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. 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. The expression levels of the transcriptional target genes predicted based on the EPA were compared with those predicted from only promoters.

Total 287 biased orientation of DNA motifs (159 FR and 153 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 ~85% 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

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

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

76 DNA binding motif sequence of YY1 does not seem to be enriched at both

77 chromatin interaction anchors (Z-score < 2), whereas DNA binding motif sequence of
78 ZNF143 is significantly enriched (Z-score > 7; Bailey et al. 2015 Figure 2a). In the
79 analysis of YY1, to identify a protein factor that might contribute to EPI, (Ji et al. 2015)
80 performed chromatin immune precipitation with mass spectrometry (ChIP-MS), using
81 antibodies directed toward histones with modifications characteristic of enhancer and
82 promoter chromatin (H3K27ac and H3K4me3, respectively). Of 26 transcription factors
83 that occupy both enhancers and promoters, four are essential based on a CRISPR
84 cell-essentiality screen and two (CTCF, YY1) are expressed in >90% of tissues
85 examined (Weintraub et al. 2017). These analyses started from the analysis of histone
86 modifications of enhancer and promote marks rather than chromatin interactions. Other
87 protein factors associated with chromatin interactions may be found from other
88 researches.

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

94 DNA binding proteins involved in chromatin interactions are supposed to affect
95 the transcription of genes in the loops formed by chromatin interactions. In my previous
96 analysis, the expression level of human putative transcriptional target genes was
97 affected, according to the criteria of enhancer-promoter association (EPA) (Fig. 1B;
98 (Osato 2018)). EPI were predicted based on EPA shortened at the genomic locations of

99 FR orientation of CTCF binding sites, and transcriptional target genes of each TF bound
100 in enhancers and promoters were predicted based on the EPI. The EPA affected the
101 expression levels of putative transcriptional target genes the most among three types of
102 EPAs, compared with the expression levels of transcriptional target genes predicted
103 from only promoters (Fig. 2). The expression levels tended to be increased in
104 monocytes and $CD4^+$ T cells, implying that enhancers activated the transcription of
105 genes, and decreased in ES and iPS cells, implying that enhancers repressed the
106 transcription of genes. These analyses suggested that enhancers affected the
107 transcription of genes significantly, when EPI were predicted properly. Other DNA
108 binding proteins involved in chromatin interactions, as well as CTCF, may locate at
109 chromatin interaction anchors with a pair of biased orientation of DNA binding motif
110 sequences, affecting the expression level of putative transcriptional target genes in the
111 loops formed by the chromatin interactions. As experimental issues of the analyses of
112 chromatin interactions, chromatin interaction data are changed, according to
113 experimental techniques, depth of DNA sequencing, and even replication sets of the
114 same cell type. Chromatin interaction data may not be saturated enough to cover all
115 chromatin interactions. Supposing these characters of DNA binding proteins associated
116 with chromatin interactions and avoiding experimental issues of the analyses of
117 chromatin interactions, here I searched for DNA motif sequences of TF and repeat DNA
118 sequences, affecting EPI and the expression level of putative transcriptional target genes
119 in $CD14^+$ monocytes and $CD4^+$ T cells of four people and other cell types without using
120 chromatin interaction data. Then, putative EPI were compared with chromatin

121 interaction data.

122

123 ## Results

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

125 Transcription factor binding sites (TFBS) were predicted using open chromatin
126 regions and DNA motif sequences of transcription factors (TF) collected from various
127 databases and journal papers (see Methods). Transcriptional target genes were predicted
128 using TFBS in promoters and enhancer-promoter association (EPA) shortened at the
129 genomic locations of DNA binding motif sequence of a TF acting as insulator such as
130 CTCF and cohesin (RAD21 and SMC3) (Fig. 1B). To find DNA motif sequences of TF
131 acting as insulator, other than CTCF and cohesin, and repeat DNA sequences affecting
132 the expression level of genes, EPI were predicted based on EPA shortened at the
133 genomic locations of the DNA motif sequence of a TF or a repeat DNA sequence, and
134 transcriptional target genes of each TF bound in enhancers and promoters were
135 predicted based on the EPI. The expression levels of the putative transcriptional target
136 genes were compared with those predicted from promoters using Mann Whiteney U test,
137 two-sided (p-value < 0.05) (Fig. 2A). The number of TF showing a significant
138 difference of expression level of their putative transcriptional target genes was counted.
139 To examine whether the orientation of a DNA motif sequence, which is supposed to act
140 as insulator sites and shorten EPA, affected the number of TF showing a significant
141 difference of expression level of their putative transcriptional target genes, the number
142 of the TF was compared among forward-reverse (FR), reverse-forward (RF), and any

143 orientation (i.e. without considering orientation) of a DNA motif sequence shortening
144 EPA, using chi-square test (p-value < 0.05). To avoid missing DNA motif sequences
145 showing a relatively weak statistical significance by multiple testing collection, the
146 above analyses were conducted using monocytes of four people independently, and
147 DNA motif sequences found in monocytes of four people in common were selected.
148 Total 287 of biased (159 FR and 153 RF) orientation of DNA binding motif sequences
149 of TF were found in monocytes of four people in common, whereas only one any
150 orientation of DNA binding motif sequence was found in monocytes of four people in
151 common (Fig. 2B; Table 1; Supplemental Table S1). FR orientation of DNA motif
152 sequences included CTCF, cohesin (RAD21 and SMC3), ZNF143 and YY1, which are
153 associated with chromatin interactions and EPI. SMARCA4 (BRG1) is associated with
154 topologically associated domain (TAD), which is higher-order chromatin organization.
155 The DNA binding motif sequence of SMARCA4 was not registered in the databases
156 used in this study. FR orientation of DNA motif sequences included SMARCC2, a
157 member of the same SWI/SNF family of proteins as SMARCA4.

158 The same analysis was conducted using DNase-seq data of $CD4^+$ T cells of four
159 people. Total 265 of biased (158 FR and 139 RF) orientation of DNA binding motif
160 sequences of TF were found in T cells of four people in common, whereas only four
161 any orientation (i.e. without considering orientation) of DNA binding motif sequences
162 were found in T cells of four people in common (Supplemental Fig. S1 and
163 Supplemental Table S2). Biased orientation of DNA motif sequences in T cells included
164 CTCF, cohesin (RAD21 and SMC3), ZNF143, YY1, and SMARC. Biased orientation

165 of DNA motif sequences in T cells and monocytes (using H3K27ac histone
166 modification marks, which is described in the next paragraph) included JUNDM2
167 (JDP2), which is involved in histone-chaperone activity, promoting nucleosome, and
168 inhibition of histone acetylation (Jin et al. 2006). JDP2 forms a homodimer or
169 heterodimer with various TF (https://en.wikipedia.org/wiki/Jun_dimerization_protein).
170 Among 287, 61 of biased orientation of DNA binding motif sequences of TF were
171 found in both monocytes and T cells in common (Supplemental Table S5). For each
172 orientation, 31 FR and 34 RF orientation of DNA binding motif sequences of TF were
173 found in both monocytes and T cells in common. Without considering the difference of
174 orientation of DNA binding motif sequences, 84 of biased orientation of DNA binding
175 motif sequences of TF were found in both monocytes and T cells. As a reason for the
176 increase of the number (84) from 61, a TF or an isoform of the same TF may bind to a
177 different DNA binding motif sequence according to cell types and/or in the same cell
178 type. About 50% or more of alternative splicing isoforms are differently expressed
179 among tissues, indicating that most alternative splicing is subject to tissue-specific
180 regulation (Wang et al. 2008) (Chen and Manley 2009) (Das et al. 2007). The same TF
181 has several DNA binding motif sequences and in some cases one of the motif sequences
182 is almost the same as the reverse complement sequence of another motif sequence of the
183 same TF. For example, RAD21 had both FR and RF orientation of DNA motif
184 sequences, but the number of the FR orientation of DNA motif sequence was relatively
185 small in the genome, and the RF orientation of DNA motif sequence was frequently
186 observed and co-localized with CTCF. I previously found that a complex of TF would

187 bind to a slightly different DNA binding motif sequence from the combination of DNA
188 binding motif sequences of TF composing the complex in *C. elegans* (Tabuchi et al.
189 2011). From another viewpoint of this study, the expression level of putative
190 transcription target genes of some TF would be different, depending on the genomic
191 locations (enhancers or promoters) of DNA binding motif sequences of the TF in
192 monocytes and T cells of four people.

193 Moreover, using open chromatin regions overlapped with H3K27ac histone
194 modification marks known as enhancer and promoter marks, the same analyses were
195 performed in monocytes and T cells. H3K27ac histone modification marks were used in
196 the analysis of EPI, but were not used in the analysis of TF as insulator like CTCF and
197 cohesin in this study, since new biased orientation of DNA motif sequences were found
198 in this criterion. When H3K27ac histone modification marks were used in the analysis
199 of TF as insulator like CTCF and cohesin, the number of biased orientation of DNA
200 motif sequences was decreased. Total 241 of biased (141 FR and 120 RF) orientation of
201 DNA binding motif sequences of TF were found in monocytes of four people in
202 common, whereas only one any orientation of DNA binding motif sequence was found
203 (Supplemental Table S3). Though the number of biased orientation of DNA motif
204 sequences was reduced, JDP2 was found in monocytes as well as T cells. CTCF,
205 RAD21, SMC3, ZNF143, YY1, and SMARCC2 were also found. For T cells using
206 H3K27ac histone modification marks, total 195 of biased (107 FR and 110 RF)
207 orientation of DNA binding motif sequences of TF were found in T cells of four people
208 in common, whereas only six any orientation of DNA binding motif sequences were

209 found (Supplemental Table S4). Though the number of biased orientation of DNA motif
210 sequences was reduced, CTCF, RAD21, SMC3, YY1, and SMARCC2 were found.
211 Scores of CTCF, RAD21, and SMC3 were increased compared with the result of T cells
212 without using H3K27ac histone modification marks, and they were ranked in the top six.
213 As summary of the results with and without H3K27ac histone modification marks, total
214 386 of biased (223 FR and 214 RF) orientation of DNA motif sequences were found in
215 monocytes of four people in common. Total 348 of biased (202 FR and 198 RF)
216 orientation of DNA motif sequences were found in T cells of four people in common.
217 Total number of these results in monocytes and T cells was 590 biased (368 FR and 359
218 RF) orientation of DNA motif sequences. Biased orientation of DNA motif sequences
219 found in both monocytes and T cells were listed in Supplemental Table S5.

220 To examine whether the biased orientation of DNA binding motif sequences of
221 TF were observed in other cell types, the same analyses were conducted in other cell
222 types. However, for other cell types, experimental data of one sample were available in
223 ENCODE database, so the analyses of DNA motif sequences were performed by
224 comparing with the result in monocytes of four people. Among the biased orientation of
225 DNA binding motif sequences found in monocytes, 69, 82, 56, and 53 DNA binding
226 motif sequences were also observed in H1-hESC, iPS, Huvec and MCF-7 respectively,
227 including CTCF and cohesin (RAD21 and SMC3) (Table 2; Supplemental Table S6).
228 The scores of DNA binding motif sequences were the highest in monocytes, and the
229 other cell types showed lower scores. The results of the analysis of DNA motif
230 sequences in $CD20^+$ B cells and macrophages did not include CTCF and cohesin,

231 because these analyses can be utilized in cells where the expression level of putative
232 transcriptional target genes of each TF show a significant difference between promoters
233 and EPA shortened at the genomic locations of a DNA motif sequence acting as
234 insulator sites. Some experimental data of a cell did not show a significant difference
235 between promoters and the EPA (Osato 2018).

236 Instead of DNA binding motif sequences of TF, repeat DNA sequences were also
237 examined. The expression levels of transcriptional target genes of each TF predicted
238 based on EPA that were shortened at the genomic locations of a repeat DNA sequence
239 were compared with those predicted from promoters. Three RF orientation of repeat
240 DNA sequences showed a significant difference of expression level of putative
241 transcriptional target genes in monocytes of four people in common (Table 3). Among
242 them, LTR16C repeat DNA sequence was observed in iPS and H1-hESC with enough
243 statistical significance considering multiple tests (p-value $< 10^{-7}$). The same as $CD14^{+}$
244 monocytes, biased orientation of repeat DNA sequences were examined in $CD4^{+}$ T cells.
245 Three FR and two RF orientation of repeat DNA sequences showed a significant
246 difference of expression level of putative transcriptional target genes in T cells of four
247 people in common (Supplemental Table S7). MIRb and MIR3 were also found in the
248 analysis using open chromatin regions overlapped with H3K27ac histone modification
249 marks, which are enhancer and promoter marks. MIR and other transposon sequences
250 are known to act as insulators and enhancers (Bejerano et al. 2006; Rebollo et al. 2012;
251 de Souza et al. 2013; Jjingo et al. 2014; Wang et al. 2015).
252

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

254 To examine the association of 287 biased (FR and RF) orientation of DNA
255 binding motif sequences of TF, co-location of the DNA binding motif sequences in
256 open chromatin regions was analyzed in monocytes. The number of open chromatin
257 regions with the same pairs of DNA binding motif sequences was counted, and when
258 the pairs of DNA binding motif sequences were enriched with statistical significance
259 (chi-square test, p-value $< 1.0 \times 10^{-10}$), they were listed (Table 4; Supplemental Table
260 S8; Fig. 3A). Open chromatin regions overlapped with histone modification of enhancer
261 and promoter marks (H3K27ac) (total 26,095 regions) showed a larger number of
262 enriched pairs of DNA motifs than all open chromatin regions (Table 4; Supplemental
263 Table S8; Fig. 3A). H3K27ac is known to be enriched at chromatin interaction anchors
264 (Phanstiel et al. 2017). As already known, CTCF was found with cohesin such as
265 RAD21 and SMC3 (Table 4; Fig. 3A). Top 30 pairs of FR and RF orientations of DNA
266 motifs co-occupied in the same open chromatin regions were shown (Table 4). Total
267 number of pairs of DNA motifs was 492, consisting of 115 unique DNA motifs, when
268 the pair of DNA motifs were observed in more than 80% of the number of open
269 chromatin regions with the DNA motifs. Biased orientation of DNA binding motif
270 sequences of TF tended to be co-localized in the same open chromatin regions.

271 To examine the association of 265 biased orientation of DNA binding motif
272 sequences of TF in $CD4^+$ T cells, co-location of the DNA binding motif sequences in
273 open chromatin regions was analyzed. Top 30 pairs of FR and RF orientations of DNA
274 motifs co-occupied in the same open chromatin regions were shown (Supplemental

275 Table S9; Supplemental Fig. S2). Total number of pairs of DNA motifs was 332,
276 consisting of 118 unique DNA motifs, when the pair of DNA motifs were observed in
277 more than 60% of the number of open chromatin regions with the DNA motifs
278 (chi-square test, p-value $< 1.0 \times 10^{-10}$). Among them, 12 pairs of DNA motif
279 sequences including a pair of CTCF, SMC3, and RAD21 in T cells were found in
280 monocytes in common (Supplemental Table S9).
281
282 **Pairs of biased orientation of DNA motif sequences surrounding genes**
283 CTCF and cohesin are found in chromatin interaction anchors and are considered
284 to form homodimers. TF are also known to form heterodimers with another TF. In this
285 study, if the DNA binding motif sequence of only the mate to a pair of TF was found in
286 EPA, EPA was shortened at one side, which is the genomic location of the DNA
287 binding motif sequence of the mate to the pair, and transcriptional target genes were
288 predicted using the EPA shortened at the side. In this analysis, the mate to both
289 heterodimer and homodimer of TF can be used to examine the effect on the expression
290 level of transcriptional target genes predicted based on the EPA shortened at one side.
291 To examine whether biased orientation of DNA motif sequences tend to be found as a
292 pair of the same TF or different TF in upstream and downstream of genes, the
293 enrichment of pairs of biased orientation of DNA motif sequences was investigated. The
294 number of genes with a pair of FR or RF orientation of DNA motif sequences was
295 counted using all pairs of the DNA motif sequences, and statistical tests (chi-square
296 test) were conducted. The total number of genes (transcripts) with a biased orientation

297 of DNA motif sequence in their upstream or/and downstream was about 20,000. This is
298 almost the same way of analysis as co-location of biased orientation of DNA motif
299 sequences, and the analysis here used the number of genes instead of open chromatin
300 regions. Fig. 3B and Supplemental Fig. S3 showed that not only pairs of the same TF,
301 but also pairs of different TF were found in upstream and downstream of genes. The
302 number of DNA motif sequences of the same TF was assigned independently in each
303 figure of heat map, so for example, 'CTCF_1' in Fig. 3B may not be the same DNA
304 motif sequence of 'CTCF_1' in another figure in Supplemental Fig. S3. DNA motif
305 sequences of CTCF and cohesin (RAD21 and SMC3) were enriched near the same
306 genes, but cohesin tended to be found with another TF near the same genes as well. The
307 number of some pairs of biased orientation of DNA motif sequences in genome
308 sequences was small, though they showed statistical significance (Supplemental
309 material 2). The result of the analysis seems to be different between monocytes and T
310 cells and between using H3K27ac histone modification marks and not using them
311 (Supplemental Fig. S3). Some of these TF would form homodimers or heterodimers
312 directory or indirectly composing a complex structure, contributing to make chromatin
313 interactions (Fig. 5).
314

315 **Comparison with chromatin interaction data**

316 To examine whether the biased orientation of DNA motif sequences is associated
317 with chromatin interactions, enhancer-promoter interactions (EPI) predicted based on
318 enhancer-promoter associations (EPA) were compared with chromatin interaction data

319 (Hi-C). Due to the resolution of Hi-C experimental data used in this study (50kb), EPI
320 were adjusted to 50kb resolution. EPI were predicted based on three types of EPA: (i)
321 EPA shortened at the genomic locations of FR or RF orientation of DNA motif
322 sequence of a TF, (ii) EPA shortened at the genomic locations of DNA motif sequence
323 of a TF acting as insulator sites such as CTCF and cohesin (RAD21, SMC3) without
324 considering their orientation, and (iii) EPA without being shortened by the genomic
325 locations of a DNA motif sequence. EPA (i) showed a significantly higher ratio of EPI
326 overlapped with chromatin interactions (Hi-C) using DNA binding motif sequences of
327 CTCF and cohesin (RAD21 and SMC3) than the other two types of EPAs (n = 4,
328 binomial test, two-sided) (Supplemental Fig. S4). Total 59 biased orientation (40 FR
329 and 23 RF) of DNA motif sequences including CTCF, cohesin, and YY1 showed a
330 significantly higher ratio of EPI overlapped with Hi-C chromatin interactions (a cutoff
331 score of CHiCAGO tool > 1) than the other types of EPAs in monocytes (Supplemental
332 Table S10). When comparing EPI predicted based on only EPA (i) and (iii) with
333 chromatin interactions, total 174 biased orientation (101 FR and 84 RF) of DNA motif
334 sequences showed a significantly higher ratio of EPI predicted based on EPA (i)
335 overlapped with the chromatin interactions than EPI predicted based on EPA (iii)
336 (Supplemental material 2). The difference between EPI predicted based on EPA (i) and
337 (ii) seemed to be difficult to distinguish using the chromatin interaction data and the
338 statistical test in some cases. However, as for the difference between EPI predicted
339 based on EPA (i) and (iii), a larger number of biased orientation of DNA motif
340 sequences was found to be correlated with chromatin interaction data. Chromatin

341 interaction data were obtained from different samples from DNase-seq, open chromatin
342 regions, so individual differences seemed to be large from the results of this analysis.
343 Since, for some DNA motif sequences of transcription factors, the number of EPI
344 overlapped with chromatin interactions was small, if higher resolution of chromatin
345 interaction data (such as HiChIP, in situ DNase Hi-C, and in situ Hi-C data, or a tool to
346 improve the resolution such as HiCPlus) is available, the number of EPI overlapped
347 with chromatin interactions would be increased and the difference of the numbers
348 among three types of EPA would be larger and more significant (Rao et al. 2014;
349 Ramani et al. 2016; Mumbach et al. 2017; Zhang et al. 2018).

350 After the analysis of $CD14^+$ monocytes, to utilize HiChIP chromatin interaction
351 data in $CD4^+$ T cells, the same analysis for $CD14^+$ monocytes was performed using
352 DNase-seq data of four donors in $CD4^+$ T cells (Mumbach et al. 2017). EPI predicted
353 based on EPA were compared with HiChIP chromatin interaction data in $CD4^+$ T cells.
354 The resolutions of HiChIP chromatin interaction data and EPI were adjusted to 5kb. EPI
355 were predicted based on the three types of EPA in the same way as $CD14^+$ monocytes
356 using top 60% of all the transcript (gene) excluding transcripts not expressed in T cells.
357 The criteria of the analysis were determined to include known DNA motif sequences
358 involved in chromatin interactions such as CTCF and cohesin in the result, and the
359 result was consistent with that using Hi-C chromatin interaction data. EPA (iii) showed
360 the highest ratio of EPI overlapped with chromatin interactions (HiChIP) using DNA
361 binding motif sequences of CTCF and cohesin (RAD21 and SMC3), compared with the
362 other two types of EPAs (n = 4, binomial test, two-sided) (Fig. 4). Total 121 biased

363 orientation (73 FR and 58 RF) of DNA motif sequences including CTCF, cohesin,
364 ZNF143, and SMARC showed a significantly higher ratio of EPI overlapped with
365 HiChIP chromatin interactions (more than 1,000 counts for each interaction) than the
366 other types of EPAs in T cells (Table 5). When comparing EPI predicted based on only
367 EPA (i) and (iii) with the chromatin interactions, total 226 biased orientation (132 FR
368 and 123 RF) of DNA motif sequences showed a significantly higher ratio of EPI
369 predicted based on EPA (i) overlapped with the chromatin interactions than EPI
370 predicted based on EPA (iii) (Supplemental material 2). As expected, the number of EPI
371 overlapped with chromatin interactions (HiChIP) was increased, compared with Hi-C
372 chromatin interactions. Most of biased orientation of DNA motif sequences (85%) were
373 found to be correlated with chromatin interactions, when comparing EPI predicted
374 based on EPA (i) and (iii) with HiChIP chromatin interactions.

375 Though the ratios of EPI overlapped with chromatin interactions were increased
376 by using many chromatin interaction data including lower score and count of chromatin
377 interactions (a cutoff score of CHiCAGO tool > 1 for Hi-C and more than 1,000 counts
378 for HiChIP), the ratios of EPI overlapped with chromatin interactions showed the same
379 tendency among the three types of EPAs. The ratio of EPI overlapped with Hi-C
380 chromatin interactions was increased using H3K27ac marks in both monocytes and T
381 cells. The ratio of EPI overlapped with HiChIP chromatin interactions was also
382 increased using H3K27ac marks. Chromatin interaction data were obtained from
383 different samples from DNase-seq, open chromatin regions in $CD4^+$ T cells, so
384 individual differences seemed to be large from the results of this analysis, and

385 (Mumbach et al. 2017) suggested that individual differences of chromatin interactions
386 were larger than those of open chromatin regions. ATAC-seq data, open chromatin
387 regions were available in $CD4^+$ T cells in the paper, however, when using ATAC-seq
388 data, the result of the analysis of biased orientation of DNA motif sequences was
389 different from DNase-seq data, and not included a part of CTCF and cohesin. Thus,
390 DNase-seq data collected from ENCODE and Blueprint projects were employed in this
391 study.
392

393 **Discussion**

394 To find DNA motif sequences of transcription factors (TF) and repeat DNA
395 sequences affecting the expression level of human putative transcriptional target genes,
396 the DNA motif sequences were searched from open chromatin regions of monocytes of
397 four people. Total 287 biased [159 forward-reverse (FR) and 153 reverse-forward (RF)]
398 orientation of DNA motif sequences of TF were found in monocytes of four people in
399 common, whereas only one any orientation (i.e. without considering orientation) of
400 DNA motif sequence of TF was found to affect the expression level of putative
401 transcriptional target genes, suggesting that enhancer-promoter association (EPA)
402 shortened at the genomic locations of FR or RF orientation of the DNA motif sequence
403 of a TF or a repeat DNA sequence is an important character for the prediction of
404 enhancer-promoter interactions (EPI) and the transcriptional regulation of genes.

405 When DNA motif sequences were searched from monocytes of one person, a
406 larger number of biased orientation of DNA motif sequences affecting the expression

407 level of human putative transcriptional target genes were found. When the number of
408 donors, from which experimental data were obtained, was increased, the number of
409 DNA motif sequences found in all people in common decreased and in some cases,
410 known transcription factors involved in chromatin interactions such as CTCF and
411 cohesin (RAD21 and SMC3) were not identified by statistical tests. This would be
412 caused by individual difference of the same cell type, low quality of experimental data,
413 and experimental errors. Moreover, though FR orientation of DNA binding motif
414 sequences of CTCF and cohesin is frequently observed at chromatin interaction anchors,
415 the percentage of FR orientation is not 100, and other orientations of the DNA binding
416 motif sequences are also observed. Though DNA binding motif sequences of CTCF and
417 cohesin are found in various open chromatin regions, DNA binding motif sequences of
418 some TF would be observed less frequently in open chromatin regions. The analyses of
419 experimental data of a number of people would avoid missing relatively weak statistical
420 significance of DNA motif sequences of TF in experimental data of each person by
421 multiple testing correction of thousands of statistical tests. A DNA motif sequence was
422 found with p-value < 0.05 in experimental data of one person and the DNA motif
423 sequence found in the same cell type of four people in common would have p-value <
424 $0.05^4 = 6.25 \times 10^{-6}$. Actually, DNA motif sequences with p-value slightly less than
425 0.05 in monocytes of one person were observed in monocytes of four people in
426 common.

427 EPI were compared with chromatin interactions (Hi-C) in monocytes. EPAs

428 shortened at the genomic locations of DNA binding motif sequences of CTCF and
429 cohesin (RAD21 and SMC3) showed a significant difference of the ratios of EPI
430 overlapped with chromatin interactions, according to three types of EPAs (see Methods).
431 Using open chromatin regions overlapped with ChIP-seq experimental data of histone
432 modification of an enhancer mark (H3K27ac), the ratio of EPI not overlapped with
433 Hi-C was reduced. (Phanstiel et al. 2017) also reported that there was an especially
434 strong enrichment for loops with H3K27 acetylation peaks at both ends (Fisher's Exact
435 Test, $p = 1.4 \times 10^{-27}$). However, the total number of EPI overlapped with chromatin
436 interactions was also reduced using H3K27ac peaks, so more chromatin interaction data
437 would be needed to obtain reliable results in this analysis. As an issue of experimental
438 data, data for chromatin interactions and open chromatin regions were came from
439 different samples and donors, so individual differences would exist in the data.
440 Moreover, the resolution of chromatin interaction data used in monocytes was about
441 50kb, thus the number of chromatin interactions was relatively small (72,284 at 50kb
442 resolution with a cutoff score of CHiCAGO tool > 1 and 16,501 with a cutoff score of
443 CHiCAGO tool > 5). EPI predicted based on EPA shortened at the genomic locations of
444 DNA binding motif sequence of TF that were found in various open chromatin regions
445 such as CTCF and cohesin (RAD21 and SMC3) tended to be overlapped with a larger
446 number of chromatin interactions than TF less frequently observed in open chromatin
447 regions. Therefore, to examine the difference of the numbers of EPI overlapped with
448 chromatin interactions, according to the three types of EPAs, the number of chromatin
449 interactions should be large enough.

450 As HiChIP chromatin interaction data were available in $CD4^+$ T cells, biased
451 orientation of DNA motif sequences of TF were examined in T cells using DNase-seq
452 data of four people. The resolutions of chromatin interactions and EPI were adjusted to
453 5kb by fragmentation of genome sequences. In monocytes, the resolution of Hi-C
454 chromatin interaction data was converted by extending anchor regions of chromatin
455 interactions to 50kb length and merging the chromatin interactions overlapped with
456 each other. Fragmentation of genome sequences may affect the classification of
457 chromatin interactions of which anchors are located near the border of a fragment, but
458 the number of chromatin interactions would not be decreased, compared with merging
459 chromatin interactions. The number of HiChIP chromatin interactions was 19,926,360
460 at 5kb resolution, 666,149 at 5kb resolution with chromatin interactions (more than
461 1,000 counts for each interaction), and 78,209 at 5kb resolution with chromatin
462 interactions (more than 6,000 counts for each interaction). As expected, the number of
463 EPI overlapped with chromatin interactions was increased, and 46 – 85% of biased
464 orientation of DNA motif sequences of TF showed a statistical significance in EPI
465 predicted based on EPA shortened at the genomic locations of the DNA motif sequence,
466 compared with the other types of EPAs or EPA not shortened. False positive predictions
467 of EPI would be decreased by using H3K27ac marks and other features. The ratio of
468 EPI overlapped with Hi-C chromatin interactions was increased using H3K27ac marks
469 in both monocytes and T cells. The ratio of EPI overlapped with HiChIP chromatin
470 interactions was also increased using H3K27ac marks. However, the number of biased
471 orientation of DNA motif sequences showing a higher ratio of EPI overlapped with

472 HiChIP chromatin interactions than the other types of EPAs was decreased using
473 H3K27ac marks (Supplemental material 2).
474 Though the number of known DNA binding proteins and TF involved in
475 chromatin interactions would be less than 30 or so, about 300 DNA binding proteins
476 showed enrichment of biased orientation of their DNA binding sequences, and known
477 DNA binding proteins involved in chromatin interactions were found in the biased
478 orientation of DNA binding sequences. JUNDM2 (JDP2) found in T cells and
479 monocytes is not known DNA binding proteins involved in chromatin interactions.
480 JDP2 is involved in histone and nucleosome-related functions such as
481 histone-chaperone activity, promoting nucleosome, and inhibition of histone acetylation
482 (Jin et al. 2006). JDP2 forms a homodimer or heterodimer with various TF
483 (https://en.wikipedia.org/wiki/Jun_dimerization_protein). Thus, the main function of
484 JDP2 may not be associated with chromatin interactions, but it forms a homodimer or
485 heterodimer with TF, potentially forming a chromatin interaction through the binding of
486 JDP2 and another TF, like chromatin interactions formed by YY1 (Weintraub et al.
487 2017). In the analysis of HiChIP chromatin interaction data, JDP2 was in the list of
488 biased orientation of DNA motif sequences showing a higher ratio of EPI overlapped
489 with HiChIP chromatin interactions than the other types of EPAs, using chromatin
490 interactions with more than 6,000 counts for each interaction (Supplemental material 2).
491 When forming a homodimer or heterodimer with another TF, TF may bind to genome
492 DNA with a specific orientation of their DNA binding sequences. From the analysis of
493 biased orientation of DNA motif sequences of TF, TF forming heterodimer would also

494 be found. If the DNA binding motif sequence of only the mate to a pair of TF was
495 found in EPA, EPA was shortened at one side, which is the genomic location of the
496 DNA binding motif sequence of the mate to the pair, and transcriptional target genes
497 were predicted using the EPA shortened at the side. In this analysis, the mate to both
498 heterodimer and homodimer of TF can be used to examine the effect on the expression
499 level of transcriptional target genes predicted based on the EPA shortened at one side.
500 For heterodimer, biased orientation of DNA motif sequences may also be found in
501 forward-forward or reverse-reverse orientation.

502 I reported that the ratio of the number of functional enrichment of putative
503 transcriptional target genes in the total number of putative transcriptional target genes
504 changed according to the types of EPAs, in the same way as gene expression level
505 (Osato 2018). JDP2 showed a higher ratio of functional enrichment of putative
506 transcriptional target genes using EPA shortened at the genomic locations of FR
507 orientation of the DNA binding motif sequence, compared with the other types of EPAs
508 (RF or any orientation) in monocytes and T cells, among monocytes, T cells and $CD20^+$
509 B cells. While the genome-wide analysis of functional enrichment of putative
510 transcriptional target genes takes computational time, the genome-wide analysis of gene
511 expression level is finished with less computational time. The analysis of gene
512 expression level can be used for a cell type and experimental data where the expression
513 levels of transcriptional target genes predicted based on EPA are significantly different
514 from those predicted from only promoter.

515 It has been reported that CTCF and cohesin-binding sites are frequently mutated

516 in cancer (Katainen et al. 2015). Some biased orientation of DNA motif sequences
517 would be associated with chromatin interactions and might be associated with diseases
518 including cancer.

519 The analyses in this study revealed novel characters of DNA binding motif
520 sequences of TF and repeat DNA sequences, and would be useful to analyze TF
521 involved in chromatin interactions and/or forming a homodimer, heterodimer or
522 complex with other TF, affecting the transcriptional regulation of genes.

523

524 ## Methods

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

526 To examine transcriptional regulatory target genes of transcription factors (TF),
527 bed files of hg38 of Blueprint DNase-seq data for $CD14^+$ monocytes of four donors
528 (EGAD00001002286; Donor ID: C0010K, C0011I, C001UY, C005PS) were obtained
529 from Blueprint project web site (http://dcc.blueprint-epigenome.eu/#/home), and the bed
530 files of hg38 were converted into those of hg19 using Batch Coordinate Conversion
531 (liftOver) web site (https://genome.ucsc.edu/util.html). Bed files of hg19 of ENCODE
532 H1-hESC (GSM816632; UCSC Accession: wgEncodeEH000556), iPSC (GSM816642;
533 UCSC Accession: wgEncodeEH001110), HUVEC (GSM1014528; UCSC Accession:
534 wgEncodeEH002460), and MCF-7 (GSM816627; UCSC Accession:
535 wgEncodeEH000579) were obtained from the ENCODE websites
536 (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeUwDgf/;
537 http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeUwDnase/).

538 As high resolution of chromatin interaction data using HiChIP became available
539 in $CD4^+$ T cells, to promote the same analysis of $CD14^+$ monocytes in $CD4^+$ T cells,
540 DNase-seq data of four donors were obtained from a public database and Blueprint
541 projects web sites. DNase-seq data of only one donor was available in Blueprint project
542 for $CD4^+$ T cells (Donor ID: S008H1), and three other DNase-seq data were obtained
543 from NCBI Gene Expression Omnibus (GEO) database. Though the peak calling of
544 DNase-seq data available in GEO database was different from other DNase-seq data in
545 ENCODE and Blueprint projects where 150bp length of peaks were usually predicted
546 using HotSpot (John et al. 2011), FASTQ files of DNase-seq data were downloaded
547 from NCBI Sequence Read Archive (SRA) database (SRR097566, SRR097618, and
548 SRR171574). Read sequences of the FASTQ files were aligned to the hg19 version of
549 the human genome reference using BWA (Li and Durbin 2009), and the BAM files
550 generated by BWA were converted into SAM files, sorted, and indexed using Samtools
551 (Li et al. 2009). Peaks of the DNase-seq data were predicted using HotSpot-4.1.1.
552 To identify transcription factor binding sites (TFBS) from the DNase-seq data,
553 TRANSFAC (2013.2), JASPAR (2010), UniPROBE, BEEML-PBM, high-throughput
554 SELEX, Human Protein-DNA Interactome, and transcription factor binding sequences
555 of ENCODE ChIP-seq data were used (Wingender et al. 1996; Newburger and Bulyk
556 2009; Portales-Casamar et al. 2010; Xie et al. 2010; Zhao and Stormo 2011; Jolma et al.
557 2013; Kheradpour and Kellis 2014). Position weight matrices of transcription factor
558 binding sequences were transformed into TRANSFAC matrices and then into MEME
559 matrices using in-house scripts and transfac2meme in MEME suite (Bailey et al. 2009).

560 Transcription factor binding sequences of TF derived from vertebrates were used for
561 further analyses. Transcription factor binding sequences were searched from each
562 narrow peak of DNase-seq data using FIMO with *p*-value threshold of 10^{-5} (Grant et al.
563 2011). TF corresponding to transcription factor binding sequences were searched
564 computationally by comparing their names and gene symbols of HGNC (HUGO Gene
565 Nomenclature Committee) -approved gene nomenclature and 31,848 UCSC known
566 canonical transcripts
567 (http://hgdownload.cse.ucsc.edu/goldenpath/hg19/database/knownCanonical.txt.gz), as
568 transcription factor binding sequences were not linked to transcript IDs such as UCSC,
569 RefSeq, and Ensembl transcripts.

570 Target genes of a TF were assigned when its TFBS was found in DNase-seq
571 narrow peaks in promoter or extended regions for enhancer-promoter association of
572 genes (EPA). Promoter and extended regions were defined as follows: promoter regions
573 were those that were within distance of ±5 kb from transcriptional start sites (TSS).
574 Promoter and extended regions were defined as per the following association rule,
575 which is the same as that defined in Figure 3A of a previous study (McLean et al.
576 2010): the single nearest gene association rule, which extends the regulatory domain to
577 the midpoint between the TSS of the gene and that of the nearest gene upstream and
578 downstream without the limitation of extension length. Extended regions for EPA were
579 shortened at the genomic locations of DNA binding sites of a TF that was the closest to
580 a transcriptional start site, and transcriptional target genes were predicted from the
581 shortened enhancer regions using TFBS. Furthermore, promoter and extended regions

582 for EPA were shortened at the genomic locations of forward–reverse (FR) orientation of
583 DNA binding sites of a TF. When forward or reverse orientation of DNA binding sites
584 were continuously located in genome sequences several times, the most external
585 forward–reverse orientation of DNA binding sites were selected. The genomic positions
586 of genes were identified using 'knownGene.txt.gz' file in UCSC bioinformatics sites
587 (Karolchik et al. 2014). The file 'knownCanonical.txt.gz' was also utilized for choosing
588 representative transcripts among various alternate forms for assigning promoter and
589 extended regions for EPA. From the list of transcription factor binding sequences and
590 transcriptional target genes, redundant transcription factor binding sequences were
591 removed by comparing the target genes of a transcription factor binding sequence and
592 its corresponding TF; if identical, one of the transcription factor binding sequences was
593 used. When the number of transcriptional target genes predicted from a transcription
594 factor binding sequence was less than five, the transcription factor binding sequence
595 was omitted.

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

600 For gene expression data, RNA-seq reads mapped onto human hg19 genome
601 sequences were obtained, including ENCODE long RNA-seq reads with poly-A of
602 H1-hESC, iPSC, HUVEC, and MCF-7 (GSM26284, GSM958733, GSM2344099,
603 GSM2344100, GSM958734, and GSM765388), and UCSF-UBC human reference

604 epigenome mapping project RNA-seq reads with poly-A of naive $CD4^+$ T cells
605 (GSM669617). Two replicates were present for H1-hESC, iPSC, HUVEC, and MCF-7,
606 and a single one for $CD4^+$ T cells. RPKMs of the RNA-seq data were calculated using
607 RSeQC (Wang et al. 2012). For monocytes, Blueprint RNA-seq RPKM data
608 (GSE58310, GSE58310_GeneExpression.csv.gz, Monocytes_Day0_RPMI) were used
609 (Saeed et al. 2014). Based on RPKM, UCSC transcripts with expression level among
610 top 30% of all the transcripts were selected in each cell type.

611 The expression level of transcriptional target genes predicted based on EPA
612 shortened at the genomic locations of DNA motif sequence of a TF or a repeat DNA
613 sequence was compared with the expression level of transcriptional target genes
614 predicted from promoter. For each DNA motif sequence shortening EPA,
615 transcriptional target genes were predicted using about 3,000 – 5,000 DNA binding
616 motif sequences of TF, and the expression level of putative transcriptional target genes
617 of each TF was compared between EPA and only promoter using Mann-Whitney test,
618 two-sided (p-value < 0.05). The number of TF showing a significant difference of
619 expression level of putative transcriptional target genes between EPA and promoter was
620 compared among forward-reverse (FR), reverse-forward (RF), and any orientation (i.e.
621 without considering orientation) of a DNA motif sequence shortening EPA using
622 chi-square test (p-value < 0.05). When a DNA motif sequence of a TF or a repeat DNA
623 sequence shortening EPA showed a significant difference of expression level of putative
624 transcriptional target genes among FR, RF, or any orientation in monocytes of four
625 people in common, the DNA motif sequence was listed.

626 Though forward-reverse orientation of DNA binding motif sequences of CTCF
627 and cohesin are frequently observed at chromatin interaction anchors, the percentage of
628 forward-reverse orientation is not 100, and other orientations of the DNA binding motif
629 sequences are also observed. Though DNA binding motif sequences of CTCF and
630 cohesin are found in various open chromatin regions, DNA binding motif sequences of
631 some transcription factors would be observed less frequently in open chromatin regions.
632 The analyses of experimental data of a number of people would avoid missing relatively
633 weak statistical significance of DNA motif sequences in experimental data of each
634 person by multiple testing correction of thousands of statistical tests. A DNA motif
635 sequence was found with p-value < 0.05 in experimental data of one person and the
636 DNA motif sequence found in the same cell type of four people in common would have
637 p-value $< 0.05^4 = 6.25 \times 10^{-6}$.
638
639 **Co-location of biased orientation of DNA motif sequences**
640 Co-location of biased orientation of DNA binding motif sequences of TF was
641 examined. The number of open chromatin regions with the same pair of DNA binding
642 motif sequences was counted, and when the pair of DNA binding motif sequences were
643 enriched with statistical significance (chi-square test, p-value $< 1.0 \times 10^{-10}$), they were
644 listed. For histone modification of an enhancer mark (H3K27ac), bed files of hg38 of
645 Blueprint ChIP-seq data for $CD14^+$ monocytes (EGAD00001001179) and $CD4^+$ T cells
646 (Donor ID: S000RD) were obtained from Blueprint web site

647 (http://dcc.blueprint-epigenome.eu/#/home), and the bed files of hg38 were converted
648 into those of hg19 using Batch Coordinate Conversion (liftOver) web site
649 (https://genome.ucsc.edu/util.html). Networks of co-locations of biased orientation of
650 DNA motif sequences were plotted using Cytoscape v3.71 with yFiles Layout
651 Algorithms for Cytoscape (Shannon et al. 2003).
652
653 **Comparison with chromatin interaction data**
654 For comparison of EPA in monocytes with chromatin interactions,
655 'PCHiC_peak_matrix_cutoff0.txt.gz' file was downloaded from 'Promoter Capture
656 Hi-C in 17 human primary blood cell types' web site (https://osf.io/u8tzp/files/), and
657 chromatin interactions for Monocytes with scores of CHiCAGO tool > 1 and
658 CHiCAGO tool > 5 were extracted from the file (Javierre et al. 2016). In the same way
659 as monocytes, Hi-C chromatin interaction data of $CD4^+$ T cells (Naive $CD4^+$ T cells,
660 nCD4) were obtained.
661 Enhancer-promoter interactions (EPI) were predicted using three types of EPAs
662 in monocytes: (i) EPA shortened at the genomic locations of FR or RF orientation of
663 DNA motif sequence of a TF, (ii) EPA shortened at the genomic locations of any
664 orientation (i.e. without considering orientation) of DNA motif sequence of a TF, and
665 (iii) EPA without being shortened by a DNA motif sequence. EPI predicted using the
666 three types of EPAs in common were removed. First, EPI predicted based on EPA (i)
667 were compared with chromatin interactions (Hi-C). The resolution of chromatin
668 interaction data used in this study was 50kb, so EPI were adjusted to 50kb before their

669 comparison. The number and ratio of EPI overlapped with chromatin interactions were
670 counted. Second, EPI were predicted based on EPA (ii), and EPI predicted based on
671 EPA (i) were removed from the EPI. The number and ratio of EPI overlapped with
672 chromatin interactions were counted. Third, EPI were predicted based on EPA (iii), and
673 EPI predicted based on EPA (i) and (ii) were removed from the EPI. The number and
674 ratio of EPI overlapped with chromatin interactions were counted. The number and ratio
675 of the EPI were compared two times between EPA (i) and (iii), and EPA (i) and (ii)
676 (binomial distribution, p-value < 0.025 for each test).

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

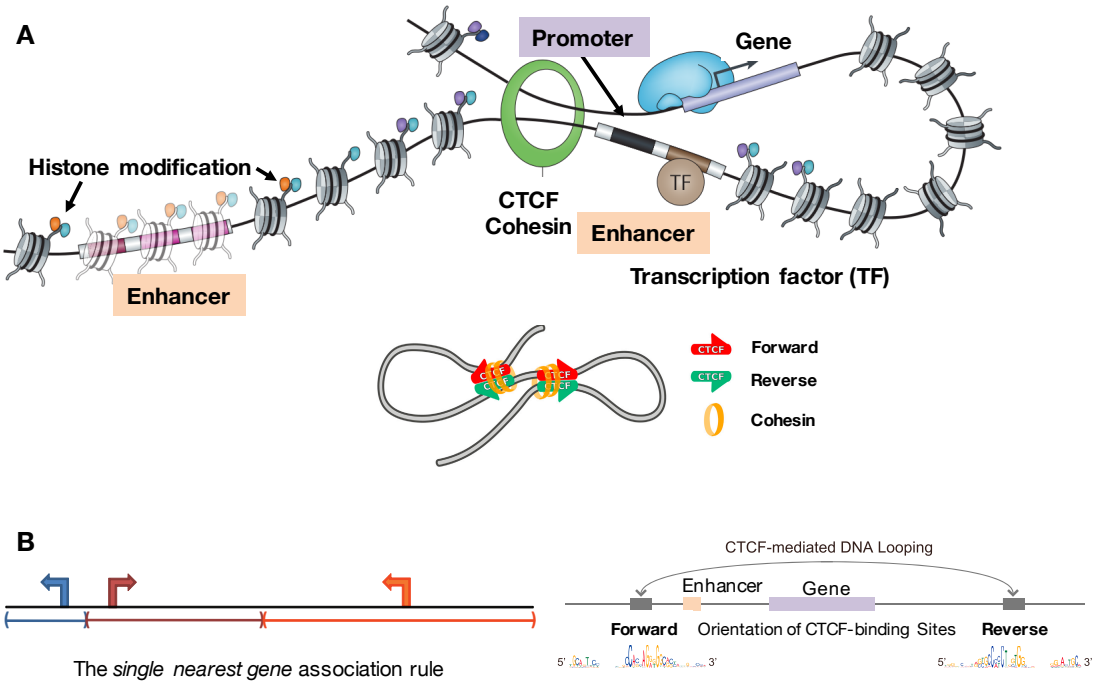
683

684 **Acknowledgements**

685 The supercomputing resource was provided by Human Genome Center of the Institute
686 of Medical Science at the University of Tokyo. Computations were partially performed
687 on the NIG supercomputer at ROIS National Institute of Genetics. Publication charges
688 for this article were funded by JSPS KAKENHI Grant Number 16K00387. This
689 research was partially supported by the Platform Project for Supporting in Drug
690 Discovery and Life Science Research（Platform for Dynamic Approaches to Living

691 System）from Japan Agency for Medical Research and Development (AMED). This
692 research was partially supported by Development of Fundamental Technologies for
693 Diagnosis and Therapy Based upon Epigenome Analysis from Japan Agency for
694 Medical Research and Development (AMED). This work was partially supported by
695 JST CREST Grant Number JPMJCR15G1, Japan.
696

697 **Figures**



698

699 **Figure 1.** Chromatin interactions and enhancer-promoter association. (A) Forward–
700 reverse orientation of CTCF-binding sites are frequently found in chromatin interaction
701 anchors. CTCF can block the interaction between enhancers and promoters limiting the
702 activity of enhancers to certain functional domains (de Wit et al. 2015; Guo et al. 2015).
703 (B) Computationally-defined regulatory domains for enhancer-promoter association
704 (McLean et al. 2010). The *single nearest gene* association rule extends the regulatory
705 domain to the midpoint between this gene's TSS and the nearest gene's TSS both
706 upstream and downstream. Enhancer-promoter association was shortened at the
707 genomic locations of forward-reverse orientation of DNA binding motif sequence of a
708 transcription factor (e.g. CTCF in the figure).

709

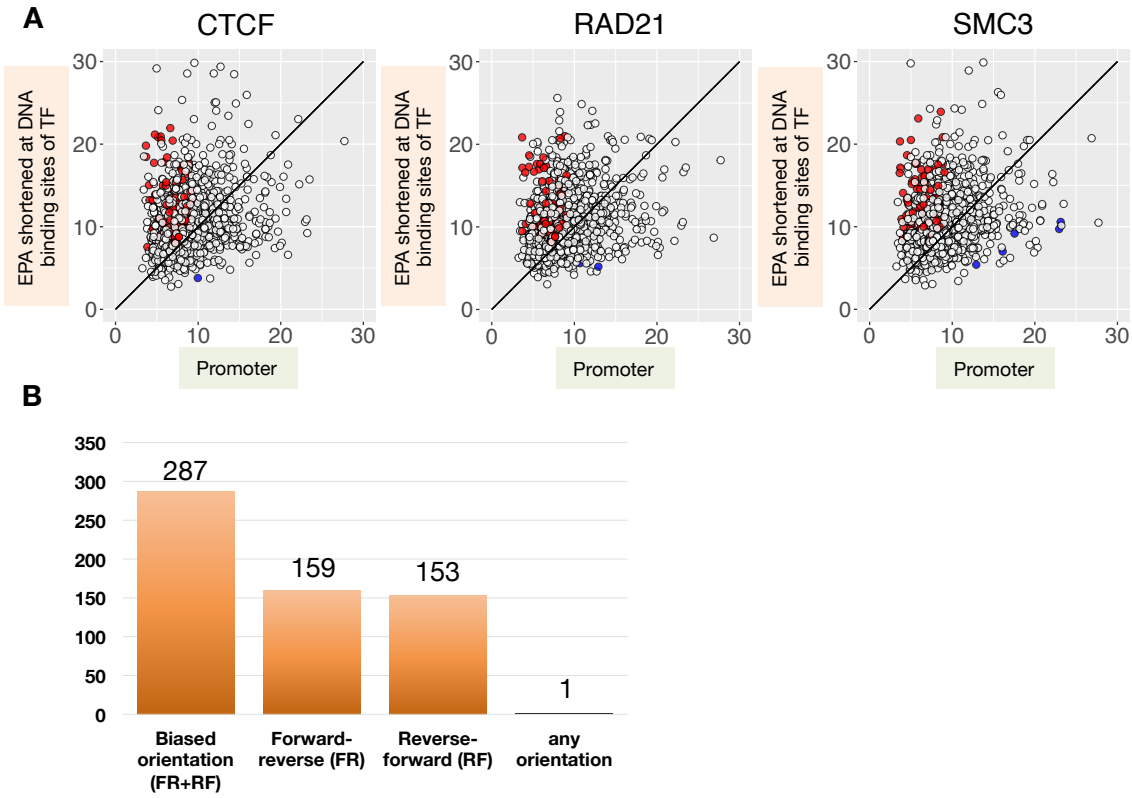
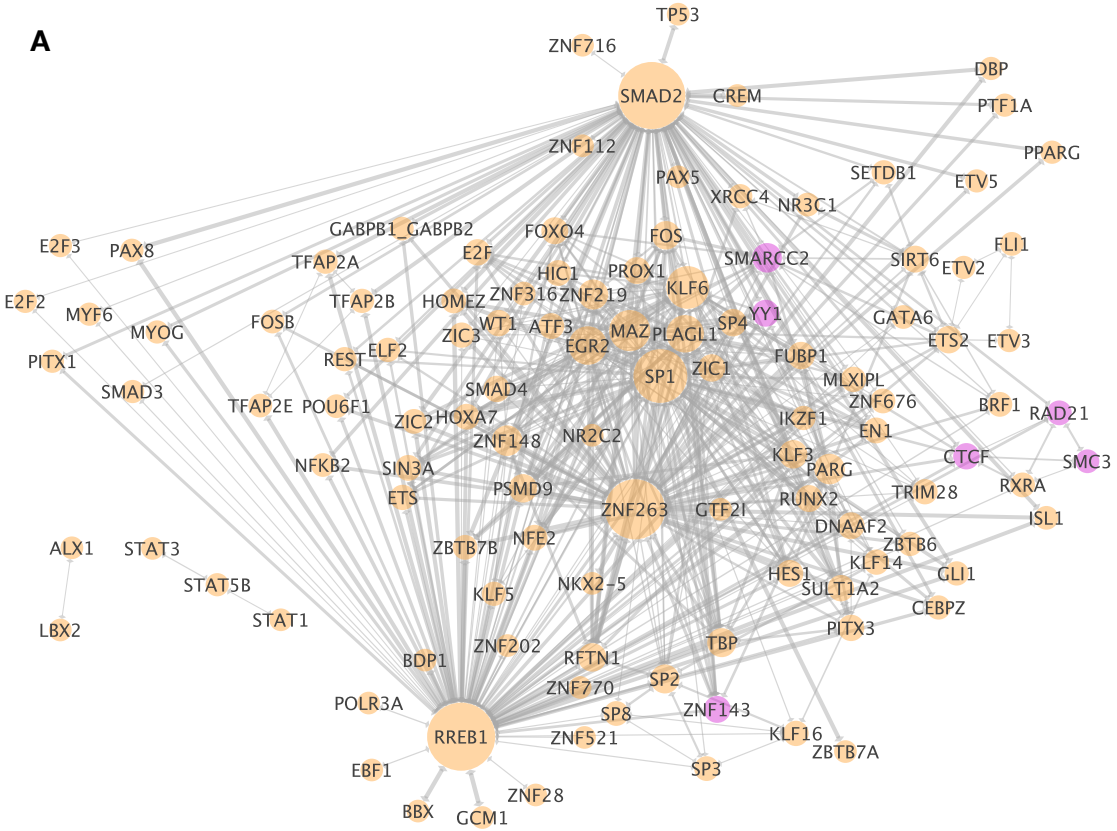
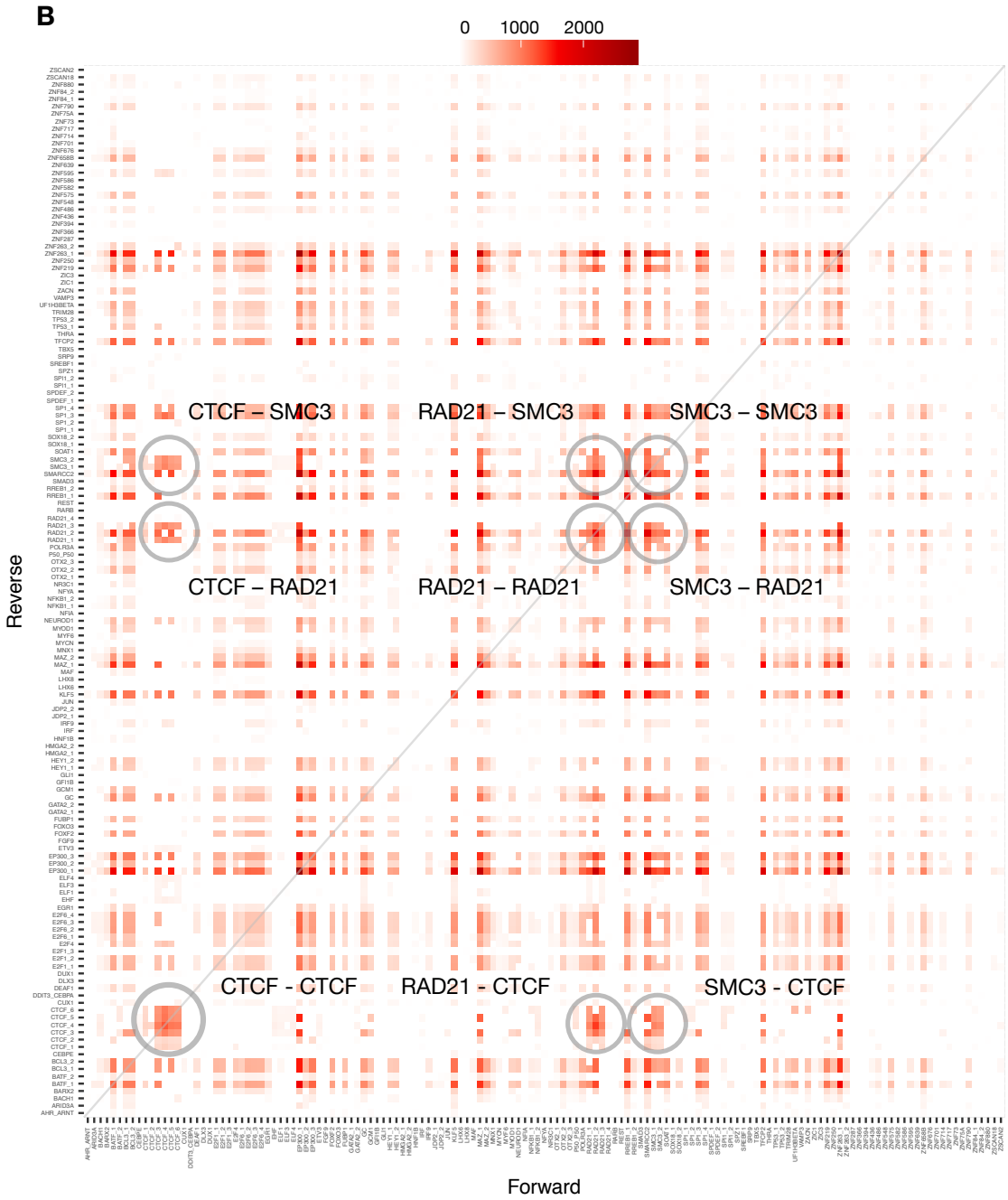


Figure 2. Biased orientation of DNA motif sequences of transcription factors affecting the transcription of genes. (A) Comparison of expression level of putative transcriptional target genes. The median expression levels of target genes of the same transcription factor binding sequences were compared 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

722 expression level of target genes predicted based on the EPA tended to be higher than
723 promoters, implying that transcription factors acted as activators of target genes. (B)
724 Biased orientation of DNA motif sequences of transcription factors in monocytes. Total
725 287 of biased orientation of DNA binding motif sequences of transcription factors were
726 found to affect the expression level of putative transcriptional target genes in monocytes
727 of four people in common, whereas only one any orientation (i.e. without considering
728 orientation) of DNA binding motif sequence was found to affect the expression level.
729



730
731



732

733 **Figure 3.** Association of biased orientation of DNA binding motif sequences of

734 transcription factors. (A) Co-location of biased orientation of DNA binding motif

735 sequences of transcription factors. Co-locations of the DNA motif sequences in open

736 chromatin regions were examined in monocytes. When the same DNA motif sequence

737 of a TF was co-localized with DNA motif sequences of more than one TF in the same or
738 different open chromatin regions, these co-locations of DNA motif sequences were
739 connected to each other. Purple circles of TFs are known to be involved in chromatin
740 interactions. The line thickness is correlated with the number of open chromatin regions
741 with co-location of the same pair of DNA motif sequences. The size of a circle is
742 correlated with the number of unique TF co-localized with the TF of the circle, which is
743 the number of edges of the node. (B) Pairs of biased orientation of DNA binding motif
744 sequences of transcription factors enriched in upstream and downstream of genes. The
745 number of genes with a pair of forward-reverse orientation of DNA motif sequences
746 was counted using all pairs of the DNA motif sequences found in open chromatin
747 regions overlapped with H3K27ac histone modification marks in T cells, and statistical
748 tests (chi-square test) were conducted. The number of genes where a pair of
749 forward-reverse orientation of DNA motif sequences were enriched with p-value < 0.05
750 was counted and shown as heat map. Gray circles show pairs of DNA motif sequences
751 of CTCF and cohesin (RAD21 and SMC3). CTCF and cohesin are known to be
752 involved in chromatin interactions and have biased orientation of DNA binding motif
753 sequences to form chromatin loops. Other pairs of forward-reverse orientation of DNA
754 motif sequences were also enriched in upstream and downstream of genes.
755

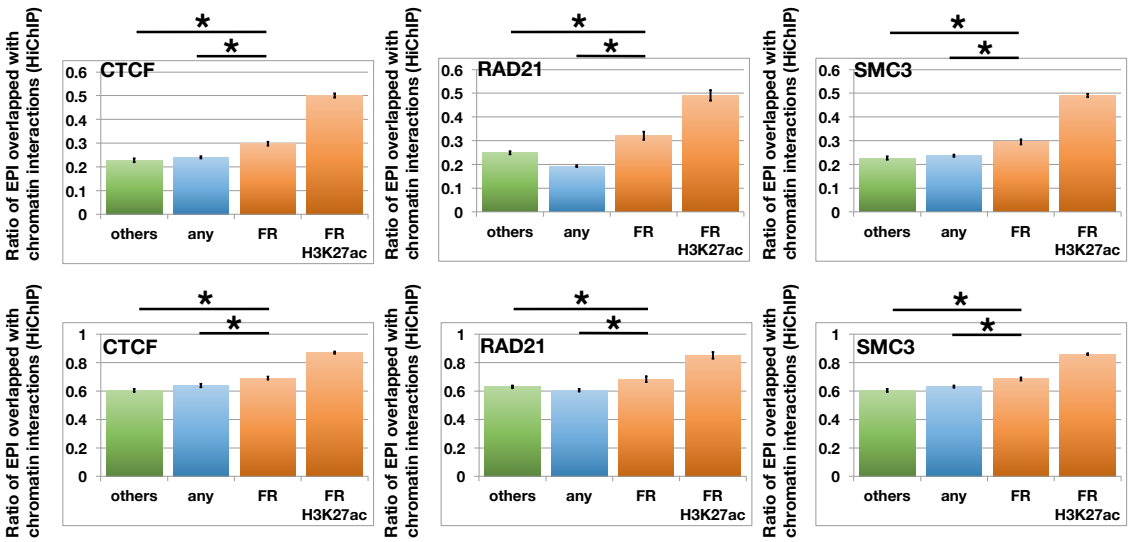


Figure 4. Comparison of enhancer-promoter interactions (EPI) with chromatin interactions in T cells. EPI were predicted based on enhancer-promoter association (EPA) shortened at the genomic locations of biased orientation of DNA binding motif sequence of a transcription factor. Total 121 biased orientation (73 FR and 58 RF) of DNA motif sequences including CTCF and cohesin showed a significantly higher ratio of EPI overlapped with HiChIP chromatin interactions 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^{-4}.

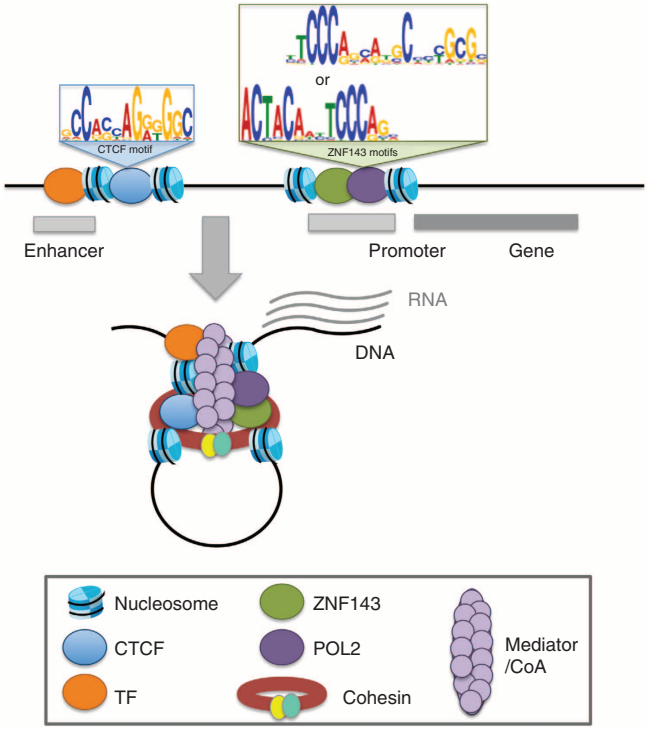


Figure 5. Schematic representation of chromatin interactions involving gene promoters. ZNF143 contributes the formation of chromatin interactions by directly binding the promoter of genes establishing looping with distal element bound by CTCF (Bailey et al. 2015).

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

DNA motif of TF	Score	TF	Score	TF	Score
SRF	75.45	STAT5A	31.24	SMARCC2	21.86
IRF7	71.05	E2F8	30.62	ZNF695	21.78
MINI20	61.68	TP53	30.56	ZNF28	21.77
CTCF	61.62	STAT5B	29.94	ZNF316	21.66
E2F	59.32	STAT3	29.85	TBP	21.60
STAT1	58.94	FOS	29.75	NFIB	21.45
ZNF317	55.97	CEBPZ	29.42	PHOX2B	20.82
CEBPG	55.46	RAD21	28.67	DEC	20.62
ZNF623	54.71	SMAD2	28.60	ZNF709	20.58
ZNF93	54.23	GATA2	28.52	YY1	20.57
HMBOX1	52.99	KLF6	28.02	EIF5A2	20.55
SMAD2_SMAD3_SMAD4	50.66	HOXB1	27.62	ZNF148	20.45
FOXN2	49.88	GLI1	27.09	PAX5	20.05
SPEF1	49.52	GCM1	26.73	POU6F1	19.93
ZNF646	43.95	ZNF676	26.66	FOXO4	19.72
TFE3	42.46	TCF3	26.32	EBF	19.26
MYB	41.98	EHF	26.32	ZNF648	19.12
STAT4	40.54	TCF12	26.29	HENMT1	19.02
ZNF716	40.39	MEF2A	26.22	PARG	18.78
KLF14	39.64	HIC1	25.34	ZNF460	18.64
FOXA1	38.13	EGR2	24.92	SLC25A20	17.83
USF	37.74	NR2F2	24.86	KLF5	17.77
EN1	37.01	PLAGL1	24.65	XRCC4	17.53
ZNF219	35.37	REST	24.53	WT1	17.34
IKZF1	33.79	ZNF143	23.67	SULT1A2	17.34
ZIC1	33.60	YBX1	22.99	FUBP1	17.33
SMC3	33.37	PRDM15	22.63	DNAAF2	17.26
ESR2	32.07	ZNF92	22.22	ZNF521	17.25
TP63	31.98	ZNF214	22.11	XBP1	16.85
ZNF836	31.66	PTF1A	22.09	SP1	16.84

Reverse-forward orientation

TF	Score	TF	Score	TF	Score
RFX5	102.30	TFAP2B	34.55	ZNF449	27.96
TCF7	75.34	TBP	34.47	ZNF233	27.66
SRF	65.52	ZNF670	34.33	SOX2	26.46
ZNF28	63.89	ZNF350	34.30	RXRA	26.07
HNF1B	62.46	ZNF687	33.82	MYF6	25.85
ZNF225	62.18	NFKB2	33.05	MAFK	25.81
CEBPB	59.69	RFX3	32.74	CEBPZ	25.73
NFY	58.96	NFE2L3	32.69	POU2F2	25.59
ZNF286B	54.73	TRIM28	32.27	ETS2	25.05
CTCF	53.16	SP3	32.22	HSF1	24.76
PAX5	52.57	FLI1	32.15	AHR	23.56
STAT1	52.11	RBAK	31.92	PAX6	23.48
ISL1	49.73	NR2F1	31.80	IRF4	22.47
ETV7	47.08	NKX3-1	31.55	ETV3	21.90
KLF16	46.45	ZIC2	31.40	FOXP2	21.72
SP1	46.03	PARP1	31.19	CDX2	21.45
SMAD3	45.27	ZNF195	31.06	NFATC1	21.24
STAT5B	44.72	BDP1	30.84	SP4	21.14
TEAD1	43.00	SOX9	30.72	USF1	20.92
ZNF579	42.94	SP8	30.49	MSX2	20.84
ZBTB2	42.28	ZNF121	30.04	PPARG	20.70
ZNF343	41.01	ZNF202	30.01	MYEF2	20.23
DOBOX5	39.21	CD59	29.85	MYC_MAX	20.22
NANOG	38.84	RXRB	29.57	ZNF711	19.84
NFE2L2	38.53	STAT5A	29.31	ZBTB18	19.60
OBOX2	38.00	SOX11	29.09	HSF2	19.51
ZNF652	37.84	WT1	29.06	DNAAF2	19.00
ZNF30	37.59	CREM	28.65	IRF3	18.88
MYOG	36.34	ZNF316	28.34	TFAP2E	18.87
GABPB1_GABPB2	35.96	HAND1_TCF3	28.07	BCL11A	18.59

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

790 MCF-7

TF	Score
RAD21	16.21
SMC3	15.63
CTCF	15.12
E2F	10.14
SLC25A20	9.62
HOXA7	8.29
STAT5B	7.55
REST	7.25
SIRT6	7.05
KLF3	6.53
YY1	5.30
PLAGL1	4.97
ZNF219	4.91
STAT1	4.43
SNTB1	4.42
ETS	3.16
XRCC4	3.15
ZBTB6	3.12
DBP	2.53
SETDB1	2.53
RREB1	2.50
E2F5	2.38
KLF5	2.33
SPEF1	2.22
ZNF143	2.20
ZNF460	2.19
FOS	2.18
GATA2	1.97
MAZ	1.95
PARG	1.94

iPS

TF	Score
ZIC3	19.32
ZNF521	18.93
TFE3	16.63
ZNF317	16.43
NKX2-5	16.25
HES1	15.34
ZBTB24	14.21
SMC3	13.76
FOS	13.69
RAD21	13.05
ZNF836	11.64
TBX5	11.48
CTCF	11.24
STAT1	11.19
RUNX2	10.28
WT1	10.27
ZIC1	9.60
MAX	8.55
KLF3	8.29
KLF14	8.25
ZNF143	7.81
E2F5	7.66
SIRT6	7.56
EN1	7.49
YY1	7.02
STAT5B	6.73
MAZ	6.39
ZNF148	6.20
TCF3	5.73
EGR2	5.45

791
792
793
794 **Table 3.** Biased orientation of repeat DNA sequences in monocytes. Score: $-\log_{10}$
795 (*p*-value).

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

Repeat DNA seq.	Score
MLT1F1	6.75

796

Reverse-forward (RF) orientation

Repeat DNA seq.	Score
LTR16C	69.39
L1ME4b	23.90
AluSg	23.61

797

798 **Table 4.** Top 30 of co-locations of biased orientation of DNA binding motif sequences
799 of transcription factors in monocytes. Co-locations of DNA motif sequence of CTCF
800 with another biased orientation of DNA motif sequence were shown in a separate table.
801 Motif 1,2: DNA binding motif sequences of transcription factors. # both: the number of
802 open chromatin regions including both Motif 1 and Motif 2. # motif 1: the number of
803 open chromatin regions including Motif 1. # motif 2: the number of open chromatin
804 regions including Motif 2. # others: the number of open chromatin regions not including
805 Motif 1 and Motif 2. Open chromatin regions overlapped with histone modification
806 marks (H3K27ac) were used (Total 26,095 regions).

807

Motif 1	Motif 2	# both	# motif 1	# motif 2	# others	*p*-value
CTCF	SMAD2	4527	781	4478	16309	0
CTCF	RREB1	4379	929	4757	16030	0
CTCF	ZNF263	4357	951	4875	15912	0
CTCF	SP1	4321	987	4052	16735	0
CTCF	RAD21	3804	1504	225	20562	0
CTCF	KLF6	3695	852	3800	17748	0
CTCF	SMC3	1259	178	1977	22681	0
CTCF	RXRA	1220	259	2559	22057	0

808

Motif 1	Motif 2	# both	# motif 1	# motif 2	# others	*p*-value
SMAD2	ZNF263	7427	1578	1805	15285	0
SMARCC2	ZNF263	7354	1192	1893	15656	0
RREB1	SP1	7208	1957	1165	15765	0
SIRT6	ZNF263	7126	1154	2121	15694	0
MAZ	SP1	7052	909	1321	16813	0
MAZ	ZNF263	7022	999	2225	15849	0
SP1	ZNF263	7009	1364	2223	15499	0
SMAD2	SP1	6925	2080	1448	15642	0
MAZ	RREB1	6920	967	2245	15963	0
SIRT6	SMAD2	6766	1514	2239	15576	0
SIRT6	SMARCC2	6668	1612	1878	15937	0
PLAGL1	SMAD2	6637	1074	2368	16016	0
MAZ	SMAD2	6627	1334	2378	15756	0
PARG	SMAD2	6602	1012	2403	16078	0
PLAGL1	ZNF263	6571	1140	2661	15723	0
EGR2	RREB1	6565	787	2600	16143	0
KLF6	RREB1	6563	932	2573	16027	0
MAZ	SMARCC2	6562	1459	1984	16090	0
RREB1	ZNF263	6551	1607	2681	15256	0
KLF6	SP1	6541	954	1832	16768	0
PLAGL1	RREB1	6496	1215	2640	15744	0
EGR2	SP1	6492	860	1881	16862	0
KLF6	SMAD2	6491	1004	2514	16086	0
PARG	ZNF263	6390	1224	2842	15639	0
PLAGL1	SP1	6318	1393	2055	16329	0
KLF6	ZNF263	6304	1191	2928	15672	0
EGR2	SMAD2	6270	1082	2735	16008	0
EGR2	ZNF263	6253	1099	2979	15764	0
EGR2	MAZ	6219	1133	1742	17001	0
PARG	RREB1	6219	1395	2917	15564	0

809 **Table 5.** Comparison of enhancer-promoter interactions (EPI) with chromatin
810 interactions in T cells. TF: DNA binding motif sequence of a transcription factor. Score:
811 $-\log_{10}$ (*p*-value). Inf: *p*-value = 0. Score 1: Comparison of EPA shortened at the
812 genomic locations of FR or RF orientation of DNA motif sequence with EPA not
813 shortened. Score 2: Comparison of EPA shortened at the genomic locations of FR or RF
814 orientation of DNA motif sequence with EPA shortened at the genomic locations of any
815 orientation of DNA motif sequence.

816

817 Forward-reverse (FR) orientation

TF	Score 1	Score 2
APEX1	13.57	3.85
AR	Inf	8.30
CDC5L	65.46	12.23
CTCF	101.79	14.94
DDIT3_CEBPA	249.72	4.78
EBF1	48.10	2.27
EGR1	97.56	6.09
EGR3	130.86	2.73
EGR4	59.94	2.91
ELF5	119.97	2.85
EP300	28.41	15.12
ETV4	270.06	2.16
FOXF2	42.79	5.16
FOXN4	76.81	4.80
FOXO3	323.31	5.49
FOXO6	63.29	1.77
GATA	104.41	2.17
GATA5	89.03	4.38
GFI1B	184.13	11.81
GLIS2	32.14	1.87
HBP1	281.23	3.05
HIC1	101.63	3.27
HOXA13	103.63	3.71
HOXA2	323.31	5.81
INSM2	323.31	2.02
IRF	206.25	2.09
IRF5	39.54	3.78
KLF1	146.64	3.51
LHX8	245.03	2.34
MAF	66.63	3.85
MINI20	102.68	3.29
MYC	39.24	2.64
MZF1	24.10	1.64
NFATC2	37.81	1.92
NFKB1	195.32	11.43
NR0B1	13.24	1.84
NR2C2	137.27	12.28
OTX2	40.38	2.77
P50_P50	131.79	6.49
PITX1	37.22	2.17

TF	Score 1	Score 2
PRDM15	Inf	3.21
RAD21	56.16	4.21
RARB	113.99	2.28
REL	104.66	7.95
RREB	117.64	3.13
RREB1	116.79	3.39
SMC3	90.56	17.59
SOAT1	22.63	1.88
SOX18	113.13	6.71
SRP9	117.49	3.47
TAL1	70.39	1.82
TCF7L2	93.32	8.22
TFCP2	260.99	4.13
THRA	145.85	4.05
TP53	323.31	9.13
TP63	172.00	3.90
YBX1	164.76	5.23
ZBTB7A	66.49	4.74
ZFP14	97.38	5.25
ZNF219	32.59	7.48
ZNF253	323.31	21.30
ZNF585A	185.37	6.09
ZNF616	129.96	8.66
ZNF639	193.43	1.77
ZNF681	323.31	35.04
ZNF701	41.48	1.65
ZNF721	225.60	5.52
ZNF75A	43.99	3.51
ZNF770	66.18	2.95
ZNF786	323.31	15.68
ZNF790	74.59	6.71
ZNF84	149.37	3.38
ZNF846	Inf	10.92

Reverse-forward (RF) orientation

TF	Score 1	Score 2
BCL11B	101.77	5.58
CEBPA	323.31	2.37
CREB3	139.16	13.51
CTCF	119.43	20.80
DBP	25.72	1.84
EGR1	247.74	1.65
EP300	135.93	6.30
ESR1	101.75	9.30
ETS1	250.92	1.76
ETV1	272.94	3.43
ETV3	323.31	2.99
FOS	101.53	9.02
FOXA	323.31	9.16
FOXA1	323.31	8.99
FOXG1	235.13	2.77
FOXN4	61.07	5.39
FOXO3	306.90	5.33
GATA1	314.93	1.91
HLTF	82.82	1.85
HOXA10	29.26	1.81
HOXB6	197.15	2.21
IRF5	Inf	6.92
KLF13	148.08	4.91
KLF8	120.36	2.97
NFE2L2	155.83	3.83
NFKB1	106.02	7.09
NKX2-1	323.31	4.01
NKX2-4	Inf	5.73
NKX2-5	323.31	5.80
NR4A1	116.75	2.99
PPARGC1A	57.39	3.36
RAD21	41.87	3.16
RUNX2	124.18	5.84
SIX6	95.00	3.28
SMARC	323.31	6.02
SRF	323.31	6.78
TAL1	124.35	4.80
TCF12	74.48	2.21
ZBTB18	240.63	9.05
ZBTB4	323.31	2.06

TF	Score 1	Score 2
ZFP2	323.31	6.24
ZNF143	134.24	1.81
ZNF174	159.72	7.65
ZNF257	172.88	3.75
ZNF28	323.31	3.53
ZNF282	100.44	4.02
ZNF300	Inf	3.91
ZNF341	94.91	5.59
ZNF44	165.34	3.45
ZNF484	301.75	12.63
ZNF526	82.28	4.67
ZNF580	55.73	2.98
ZNF625	323.31	2.31
ZNF658	92.95	4.92
ZNF718	165.94	2.09
ZNF721	86.39	7.28
ZNF773	219.31	13.43
ZNF777	137.69	1.96

818

819 **References**

820

821 Bailey SD, Zhang X, Desai K, Aid M, Corradin O, Cowper-Sal Lari R, Akhtar-Zaidi B,
822 Scacheri PC, Haibe-Kains B, Lupien M. 2015. ZNF143 provides sequence
823 specificity to secure chromatin interactions at gene promoters. *Nat Commun* **2**:
824 6186.
825 Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble
826 WS. 2009. MEME SUITE: tools for motif discovery and searching. *Nucleic*
827 *acids research* **37**: W202-208.
828 Barutcu AR, Lajoie BR, Fritz AJ, McCord RP, Nickerson JA, van Wijnen AJ, Lian JB,
829 Stein JL, Dekker J, Stein GS et al. 2016. SMARCA4 regulates gene expression
830 and higher-order chromatin structure in proliferating mammary epithelial cells.
831 *Genome research* **26**: 1188-1201.
832 Bejerano G, Lowe CB, Ahituv N, King B, Siepel A, Salama SR, Rubin EM, Kent WJ,
833 Haussler D. 2006. A distal enhancer and an ultraconserved exon are derived
834 from a novel retroposon. *Nature* **441**: 87-90.
835 Chen M, Manley JL. 2009. Mechanisms of alternative splicing regulation: insights from
836 molecular and genomics approaches. *Nat Rev Mol Cell Biol* **10**: 741-754.
837 Das D, Clark TA, Schweitzer A, Yamamoto M, Marr H, Arribere J, Minovitsky S,
838 Poliakov A, Dubchak I, Blume JE et al. 2007. A correlation with exon
839 expression approach to identify cis-regulatory elements for tissue-specific
840 alternative splicing. *Nucleic acids research* **35**: 4845-4857.
841 de Souza FS, Franchini LF, Rubinstein M. 2013. Exaptation of transposable elements
842 into novel cis-regulatory elements: is the evidence always strong? *Mol Biol Evol*
843 **30**: 1239-1251.
844 de Wit E, Vos ES, Holwerda SJ, Valdes-Quezada C, Verstegen MJ, Teunissen H,
845 Splinter E, Wijchers PJ, Krijger PH, de Laat W. 2015. CTCF Binding Polarity
846 Determines Chromatin Looping. *Molecular cell* **60**: 676-684.
847 Grant CE, Bailey TL, Noble WS. 2011. FIMO: scanning for occurrences of a given
848 motif. *Bioinformatics (Oxford, England)* **27**: 1017-1018.
849 Guo Y, Xu Q, Canzio D, Shou J, Li J, Gorkin DU, Jung I, Wu H, Zhai Y, Tang Y et al.
850 2015. CRISPR Inversion of CTCF Sites Alters Genome Topology and

851 Enhancer/Promoter Function. *Cell* **162**: 900-910.
852 Javierre BM, Burren OS, Wilder SP, Kreuzhuber R, Hill SM, Sewitz S, Cairns J,
853 Wingett SW, Varnai C, Thiecke MJ et al. 2016. Lineage-Specific Genome
854 Architecture Links Enhancers and Non-coding Disease Variants to Target Gene
855 Promoters. *Cell* **167**: 1369-1384 e1319.
856 Jeong M, Huang X, Zhang X, Su J, Shamim M, Bochkov I, Reyes J, Jung H, Heikamp
857 E, Presser Aiden A et al. 2017. A Cell Type-Specific Class of Chromatin Loops
858 Anchored at Large DNA Methylation Nadirs. *bioRxiv*.
859 Ji X, Dadon DB, Abraham BJ, Lee TI, Jaenisch R, Bradner JE, Young RA. 2015.
860 Chromatin proteomic profiling reveals novel proteins associated with
861 histone-marked genomic regions. *Proc Natl Acad Sci U S A* **112**: 3841-3846.
862 Jin C, Kato K, Chimura T, Yamasaki T, Nakade K, Murata T, Li H, Pan J, Zhao M, Sun
863 K et al. 2006. Regulation of histone acetylation and nucleosome assembly by
864 transcription factor JDP2. *Nat Struct Mol Biol* **13**: 331-338.
865 Jjingo D, Conley AB, Wang J, Marino-Ramirez L, Lunyak VV, Jordan IK. 2014.
866 Mammalian-wide interspersed repeat (MIR)-derived enhancers and the
867 regulation of human gene expression. *Mob DNA* **5**: 14.
868 John S, Sabo PJ, Thurman RE, Sung MH, Biddie SC, Johnson TA, Hager GL,
869 Stamatoyannopoulos JA. 2011. Chromatin accessibility pre-determines
870 glucocorticoid receptor binding patterns. *Nature genetics* **43**: 264-268.
871 Jolma A, Yan J, Whitington T, Toivonen J, Nitta KR, Rastas P, Morgunova E, Enge M,
872 Taipale M, Wei G et al. 2013. DNA-binding specificities of human transcription
873 factors. *Cell* **152**: 327-339.
874 Karolchik D, Barber GP, Casper J, Clawson H, Cline MS, Diekhans M, Dreszer TR,
875 Fujita PA, Guruvadoo L, Haeussler M et al. 2014. The UCSC Genome Browser
876 database: 2014 update. *Nucleic acids research* **42**: D764-770.
877 Katainen R, Dave K, Pitkanen E, Palin K, Kivioja T, Valimaki N, Gylfe AE,
878 Ristolainen H, Hanninen UA, Cajuso T et al. 2015. CTCF/cohesin-binding sites
879 are frequently mutated in cancer. *Nature genetics* **47**: 818-821.
880 Kheradpour P, Kellis M. 2014. Systematic discovery and characterization of regulatory
881 motifs in ENCODE TF binding experiments. *Nucleic acids research* **42**:
882 2976-2987.
883 Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler

884 transform. *Bioinformatics (Oxford, England)* **25**: 1754-1760.
885 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G,
886 Durbin R, Genome Project Data Processing S. 2009. The Sequence
887 Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)* **25**:
888 2078-2079.
889 McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM,
890 Bejerano G. 2010. GREAT improves functional interpretation of cis-regulatory
891 regions. *Nature biotechnology* **28**: 495-501.
892 Mumbach MR, Satpathy AT, Boyle EA, Dai C, Gowen BG, Cho SW, Nguyen ML,
893 Rubin AJ, Granja JM, Kazane KR et al. 2017. Enhancer connectome in primary
894 human cells identifies target genes of disease-associated DNA elements. *Nature*
895 *genetics* **49**: 1602-1612.
896 Newburger DE, Bulyk ML. 2009. UniPROBE: an online database of protein binding
897 microarray data on protein-DNA interactions. *Nucleic acids research* **37**:
898 D77-82.
899 Osato N. 2018. Characteristics of functional enrichment and gene expression level of
900 human putative transcriptional target genes. *BMC Genomics* **19**: 957.
901 Phanstiel DH, Van Bortle K, Spacek D, Hess GT, Shamim MS, Machol I, Love MI,
902 Aiden EL, Bassik MC, Snyder MP. 2017. Static and Dynamic DNA Loops form
903 AP-1-Bound Activation Hubs during Macrophage Development. *Molecular cell*
904 **67**: 1037-1048.e1036.
905 Portales-Casamar E, Thongjuea S, Kwon AT, Arenillas D, Zhao X, Valen E, Yusuf D,
906 Lenhard B, Wasserman WW, Sandelin A. 2010. JASPAR 2010: the greatly
907 expanded open-access database of transcription factor binding profiles. *Nucleic*
908 *acids research* **38**: D105-110.
909 Ramani V, Cusanovich DA, Hause RJ, Ma W, Qiu R, Deng X, Blau CA, Disteche CM,
910 Noble WS, Shendure J et al. 2016. Mapping 3D genome architecture through in
911 situ DNase Hi-C. *Nat Protoc* **11**: 2104-2121.
912 Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn
913 AL, Machol I, Omer AD, Lander ES et al. 2014. A 3D map of the human
914 genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**:
915 1665-1680.
916 Rebollo R, Romanish MT, Mager DL. 2012. Transposable elements: an abundant and

917 natural source of regulatory sequences for host genes. *Annu Rev Genet* **46**:
918 21-42.
919 Saeed S, Quintin J, Kerstens HH, Rao NA, Aghajanirefah A, Matarese F, Cheng SC,
920 Ratter J, Berentsen K, van der Ent MA et al. 2014. Epigenetic programming of
921 monocyte-to-macrophage differentiation and trained innate immunity. *Science*
922 *(New York, NY)* **345**: 1251086.
923 Schreiber J, Libbrecht M, Bilmes J, Noble W. 2017. Nucleotide sequence and DNaseI
924 sensitivity are predictive of 3D chromatin architecture. *bioRxiv*.
925 Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N,
926 Schwikowski B, Ideker T. 2003. Cytoscape: a software environment for
927 integrated models of biomolecular interaction networks. *Genome research* **13**:
928 2498-2504.
929 Tabuchi TM, Deplancke B, Osato N, Zhu LJ, Barrasa MI, Harrison MM, Horvitz HR,
930 Walhout AJ, Hagstrom KA. 2011. Chromosome-biased binding and gene
931 regulation by the Caenorhabditis elegans DRM complex. *PLoS Genet* **7**:
932 e1002074.
933 Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF,
934 Schroth GP, Burge CB. 2008. Alternative isoform regulation in human tissue
935 transcriptomes. *Nature* **456**: 470-476.
936 Wang J, Vicente-Garcia C, Seruggia D, Molto E, Fernandez-Minan A, Neto A, Lee E,
937 Gomez-Skarmeta JL, Montoliu L, Lunyak VV et al. 2015. MIR retrotransposon
938 sequences provide insulators to the human genome. *Proc Natl Acad Sci U S A*
939 **112**: E4428-4437.
940 Wang L, Wang S, Li W. 2012. RSeQC: quality control of RNA-seq experiments.
941 *Bioinformatics (Oxford, England)* **28**: 2184-2185.
942 Weintraub AS, Li CH, Zamudio AV, Sigova AA, Hannett NM, Day DS, Abraham BJ,
943 Cohen MA, Nabet B, Buckley DL et al. 2017. YY1 Is a Structural Regulator of
944 Enhancer-Promoter Loops. *Cell* **171**: 1573-1588 e1528.
945 Wingender E, Dietze P, Karas H, Knuppel R. 1996. TRANSFAC: a database on
946 transcription factors and their DNA binding sites. *Nucleic acids research* **24**:
947 238-241.
948 Xie Z, Hu S, Blackshaw S, Zhu H, Qian J. 2010. hPDI: a database of experimental
949 human protein-DNA interactions. *Bioinformatics (Oxford, England)* **26**:

950 287-289.
951 Zhang H, Li F, Jia Y, Xu B, Zhang Y, Li X, Zhang Z. 2017. Characteristic arrangement
952 of nucleosomes is predictive of chromatin interactions at kilobase resolution.
953 *Nucleic acids research* **45**: 12739-12751.
954 Zhang Y, An L, Xu J, Zhang B, Zheng WJ, Hu M, Tang J, Yue F. 2018. Enhancing
955 Hi-C data resolution with deep convolutional neural network HiCPlus. *Nat*
956 *Commun* **9**: 750.
957 Zhao Y, Stormo GD. 2011. Quantitative analysis demonstrates most transcription
958 factors require only simple models of specificity. *Nature biotechnology* **29**:
959 480-483.
960