formation of**
824 **long-distance chromatin interactions.** *J Exp Med* 2008, **205:**785-798.
825 20. Quigley IK, Kintner C: **Rfx2 Stabilizes Foxj1 Binding at Chromatin Loops to**
826 **Enable Multiciliated Cell Gene Expression.** *PLoS Genet* 2017, **13:**e1006538.
827 21. Mimura I, Nangaku M, Kanki Y, Tsutsumi S, Inoue T, Kohro T, Yamamoto S,
828 Fujita T, Shimamura T, Suehiro J, et al: **Dynamic change of chromatin**
829 **conformation in response to hypoxia enhances the expression of GLUT3**
830 **(SLC2A3) by cooperative interaction of hypoxia-inducible factor 1 and**
831 **KDM3A.** *Mol Cell Biol* 2012, **32:**3018-3032.

22. Deshane J, Kim J, Bolisetty S, Hock TD, Hill-Kapturczak N, Agarwal A: **Sp1 regulates chromatin looping between an intronic enhancer and distal promoter of the human heme oxygenase-1 gene in renal cells.** *J Biol Chem* 2010, **285:**16476-16486.
23. Park JH, Choi Y, Song MJ, Park K, Lee JJ, Kim HP: **Dynamic Long-Range Chromatin Interaction Controls Expression of IL-21 in CD4+ T Cells.** *J Immunol* 2016, **196:**4378-4389.
24. D'Ippolito AM, McDowell IC, Barrera A, Hong LK, Leichter SM, Bartelt LC, Vockley CM, Majoros WH, Safi A, Song L, et al: **Pre-established Chromatin Interactions Mediate the Genomic Response to Glucocorticoids.** *Cell Syst* 2018, **7:**146-160 e147.
25. Kim DE, Procopio MG, Ghosh S, Jo SH, Goruppi S, Magliozzi F, Bordignon P, Neel V, Angelino P, Dotto GP: **Convergent roles of ATF3 and CSL in chromatin control of cancer-associated fibroblast activation.** *J Exp Med* 2017, **214:**2349-2368.
26. Ramos Pittol JM, Oruba A, Mittler G, Saccani S, van Essen D: **Zbtb7a is a transducer for the control of promoter accessibility by NF-kappa B and multiple other transcription factors.** *PLoS Biol* 2018, **16:**e2004526.
27. Sherwood RI, Hashimoto T, O'Donnell CW, Lewis S, Barkal AA, van Hoff JP, Karun V, Jaakkola T, Gifford DK: **Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape.** *Nat Biotechnol* 2014, **32:**171-178.
28. Rothstein M, Simoes-Costa M: **Heterodimerization of TFAP2 pioneer factors drives epigenomic remodeling during neural crest specification.** *Genome Res* 2020, **30:**35-48.
29. Knoepfler PS, Zhang XY, Cheng PF, Gafken PR, McMahon SB, Eisenman RN: **Myc influences global chromatin structure.** *EMBO J* 2006, **25:**2723-2734.
30. Sabo A, Amati B: **Genome recognition by MYC.** *Cold Spring Harb Perspect Med* 2014, **4**.
31. Yang H, Schramek D, Adam RC, Keyes BE, Wang P, Zheng D, Fuchs E: **ETS family transcriptional regulators drive chromatin dynamics and malignancy in squamous cell carcinomas.** *Elife* 2015, **4:**e10870.
32. Li L, Bhatia R: **The controversial role of Sirtuins in tumorigenesis - SIRT7**

865 **joins the debate.** *Cell Res* 2013, **23:**10-12.
866 33. Lee JW, Kim DM, Jang JW, Park TG, Song SH, Lee YS, Chi XZ, Park IY, Hyun
867 JW, Ito Y, Bae SC: **RUNX3 regulates cell cycle-dependent chromatin**
868 **dynamics by functioning as a pioneer factor of the restriction-point.** *Nat*
869 *Commun* 2019, **10:**1897.
870 34. Huang S, Li X, Yusufzai TM, Qiu Y, Felsenfeld G: **USF1 recruits histone**
871 **modification complexes and is critical for maintenance of a chromatin**
872 **barrier.** *Mol Cell Biol* 2007, **27:**7991-8002.
873 35. Li Y, Schulz VP, Deng C, Li G, Shen Y, Tusi BK, Ma G, Stees J, Qiu Y, Steiner
874 LA, et al: **Setd1a and NURF mediate chromatin dynamics and gene**
875 **regulation during erythroid lineage commitment and differentiation.**
876 *Nucleic Acids Res* 2016, **44:**7173-7188.
877 36. Swanson HI: **DNA binding and protein interactions of the AHR/ARNT**
878 **heterodimer that facilitate gene activation.** *Chem Biol Interact* 2002,
879 **141:**63-76.
880 37. Volkel S, Stielow B, Finkernagel F, Stiewe T, Nist A, Suske G: **Zinc finger**
881 **independent genome-wide binding of Sp2 potentiates recruitment of**
882 **histone-fold protein Nf-y distinguishing it from Sp1 and Sp3.** *PLoS Genet*
883 2015, **11:**e1005102.
884 38. Frob F, Wegner M: **The role of chromatin remodeling complexes in Schwann**
885 **cell development.** *Glia* 2020, **68:**1596-1603.
886 39. Wust HM, Wegener A, Frob F, Hartwig AC, Wegwitz F, Kari V, Schimmel M,
887 Tamm ER, Johnsen SA, Wegner M, Sock E: **Egr2-guided histone H2B**
888 **monoubiquitination is required for peripheral nervous system myelination.**
889 *Nucleic Acids Res* 2020.
890 40. Mager GM, Ward RM, Srinivasan R, Jang SW, Wrabetz L, Svaren J: **Active**
891 **gene repression by the Egr2.NAB complex during peripheral nerve**
892 **myelination.** *J Biol Chem* 2008, **283:**18187-18197.
893 41. Blattler A, Yao L, Wang Y, Ye Z, Jin VX, Farnham PJ: **ZBTB33 binds**
894 **unmethylated regions of the genome associated with actively expressed**
895 **genes.** *Epigenetics Chromatin* 2013, **6:**13.
896 42. Oldfield AJ, Yang P, Conway AE, Cinghu S, Freudenberg JM, Yellaboina S,
897 Jothi R: **Histone-fold domain protein NF-Y promotes chromatin accessibility**

898 **for cell type-specific master transcription factors.** *Mol Cell* 2014,
899 **55:**708-722.
900 43. Nardini M, Gnesutta N, Donati G, Gatta R, Forni C, Fossati A, Vonrhein C,
901 Moras D, Romier C, Bolognesi M, Mantovani R: **Sequence-specific**
902 **transcription factor NF-Y displays histone-like DNA binding and H2B-like**
903 **ubiquitination.** *Cell* 2013, **152:**132-143.
904 44. Laurette P, Strub T, Koludrovic D, Keime C, Le Gras S, Seberg H, Van Otterloo
905 E, Imrichova H, Siddaway R, Aerts S, et al: **Transcription factor MITF and**
906 **remodeller BRG1 define chromatin organisation at regulatory elements in**
907 **melanoma cells.** *Elife* 2015, **4**.
908 45. Levy C, Fisher DE: **Dual roles of lineage restricted transcription factors: the**
909 **case of MITF in melanocytes.** *Transcription* 2011, **2:**19-22.
910 46. Ma C, Karwacki-Neisius V, Tang H, Li W, Shi Z, Hu H, Xu W, Wang Z, Kong L,
911 Lv R, et al: **Nono, a Bivalent Domain Factor, Regulates Erk Signaling and**
912 **Mouse Embryonic Stem Cell Pluripotency.** *Cell Rep* 2016, **17:**997-1007.
913 47. Riggi N, Knoechel B, Gillespie SM, Rheinbay E, Boulay G, Suva ML, Rossetti
914 NE, Boonseng WE, Oksuz O, Cook EB, et al: **EWS-FLI1 utilizes divergent**
915 **chromatin remodeling mechanisms to directly activate or repress enhancer**
916 **elements in Ewing sarcoma.** *Cancer Cell* 2014, **26:**668-681.
917 48. Blais A, Dynlacht BD: **E2F-associated chromatin modifiers and cell cycle**
918 **control.** *Curr Opin Cell Biol* 2007, **19:**658-662.
919 49. Taubert S, Gorrini C, Frank SR, Parisi T, Fuchs M, Chan HM, Livingston DM,
920 Amati B: **E2F-dependent histone acetylation and recruitment of the Tip60**
921 **acetyltransferase complex to chromatin in late G1.** *Mol Cell Biol* 2004,
922 **24:**4546-4556.
923 50. Subtil-Rodriguez A, Vazquez-Chavez E, Ceballos-Chavez M,
924 Rodriguez-Paredes M, Martin-Subero JI, Esteller M, Reyes JC: **The chromatin**
925 **remodeller CHD8 is required for E2F-dependent transcription activation of**
926 **S-phase genes.** *Nucleic Acids Res* 2014, **42:**2185-2196.
927 51. Choi J, Jeon S, Choi S, Park K, Seong RH: **The SWI/SNF chromatin**
928 **remodeling complex regulates germinal center formation by repressing**
929 **Blimp-1 expression.** *Proc Natl Acad Sci U S A* 2015, **112:**E718-727.
930 52. Smith E, Shilatifard A: **The chromatin signaling pathway: diverse**

931 **mechanisms of recruitment of histone-modifying enzymes and varied**
932 **biological outcomes.** *Mol Cell* 2010, **40:**689-701.
933 53. Magri L, Swiss VA, Jablonska B, Lei L, Pedre X, Walsh M, Zhang W, Gallo V,
934 Canoll P, Casaccia P: **E2F1 coregulates cell cycle genes and chromatin**
935 **components during the transition of oligodendrocyte progenitors from**
936 **proliferation to differentiation.** *J Neurosci* 2014, **34:**1481-1493.
937 54. Sun Z, Xu X, He J, Murray A, Sun MA, Wei X, Wang X, McCoig E, Xie E,
938 Jiang X, et al: **EGR1 recruits TET1 to shape the brain methylome during**
939 **development and upon neuronal activity.** *Nat Commun* 2019, **10:**3892.
940 55. Ooi L, Belyaev ND, Miyake K, Wood IC, Buckley NJ: **BRG1 chromatin**
941 **remodeling activity is required for efficient chromatin binding by repressor**
942 **element 1-silencing transcription factor (REST) and facilitates**
943 **REST-mediated repression.** *J Biol Chem* 2006, **281:**38974-38980.
944 56. Cunliffe VT: **Eloquent silence: developmental functions of Class I histone**
945 **deacetylases.** *Curr Opin Genet Dev* 2008, **18:**404-410.
946 57. Xu Y, Sengupta PK, Seto E, Smith BD: **Regulatory factor for X-box family**
947 **proteins differentially interact with histone deacetylases to repress collagen**
948 **alpha2(I) gene (COL1A2) expression.** *J Biol Chem* 2006, **281:**9260-9270.
949 58. Saunders A, Huang X, Fidalgo M, Reimer MH, Jr., Faiola F, Ding J,
950 Sanchez-Priego C, Guallar D, Saenz C, Li D, Wang J: **The SIN3A/HDAC**
951 **Corepressor Complex Functionally Cooperates with NANOG to Promote**
952 **Pluripotency.** *Cell Rep* 2017, **18:**1713-1726.
953 59. Salas-Armenteros I, Perez-Calero C, Bayona-Feliu A, Tumini E, Luna R,
954 Aguilera A: **Human THO-Sin3A interaction reveals new mechanisms to**
955 **prevent R-loops that cause genome instability.** *EMBO J* 2017, **36:**3532-3547.
956 60. Adams GE, Chandru A, Cowley SM: **Co-repressor, co-activator and general**
957 **transcription factor: the many faces of the Sin3 histone deacetylase (HDAC)**
958 **complex.** *Biochem J* 2018, **475:**3921-3932.
959 61. Biddie SC, John S, Sabo PJ, Thurman RE, Johnson TA, Schiltz RL, Miranda TB,
960 Sung MH, Trump S, Lightman SL, et al: **Transcription factor AP1 potentiates**
961 **chromatin accessibility and glucocorticoid receptor binding.** *Mol Cell* 2011,
962 **43:**145-155.
963 62. Orlova A, Wagner C, de Araujo ED, Bajusz D, Neubauer HA, Herling M,

964 Gunning PT, Keseru GM, Moriggl R: **Direct Targeting Options for STAT3**
965 **and STAT5 in Cancer.** *Cancers (Basel)* 2019, **11**.
966 63. Mumbach MR, Satpathy AT, Boyle EA, Dai C, Gowen BG, Cho SW, Nguyen
967 ML, Rubin AJ, Granja JM, Kazane KR, et al: **Enhancer connectome in**
968 **primary human cells identifies target genes of disease-associated DNA**
969 **elements.** *Nat Genet* 2017, **49:**1602-1612.
970 64. Yardimci GG, Ozadam H, Sauria MEG, Ursu O, Yan KK, Yang T, Chakraborty
971 A, Kaul A, Lajoie BR, Song F, et al: **Measuring the reproducibility and**
972 **quality of Hi-C data.** *Genome Biol* 2019, **20:**57.
973 65. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T,
974 Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, et al: **Comprehensive**
975 **mapping of long-range interactions reveals folding principles of the human**
976 **genome.** *Science* 2009, **326:**289-293.
977 66. Jung I, Schmitt A, Diao Y, Lee AJ, Liu T, Yang D, Tan C, Eom J, Chan M, Chee
978 S, et al: **A compendium of promoter-centered long-range chromatin**
979 **interactions in the human genome.** *Nat Genet* 2019, **51:**1442-1449.
980 67. Krietenstein N, Abraham S, Venev SV, Abdennur N, Gibcus J, Hsieh TS, Parsi
981 KM, Yang L, Maehr R, Mirny LA, et al: **Ultrastructural Details of**
982 **Mammalian Chromosome Architecture.** *Mol Cell* 2020, **78:**554-565 e557.
983 68. Purmann A, Toedling J, Schueler M, Carninci P, Lehrach H, Hayashizaki Y,
984 Huber W, Sperling S: **Genomic organization of transcriptomes in mammals:**
985 **Coregulation and cofunctionality.** *Genomics* 2007, **89:**580-587.
986 69. Xie X, Mikkelsen TS, Gnirke A, Lindblad-Toh K, Kellis M, Lander ES:
987 **Systematic discovery of regulatory motifs in conserved regions of the**
988 **human genome, including thousands of CTCF insulator sites.** *Proc Natl*
989 *Acad Sci U S A* 2007, **104:**7145-7150.
990 70. Ansari AZ, Peterson-Kaufman KJ: **A partner evokes latent differences**
991 **between Hox proteins.** *Cell* 2011, **147:**1220-1221.
992 71. Slattery M, Riley T, Liu P, Abe N, Gomez-Alcala P, Dror I, Zhou T, Rohs R,
993 Honig B, Bussemaker HJ, Mann RS: **Cofactor binding evokes latent**
994 **differences in DNA binding specificity between Hox proteins.** *Cell* 2011,
995 **147:**1270-1282.
996 72. Merabet S, Mann RS: **To Be Specific or Not: The Critical Relationship**

997 **Between Hox And TALE Proteins.** *Trends Genet* 2016, **32:**334-347.
998 73. Yoon SO, Chikaraishi DM: **Tissue-specific transcription of the rat tyrosine**
999 **hydroxylase gene requires synergy between an AP-1 motif and an**
1000 **overlapping E box-containing dyad.** *Neuron* 1992, **9:**55-67.
1001 74. Ackerman SL, Minden AG, Williams GT, Bobonis C, Yeung CY: **Functional**
1002 **significance of an overlapping consensus binding motif for Sp1 and Zif268**
1003 **in the murine adenosine deaminase gene promoter.** *Proc Natl Acad Sci U S A*
1004 1991, **88:**7523-7527.
1005 75. Chatterjee R, Zhao J, He X, Shlyakhtenko A, Mann I, Waterfall JJ, Meltzer P,
1006 Sathyanarayana BK, FitzGerald PC, Vinson C: **Overlapping ETS and CRE**
1007 **Motifs ((G/C)CGGAAGTGACGTCA) preferentially bound by GABPalpha**
1008 **and CREB proteins.** *G3 (Bethesda)* 2012, **2:**1243-1256.
1009 76. He X, Syed KS, Tillo D, Mann I, Weirauch MT, Vinson C: **GABPalpha**
1010 **Binding to Overlapping ETS and CRE DNA Motifs Is Enhanced by**
1011 **CREB1: Custom DNA Microarrays.** *G3 (Bethesda)* 2015, **5:**1909-1918.
1012 77. Kin K, Chen X, Gonzalez-Garay M, Fakhouri WD: **The effect of non-coding**
1013 **DNA variations on P53 and cMYC competitive inhibition at cis-overlapping**
1014 **motifs.** *Hum Mol Genet* 2016, **25:**1517-1527.
1015 78. Wei Z, Gao F, Kim S, Yang H, Lyu J, An W, Wang K, Lu W: **Klf4 organizes**
1016 **long-range chromosomal interactions with the oct4 locus in reprogramming**
1017 **and pluripotency.** *Cell Stem Cell* 2013, **13:**36-47.
1018 79. Hsu SC, Gilgenast TG, Bartman CR, Edwards CR, Stonestrom AJ, Huang P,
1019 Emerson DJ, Evans P, Werner MT, Keller CA, et al: **The BET Protein BRD2**
1020 **Cooperates with CTCF to Enforce Transcriptional and Architectural**
1021 **Boundaries.** *Mol Cell* 2017, **66:**102-116 e107.
1022 80. Wang J, Zhuang J, Iyer S, Lin X, Whitfield TW, Greven MC, Pierce BG, Dong
1023 X, Kundaje A, Cheng Y, et al: **Sequence features and chromatin structure**
1024 **around the genomic regions bound by 119 human transcription factors.**
1025 *Genome Res* 2012, **22:**1798-1812.
1026 81. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF,
1027 Schroth GP, Burge CB: **Alternative isoform regulation in human tissue**
1028 **transcriptomes.** *Nature* 2008, **456:**470-476.
1029 82. Chen M, Manley JL: **Mechanisms of alternative splicing regulation: insights**

1030 **from molecular and genomics approaches.** *Nat Rev Mol Cell Biol* 2009,
1031 **10:**741-754.
1032 83. Das D, Clark TA, Schweitzer A, Yamamoto M, Marr H, Arribere J, Minovitsky S,
1033 Poliakov A, Dubchak I, Blume JE, Conboy JG: **A correlation with exon**
1034 **expression approach to identify cis-regulatory elements for tissue-specific**
1035 **alternative splicing.** *Nucleic Acids Res* 2007, **35:**4845-4857.
1036 84. Tabuchi TM, Deplancke B, Osato N, Zhu LJ, Barrasa MI, Harrison MM, Horvitz
1037 HR, Walhout AJ, Hagstrom KA: **Chromosome-biased binding and gene**
1038 **regulation by the Caenorhabditis elegans DRM complex.** *PLoS Genet* 2011,
1039 **7:**e1002074.
1040 85. Jolma A, Yin Y, Nitta KR, Dave K, Popov A, Taipale M, Enge M, Kivioja T,
1041 Morgunova E, Taipale J: **DNA-dependent formation of transcription factor**
1042 **pairs alters their binding specificity.** *Nature* 2015, **527:**384-388.
1043 86. Wang J, Vicente-Garcia C, Seruggia D, Molto E, Fernandez-Minan A, Neto A,
1044 Lee E, Gomez-Skarmeta JL, Montoliu L, Lunyak VV, Jordan IK: **MIR**
1045 **retrotransposon sequences provide insulators to the human genome.** *Proc*
1046 *Natl Acad Sci U S A* 2015, **112:**E4428-4437.
1047 87. Katainen R, Dave K, Pitkanen E, Palin K, Kivioja T, Valimaki N, Gylfe AE,
1048 Ristolainen H, Hanninen UA, Cajuso T, et al: **CTCF/cohesin-binding sites are**
1049 **frequently mutated in cancer.** *Nat Genet* 2015, **47:**818-821.
1050 88. Wingender E, Dietze P, Karas H, Knuppel R: **TRANSFAC: a database on**
1051 **transcription factors and their DNA binding sites.** *Nucleic Acids Res* 1996,
1052 **24:**238-241.
1053 89. Portales-Casamar E, Thongjuea S, Kwon AT, Arenillas D, Zhao X, Valen E,
1054 Yusuf D, Lenhard B, Wasserman WW, Sandelin A: **JASPAR 2010: the greatly**
1055 **expanded open-access database of transcription factor binding profiles.**
1056 *Nucleic Acids Res* 2010, **38:**D105-110.
1057 90. Newburger DE, Bulyk ML: **UniPROBE: an online database of protein**
1058 **binding microarray data on protein-DNA interactions.** *Nucleic Acids Res*
1059 2009, **37:**D77-82.
1060 91. Zhao Y, Stormo GD: **Quantitative analysis demonstrates most transcription**
1061 **factors require only simple models of specificity.** *Nat Biotechnol* 2011,
1062 **29:**480-483.

1063 92. Jolma A, Yan J, Whitington T, Toivonen J, Nitta KR, Rastas P, Morgunova E,
1064 Enge M, Taipale M, Wei G, et al: **DNA-binding specificities of human**
1065 **transcription factors.** *Cell* 2013, **152:**327-339.
1066 93. Kheradpour P, Kellis M: **Systematic discovery and characterization of**
1067 **regulatory motifs in ENCODE TF binding experiments.** *Nucleic Acids Res*
1068 2014, **42:**2976-2987.
1069 94. Kulakovskiy IV, Vorontsov IE, Yevshin IS, Sharipov RN, Fedorova AD,
1070 Rumynskiy EI, Medvedeva YA, Magana-Mora A, Bajic VB, Papatsenko DA, et
1071 al: **HOCOMOCO: towards a complete collection of transcription factor**
1072 **binding models for human and mouse via large-scale ChIP-Seq analysis.**
1073 *Nucleic Acids Res* 2018, **46:**D252-D259.
1074 95. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum
1075 C, Myers RM, Brown M, Li W, Liu XS: **Model-based analysis of ChIP-Seq**
1076 **(MACS).** *Genome Biol* 2008, **9:**R137.
1077 96. van Helden J: **Regulatory sequence analysis tools.** *Nucleic Acids Res* 2003,
1078 **31:**3593-3596.
1079 97. Kel AE, Gossling E, Reuter I, Cheremushkin E, Kel-Margoulis OV, Wingender
1080 E: **MATCH: A tool for searching transcription factor binding sites in DNA**
1081 **sequences.** *Nucleic Acids Res* 2003, **31:**3576-3579.
1082 98. Karolchik D, Barber GP, Casper J, Clawson H, Cline MS, Diekhans M, Dreszer
1083 TR, Fujita PA, Guruvadoo L, Haeussler M, et al: **The UCSC Genome Browser**
1084 **database: 2014 update.** *Nucleic Acids Res* 2014, **42:**D764-770.
1085 99. Li H, Durbin R: **Fast and accurate short read alignment with**
1086 **Burrows-Wheeler transform.** *Bioinformatics* 2009, **25:**1754-1760.
1087 100. Wang L, Wang S, Li W: **RSeQC: quality control of RNA-seq experiments.**
1088 *Bioinformatics* 2012, **28:**2184-2185.
1089 101. Neph S, Kuehn MS, Reynolds AP, Haugen E, Thurman RE, Johnson AK, Rynes
1090 E, Maurano MT, Vierstra J, Thomas S, et al: **BEDOPS: high-performance**
1091 **genomic feature operations.** *Bioinformatics* 2012, **28:**1919-1920.

1092