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

The expression level of the target genes predicted based on EPA was compared with target genes predicted from only promoters. Total 351 biased orientation of DNA motifs [192 forward-reverse (FR) and 179 reverse-forward (RF) orientation, the reverse

complement sequences of some DNA motifs are also registered in databases, so the total number is smaller than the number of FR and RF] affected the expression level of putative transcriptional target genes significantly in $CD14^+$ 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 biased orientation. Transposon sequences, which are known to be involved in insulators and enhancers, showed biased orientation. Moreover, EPI predicted using FR or RF orientation of DNA motif sequences were overlapped with chromatin interaction data (Hi-C and HiChIP) more than other EPA.

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

Background

Chromatin interactions have important roles for enhancer-promoter interactions (EPI) and regulating the transcription of genes. CTCF and cohesin proteins are located at the anchors of chromatin interactions, forming their loop structures. CTCF has insulator function limiting the activity of enhancers into the loops (Fig. 1A). DNA binding sequences of CTCF indicate their orientation bias at chromatin interaction anchors – forward-reverse (FR) orientation is frequently observed (de Wit et al. 2015; Guo et al. 2015). About 40% - 80% of chromatin interaction anchors of Hi-C and in situ

55 Hi-C experiments include DNA binding motif sequences of CTCF. However, it has been
56 reported that long range of chromatin interactions tends to include less CTCF at their
57 anchors (Jeong et al. 2017). Other DNA binding proteins such as ZNF143, YY1, and
58 SMARCA4 (BRG1) are found to be associated with chromatin interactions and EPI
59 (Bailey et al. 2015; Barutcu et al. 2016; Weintraub et al. 2017). CTCF, cohesin, ZNF143,
60 YY1 and SMARCA4 have other biological functions as well as chromatin interactions
61 and enhancer-promoter interactions. The DNA binding motif sequences of the
62 transcription factors are found in open chromatin regions near transcriptional start sites
63 (TSS) as well as chromatin interaction anchors.

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

71 DNA binding motif sequence of YY1 does not seem to be enriched at both
72 chromatin interaction anchors (Z-score < 2), whereas DNA binding motif sequence of
73 ZNF143 is significantly enriched (Z-score > 7; Bailey et al. 2015 Figure 2a). In the
74 analysis of YY1, to identify a protein factor that might contribute to EPI, (Ji et al. 2015)
75 performed chromatin immune precipitation with mass spectrometry (ChIP-MS), using
76 antibodies directed toward histones with modifications characteristic of enhancer and

77 promoter chromatin (H3K27ac and H3K4me3, respectively). Of 26 transcription factors
78 that occupy both enhancers and promoters, four are essential based on a CRISPR
79 cell-essentiality screen and two (CTCF, YY1) are expressed in >90% of tissues
80 examined (Weintraub et al. 2017). These analyses started from the analysis of histone
81 modifications of an enhancer mark rather than chromatin interactions. Other protein
82 factors associated with chromatin interactions may be found from other researches.

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

88 DNA binding proteins involved in chromatin interactions are supposed to affect
89 the transcription of genes in the loop formed by chromatin interactions. In my previous
90 analyses, the expression level of human putative transcriptional target genes was
91 affected, according to the criteria of enhancer-promoter association (EPA) (Fig. 1B;
92 (Osato 2018)). EPI predicted based on EPA shortened at the genomic locations of
93 forward-reverse orientation of CTCF binding sites affected the expression level of
94 putative transcriptional target genes the most, compared with the expression level of
95 transcriptional target genes predicted from only promoters (Fig. 1B). The expression
96 level tended to be increased in monocytes and $CD4^+$ T cells, implying that enhancers
97 activated the transcription of genes, and decreased in ES and iPS cells, implying that
98 enhancers repressed the transcription of genes. These analyses suggested that enhancers

99 affected the transcription of genes significantly, when EPI were predicted properly.
100 Other DNA binding proteins involved in chromatin interactions, as well as CTCF,
101 would locate at chromatin interaction anchors with a pair of biased orientation of DNA
102 binding motif sequences, affecting the expression level of putative transcriptional target
103 genes in the loop formed by the chromatin interactions. As experimental issues of the
104 analyses of chromatin interactions, chromatin interaction data are changed, according to
105 experimental techniques, depth of DNA sequencing, and even replication sets of the
106 same cell type. Chromatin interaction data may not be saturated enough to cover all
107 chromatin interactions. Supposing these characters of DNA binding proteins associated
108 with chromatin interactions and avoiding experimental issues of the analyses of
109 chromatin interactions, here I searched for DNA motif sequences of transcription factors
110 and repeat DNA sequences, affecting EPI and the expression level of putative
111 transcriptional target genes in $CD14^+$ monocytes and $CD4^+$ T cells of four people and
112 other cell types without using chromatin interaction data. Then, putative EPI were
113 compared with chromatin interaction data.
114

115 ## Results

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

117 Transcription factor binding sites (TFBS) were predicted using open chromatin
118 regions and DNA motif sequences of transcription factors collected from various
119 databases and journal papers (see Methods). Transcriptional target genes were predicted
120 using TFBS in promoters and enhancer-promoter association (EPA) shortened at the

121 genomic locations of DNA binding motif sequences of transcription factors such as
122 CTCF and cohesins (RAD21 and SMC3). To find DNA motif sequences of transcription
123 factors, other than CTCF and cohesin, and repeat DNA sequences affecting the
124 expression level of putative transcriptional target genes, transcriptional target genes of
125 each transcription factor were predicted based on EPA shortened at the genomic
126 locations of DNA motif sequences, and the expression level of the putative
127 transcriptional target genes was compared with those predicted from promoters using
128 Mann Whiteney U test (p-value < 0.05). The number of transcription factors showing a
129 significant difference of expression level of putative transcriptional target genes was
130 counted. To examine whether the orientation of DNA motif sequences shortening EPA
131 affected the number of transcription factors showing a significant difference of
132 expression level of putative transcriptional target genes, the number of the transcription
133 factors was compared among forward-reverse, reverse-forward, and any orientation of
134 DNA motif sequences shortening EPA, using chi-square test (p-value < 0.05). To avoid
135 missing DNA motif sequences showing a relatively weak statistical significance by
136 multiple testing collection, the above analyses were conducted using monocytes of four
137 people independently, and DNA motif sequences found in monocytes of four people in
138 common were selected. Total 351 of biased (192 forward-reverse and 179
139 reverse-forward) orientation of DNA binding motif sequences of transcription factors
140 have been found in monocytes of four people in common, whereas only one any
141 orientation (i.e. without considering orientation) of DNA binding motif sequences have
142 been found in monocytes of four people in common (Fig. 2; Table 1; Supplemental

143 Table S1). Forward-reverse orientation of DNA motif sequences included CTCF,
144 cohesin (RAD21 and SMC3), ZNF143 and YY1, which are associated with chromatin
145 interactions and enhancer-promoter interactions (EPI). SMARCA4 (BRG1) is
146 associated with topologically associated domain (TAD), which is higher-order
147 chromatin organization. The DNA binding motif sequences of SMARCA4 was not
148 registered in the databases used in this study. Forward-reverse orientation of DNA motif
149 sequences included SMRC2, a member of the same SWI/SNF family of proteins as
150 SMARCA4.

151 The same analysis was conducted using DNase-seq data of $CD4^+$ T cells of four
152 people. Total 322 of biased (186 forward-reverse and 166 reverse-forward) orientation
153 of DNA binding motif sequences of transcription factors have been found in T cells of
154 four people in common, whereas only five any orientation (i.e. without considering
155 orientation) of DNA binding motif sequences have been found in T cells of four people
156 in common (Supplemental Table S2). Biased orientation of DNA motif sequences in T
157 cells included CTCF, cohesion (RAD21 and SMC3), YY1, and SMARC. Biased
158 orientation of DNA motif sequences in T cells included JUNDM2 (JDP2), which is
159 involved in histone-chaperone activity, promoting nucleosome, and inhibition of histone
160 acetylation (Jin et al. 2006). JDP2 forms a homodimer or heterodimer with various
161 transcription factors (https://en.wikipedia.org/wiki/Jun_dimerization_protein). Among
162 322, 53 of biased orientation of DNA binding motif sequences of transcription factors
163 were found in both monocytes and T cells in common (Supplemental Table S3). For
164 each orientation, 27 forward-reverse and 29 reverse-forward orientation of DNA

165 binding motif sequences of transcription factors were found in both monocytes and T
166 cells in common. Without considering the difference of orientation of DNA binding
167 motif sequences, 79 of biased orientation of DNA binding motif sequences of
168 transcription factors were found in both monocytes and T cells. As a reason for the
169 increase of the number from 53, a transcription factor or an isoform of the same
170 transcription factor may bind to a different DNA binding motif sequence according to
171 cell types and/or in the same cell type. The same transcription factor has several DNA
172 binding motif sequences and in some cases one of the motif sequence is almost the same
173 as the reverse complement sequence of another motif sequence of the same transcription
174 factor. I previously found that a complex of transcription factors would bind to a
175 slightly different DNA binding motif sequence from the combination of DNA binding
176 motif sequences of transcription factors consisting of the complex in *C. elegans*
177 (Tabuchi et al. 2011). From another viewpoint of this study, the expression level of
178 putative transcription target genes of some transcription factors would be different,
179 depending on the genomic locations (enhancers or promoters) of DNA binding motif
180 sequences of the transcription factors in monocytes and T cells of four people.

181 To examine whether the biased orientation of DNA binding motif sequences of
182 transcription factors was observed in other cell types, the same analyses were conducted
183 in other cell types. However, for other cell types, experimental data of one sample were
184 available in ENCODE database, so the analyses of DNA motif sequences were
185 performed by comparing with the result in monocytes of four people. Among the biased
186 orientation of DNA binding motif sequences found in monocytes, 79, 76, 91, 61, and 59

187 DNA binding motif sequences were also observed in $CD4^+$ T cells, H1-hESC, iPS,
188 Huvec and MCF-7 respectively, including CTCF and cohesin (RAD21 and SMC3)
189 (Table 2; Supplemental Table S4). The scores of DNA binding motif sequences were the
190 highest in monocytes, and the other cell types showed lower scores. The results of the
191 analysis of DNA motif sequences in $CD20^+$ B cells and macrophages did not include
192 CTCF and cohesin, because these analyses can be utilized in cells where the expression
193 level of putative transcriptional target genes of each transcription factor show a
194 significant difference between promoters and EPA shortened at the genomic locations of
195 DNA motif sequences. Some experimental data of a cell did not show a significant
196 difference between promoters and EPA (Osato 2018).

197 Instead of DNA binding motif sequences of transcription factors, repeat DNA
198 sequences were also examined. The expression level of transcriptional target genes
199 predicted based on EPA shortened at the genomic locations of repeat DNA sequences
200 was compared with the expression level of transcriptional target genes predicted from
201 promoters. Three reverse-forward orientation of repeat DNA sequences showed a
202 significant difference of expression level of putative transcriptional target genes in
203 monocytes of four people in common (Table 3). Among them, LTR16C repeat DNA
204 sequence was observed in iPS and H1-hESC with enough statistical significance
205 considering multiple tests (p-value $< 10^{-7}$). The same as $CD14^+$ monocytes, biased
206 orientation of repeat DNA sequences was examined in $CD4^+$ T cells. Three
207 forward-reverse and two reverse-forward orientation of repeat DNA sequences showed
208 a significant difference of expression level of putative transcriptional target genes in T

209 cells of four people in common (Supplemental Table S5). MIR and other transposon
210 sequences are known to be involved in insulators and enhancers (Bejerano et al. 2006;
211 Rebollo et al. 2012; de Souza et al. 2013; Jjingo et al. 2014; Wang et al. 2015).
212
213 **Co-location of biased orientation of DNA motif sequences**
214 To examine association of 351 biased (forward-reverse and reverse-forward)
215 orientation of DNA binding motif sequences of transcription factors, co-location of the
216 DNA binding motif sequences in open chromatin regions in monocytes was analyzed.
217 The number of open chromatin regions with the same pairs of DNA binding motif
218 sequences was counted, and when the pairs of DNA binding motif sequences were
219 enriched with statistical significance (chi-square test, p-value $< 1.0 \times 10^{-10}$), they were
220 listed (Table 4; Supplemental Table S6). Open chromatin regions overlapped with
221 histone modification of an enhancer mark (H3K27ac) (total 26,095 regions) showed a
222 larger number of enriched pairs of DNA motifs than all open chromatin regions (Table
223 4; Supplemental Table S6). H3K27ac is known to be enriched at chromatin interaction
224 anchors (Phanstiel et al. 2017). As already known, CTCF was found with cohesin such
225 as RAD21 and SMC3 (Table 4). Top 30 pairs of FR and RF orientations of DNA motifs
226 co-occupied in the same open chromatin regions were shown (Table 4). Total number of
227 pairs of DNA motifs was 569, consisting of 137 unique DNA motifs, when the pair of
228 DNA motifs were observed in more than 80% of the number of open chromatin regions
229 with the DNA motifs.
230 The same as $CD14^+$ monocytes, to examine association of 322 biased orientation

231 of DNA binding motif sequences of transcription factors, co-location of the DNA
232 binding motif sequences in open chromatin regions in $CD4^+$ T cells was analyzed. Top
233 30 pairs of FR and RF orientations of DNA motifs co-occupied in the same open
234 chromatin regions were shown (Supplemental Table S7). Total number of pairs of DNA
235 motifs was 458, consisting of 150 unique DNA motifs, when the pair of DNA motifs
236 were observed in more than 60% of the number of open chromatin regions with the
237 DNA motifs (chi-square test, p-value $< 1.0 \times 10^{-10}$). Among them, 13 pairs of DNA
238 motif sequences including a pair of CTCF, SMC3, and RAD21 in T cells were found in
239 monocytes in common (Supplemental Table S7).
240
241 **Comparison with chromatin interaction data**
242 To examine whether the biased orientation of DNA motif sequences is associated
243 with chromatin interactions, enhancer-promoter interactions (EPI) predicted based on
244 enhancer-promoter associations (EPA) were compared with chromatin interaction data
245 (Hi-C). Due to the resolution of Hi-C experimental data used in this study (50kb), EPI
246 were adjusted to 50kb resolution. EPI were predicted based on three types of EPA: (i)
247 EPA without being shortened by DNA motif sequences, (ii) EPA shortened at the
248 genomic locations of DNA motif sequences of transcription factors such as CTCF and
249 cohesin (RAD21, SMC3) without considering their orientation, and (iii) EPA shortened
250 at the genomic locations of forward-reverse or reverse-forward orientation of DNA
251 motif sequences of transcription factors. EPA (iii) showed the highest ratio of EPI
252 overlapped with chromatin interactions (Hi-C) using DNA binding motif sequences of

253 CTCF and cohesin (RAD21 and SMC3), compared with the other two types of EPA
254 (Fig. 3). Total 62 biased orientation (41 FR and 24 RF) of DNA motif sequences
255 including CTCF and cohesin showed a higher ratio of EPI overlapped with Hi-C than
256 the other types of EPA in monocytes (Table 5). Chromatin interaction data were
257 obtained using different samples from DNase-seq, open chromatin regions, so
258 individual differences seemed to be large from the results of this analysis. Since, for
259 some DNA motif sequences of transcription factors, the number of EPI overlapped with
260 chromatin interactions was small, if higher resolution of chromatin interaction data
261 (such as HiChIP, in situ DNase Hi-C, and in situ Hi-C data, or a tool to improve the
262 resolution such as HiCPlus) is available, the number of EPI overlapped with chromatin
263 interactions would be increased and the difference of the numbers among three types of
264 EPA might be larger and more significant than the current results (Rao et al. 2014;
265 Ramani et al. 2016; Mumbach et al. 2017; Zhang et al. 2018).

266 After the analysis of $CD14^+$ monocytes, to utilize HiChIP chromatin interaction
267 data in $CD4^+$ T cells, the same analysis for $CD14^+$ monocytes was performed using
268 DNase-seq data of four donors in $CD4^+$ T cells (Mumbach et al. 2017). EPI predicted
269 based on EPA were compared with HiChIP chromatin interaction data in $CD4^+$ T cells.
270 The resolutions of HiChIP chromatin interaction data and EPI were adjusted to 1kb. EPI
271 were predicted based on the three types of EPA in the same way as $CD14^+$ monocytes.
272 EPA (iii) showed the highest ratio of EPI overlapped with chromatin interactions
273 (HiChIP) using DNA binding motif sequences of CTCF and cohesin (RAD21 and
274 SMC3), compared with the other two types of EPA (Supplemental Fig. S2). Total 150

275 biased orientation (92 FR and 66 RF) of DNA motif sequences including CTCF and
276 cohesin showed a higher ratio of EPI overlapped with HiChIP than the other types of
277 EPA in monocytes (Supplemental Table S8). As expected, the number of EPI
278 overlapped with chromatin interactions (HiChIP) was increased, compared with Hi-C
279 chromatin interactions. However, the same as $CD14^+$ monocytes, chromatin interaction
280 data were obtained using different samples from DNase-seq, open chromatin regions in
281 $CD4^+$ T cells, so individual differences seemed to be large from the results of this
282 analysis, and (Mumbach et al. 2017) suggested that individual differences of chromatin
283 interactions were larger than those of open chromatin regions. ATAC-seq data, open
284 chromatin regions were available in $CD4^+$ T cells in the paper, however, when using
285 ATAC-seq data, the result of the analysis of biased orientation of DNA motif sequences
286 was different from DNase-seq data, and not included a part of CTCF and cohesin. Thus,
287 DNase-seq data collected from ENCODE and Blueprint projects were employed in this
288 study.
289

290 **Discussion**

291 To find DNA motif sequences of transcription factors and repeat DNA sequences
292 affecting human putative transcriptional target genes, DNA motif sequences were
293 searched from open chromatin regions of monocytes of four people. Total 351 biased
294 (192 forward-reverse and 179 reverse-forward) orientation of DNA motif sequences
295 were found in monocytes of four people in common, whereas only one any orientation
296 (i.e. without considering orientation) of DNA motif sequence was found to affect

297 putative transcriptional target genes, suggesting that EPA shortened at the genomic
298 locations of forward-reverse or reverse-forward orientation of DNA motif sequences of
299 transcription factors is an important character for EPI and transcription of genes.

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

319 0.00000625. Actually, DNA motif sequences with *p*-value slightly less than 0.05 in
320 monocytes of one person were observed in monocytes of four people in common.

321 EPI were compared with chromatin interactions (Hi-C) in monocytes. EPA
322 shortened at the genomic locations of DNA binding motif sequences of CTCF and
323 cohesin (RAD21 and SMC3) showed a significant difference of the ratios of EPI
324 overlapped with chromatin interactions, according to three types of EPA (FR, any, and
325 others). Using open chromatin regions overlapped with ChIP-seq experimental data of
326 histone modification of an enhancer mark (H3K27ac), the ratio of EPI not overlapped
327 with Hi-C was reduced and the difference of the ratio of EPI overlapped with chromatin
328 interactions among the three types of EPA became larger (data not shown). (Phanstiel et
329 al. 2017) also reported that there was an especially strong enrichment for loops with
330 H3K27 acetylation peaks at both ends (Fisher's Exact Test, $p = 1.4 \times 10^{-27}$). However,
331 the total number of EPI overlapped with chromatin interactions was also reduced using
332 H3K27ac peaks, so more chromatin interaction data would be needed to obtain reliable
333 results in this study. As an issue of experimental data, chromatin interaction data were
334 obtained using different samples from DNase-seq, open chromatin regions, so
335 individual differences were not reflected to chromatin interaction data. Moreover, the
336 resolution of chromatin interaction data used in monocytes is about 50kb, thus the
337 number of chromatin interactions was relatively small (72,284 at 50kb resolution with
338 cutoff score of CHiCAGO tool > 1). EPA shortened at the genomic locations of DNA
339 binding motif sequences of transcription factors found in various open chromatin
340 regions such as CTCF and cohesin (RAD21 and SMC3) tend to be overlapped with a

341 larger number of chromatin interactions than DNA binding motif sequences of
342 transcription factors less frequently observed in open chromatin regions. Therefore, to
343 examine the difference of the numbers of EPI overlapped with chromatin interactions,
344 according to three types of EPA (FR or RF, any, and others), the number of chromatin
345 interactions should be large enough.

346 As HiChIP chromatin interaction data were available in $CD4^+$ T cells, biased
347 orientations of DNA motif sequences of transcription factors were examined in T cells
348 using DNase-seq data of four people. The resolutions of chromatin interactions and EPI
349 were adjusted to 1kb by fragmentation of genome sequences. In monocytes, the
350 resolution of Hi-C chromatin interaction data was converted by extending anchor
351 regions of chromatin interactions to 50kb length and merging chromatin interactions
352 overlapped with each other. Fragmentation of genome sequences may affect the
353 classification of chromatin interactions of which anchors are located near the border of a
354 fragment, but the number of chromatin interactions would not be decreased, compared
355 with merging chromatin interactions. The number of HiChIP chromatin interactions was
356 28,918,687 at 1kb resolution. As expected, the number of EPI overlapped with
357 chromatin interactions was increased, and about half of biased orientation of DNA motif
358 sequences of transcription factors showed a statistical significance in EPI predicted
359 based on EPA shortened at the genomic locations of the DNA motif sequences,
360 compared with the other types of EPA. But the difference of the ratios of EPI
361 overlapped with chromatin interactions, according to three types of EPA (FR or RF, any,
362 and others), became smaller in T cells than monocytes (Supplemental Fig. S2), due to

363 the increase of chromatin interactions. False positive predictions of EPI would be
364 decreased by using H3K27ac marks and other features, as suggested in the analysis in
365 monocytes.

366 Though the number of known DNA binding proteins and transcription factors
367 involved in chromatin interactions would be less than 20 or so, about 300 DNA binding
368 proteins showed enrichment of biased orientation of their DNA binding sequences, and
369 known DNA binding proteins involved in chromatin interactions were found in the
370 biased orientation of DNA binding sequences. JUNDM2 (JDP2) found in T cells is not
371 known DNA binding proteins involved in chromatin interactions. JDP2 is involved in
372 histone and nucleosome-related functions such as histone-chaperone activity, promoting
373 nucleosome, and inhibition of histone acetylation (Jin et al. 2006). JDP2 forms a
374 homodimer or heterodimer with various transcription factors
375 (https://en.wikipedia.org/wiki/Jun_dimerization_protein). Thus, the main function of
376 JDP2 may not be associated with chromatin interactions, but it forms a homodimer or
377 heterodimer with transcription factors, potentially forming a chromatin interaction
378 through the binding of JDP2 and another transcription factor, like chromatin interactions
379 formed by YY1 transcription factor (Weintraub et al. 2017). When forming a
380 homodimer or heterodimer with another transcription factor, transcription factors may
381 bind to genome DNA with a specific orientation of their DNA binding sequences. From
382 the analysis of biased orientation of DNA motif sequences of transcription factors, such
383 transcription factors would also be found. If the DNA binding motif sequence of only
384 the mate to a pair of transcription factors was found in EPA, EPA was shortened at one

385 side, the genomic location of the DNA binding motif sequence of the mate to the pair,
386 and transcriptional target genes were predicted using the EPA shortened at one side. In
387 this analysis, the mate to both heterodimer and homodimer of transcription factors can
388 be used to examine the effect on the expression level of transcriptional target genes
389 predicted based on the EPA shortened at one side.

390 I reported that the ratio of the number of functional enrichment of putative
391 transcriptional target genes in the total number putative transcriptional target genes
392 changed according to the types of EPA, in the same way as gene expression level (Osato
393 2018). JDP2 showed a higher ratio of functional enrichment of putative transcriptional
394 target genes using EPA shortened at genomic locations of forward-reverse orientation of
395 the DNA binding motif sequence, compared with the other types of EPA
396 (reverse-forward or any orientation) in monocytes and T cells, among monocytes, T
397 cells and $CD20^+$ B cells. While the genome-wide analysis of functional enrichments of
398 putative transcriptional target genes takes computational time, the genome-wide
399 analysis of gene expression level is finished with less computational time. The analysis
400 of gene expression level can be used for a cell type and experimental data where the
401 expression levels of transcriptional target genes predicted from EPA are significantly
402 different from those predicted from only promoter.

403 The analyses in this study revealed novel characters of DNA binding motif
404 sequences of transcription factors and repeat DNA sequences, and would be useful to
405 analyze transcription factors involved in chromatin interactions and/or forming a
406 homodimer, heterodimer or complex with other transcription factors, affecting the

407 expression and transcriptional regulation of genes.

408

409 ## Methods

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

411 To examine transcriptional regulatory target genes, bed files of hg38 of Blueprint
412 DNase-seq data for $CD14^+$ monocytes of four donors (EGAD00001002286; Donor ID:
413 C0010K, C0011I, C001UY, C005PS) were obtained from Blueprint project web site
414 (http://dcc.blueprint-epigenome.eu/#/home), and the bed files of hg38 were converted
415 into those of hg19 using Batch Coordinate Conversion (liftOver) web site
416 (https://genome.ucsc.edu/util.html). Bed files of hg19 of ENCODE
417 $CD4^+$_Naive_Wb11970640 (GSM1014537; UCSC Accession: wgEncodeEH003156),
418 H1-hESC (GSM816632; UCSC Accession: wgEncodeEH000556), iPSC (GSM816642;
419 UCSC Accession: wgEncodeEH001110), HUVEC (GSM1014528; UCSC Accession:
420 wgEncodeEH002460), and MCF-7 (GSM816627; UCSC Accession:
421 wgEncodeEH000579) were obtained from the ENCODE websites
422 (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeUwDgf/;
423 http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeUwDnase/).

424 As high resolution of chromatin interaction data using HiChIP became available
425 in $CD4^+$ T cells, to promote the same analysis of $CD14^+$ monocytes in $CD4^+$ T cells,
426 DNase-seq data of four donors were obtained from ENCODE and Blueprint projects
427 web sites. DNase-seq data of only one donor was available in Blueprint project for
428 $CD4^+$ T cells (Donor ID: S008H1), and three other DNase-seq data were obtained from

429 ENCODE project. Though the peak calling of DNase-seq data available in ENCODE
430 database was different from other DNase-seq data in ENCODE and Blueprint projects
431 where 150bp length of peaks were usually predicted using HotSpot (John et al. 2011),
432 FASTQ files of DNase-seq data were downloaded from NCBI Sequence Read Archive
433 (SRA) database (SRR097566, SRR097618, and SRR171574). Read sequences of
434 FASTQ files were aligned to the hg19 version of the human genome reference using
435 BWA (Li and Durbin 2009), and the BAM files generated by BWA were converted into
436 SAM files, sorted, and indexed using Samtools (Li et al. 2009). Peaks of the DNase-seq
437 data were predicted using HotSpot-4.1.1.

438 To identify transcription factor binding sites (TFBS) from the DNase-seq data,
439 TRANSFAC (2013.2), JASPAR (2010), UniPROBE, BEEML-PBM, high-throughput
440 SELEX, Human Protein-DNA Interactome, and transcription factor binding sequences
441 of ENCODE ChIP-seq data were used (Wingender et al. 1996; Newburger and Bulyk
442 2009; Portales-Casamar et al. 2010; Xie et al. 2010; Zhao and Stormo 2011; Jolma et al.
443 2013; Kheradpour and Kellis 2014). Position weight matrices of transcription factor
444 binding sequences were transformed into TRANSFAC matrices and then into MEME
445 matrices using in-house scripts and transfac2meme in MEME suite (Bailey et al. 2009).
446 Transcription factor binding sequences of transcription factors derived from vertebrates
447 were used for further analyses. Searches were conducted for transcription factor binding
448 sequences from each narrow peak using FIMO with *p*-value threshold of 10^{-5} (Grant et
449 al. 2011). Transcription factors corresponding to transcription factor binding sequences
450 were searched computationally by comparing their names and gene symbols of HGNC

451 (HUGO Gene Nomenclature Committee) -approved gene nomenclature and 31,848
452 UCSC known canonical transcripts
453 (http://hgdownload.cse.ucsc.edu/goldenpath/hg19/database/knownCanonical.txt.gz), as
454 transcription factor binding sequences were not linked to transcript IDs such as UCSC,
455 RefSeq, and Ensembl transcripts.

456 Target genes of a transcription factor were assigned when its TFBS was found in
457 DNase-seq narrow peaks in promoter or extended regions for enhancer-promoter
458 association of genes (EPA). Promoter and extended regions were defined as follows:
459 promoter regions were those that were within distances of ±5 kb from transcriptional
460 start sites (TSS). Promoter and extended regions were defined as per the following
461 association rule, which is the same as that defined in Figure 3A of a previous study
462 (McLean et al. 2010): the single nearest gene association rule, which extends the
463 regulatory domain to the midpoint between the TSS of the gene and that of the nearest
464 gene upstream and downstream without the limitation of extension length. Extended
465 regions for EPA were shortened at the genomic locations of DNA binding sites of
466 transcription factors that were the closest to a transcriptional start site, and
467 transcriptional target genes were predicted from the shortened enhancer regions using
468 TFBS. Furthermore, promoter and extended regions for EPA were shortened at the
469 genomic locations of forward–reverse orientation of DNA binding sites of transcription
470 factors. When forward or reverse orientation of DNA binding sites were continuously
471 located in genome sequences several times, the most external forward–reverse
472 orientation of DNA binding sites were selected. The genomic positions of genes were

473 identified using ‘knownGene.txt.gz’ file in UCSC bioinformatics sites (Karolchik et al.
474 2014). The file ‘knownCanonical.txt.gz’ was also utilized for choosing representative
475 transcripts among various alternate forms for assigning promoter and extended regions
476 for EPA. From the list of transcription factor binding sequences and transcriptional
477 target genes, redundant transcription factor binding sequences were removed by
478 comparing the target genes of a transcription factor binding sequence and its
479 corresponding transcription factor; if identical, one of the transcription factor binding
480 sequences was used. When the number of transcriptional target genes predicted from a
481 transcription factor binding sequence was less than five, the transcription factor binding
482 sequence was omitted.

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

487 For gene expression data, RNA-seq reads mapped onto human hg19 genome
488 sequences were obtained, including ENCODE long RNA-seq reads with poly-A of
489 H1-hESC, iPSC, HUVEC, and MCF-7 (GSM26284, GSM958733, GSM2344099,
490 GSM2344100, GSM958734, and GSM765388), and UCSF-UBC human reference
491 epigenome mapping project RNA-seq reads with poly-A of naive $CD4^+$ T cells
492 (GSM669617). Two replicates were present for H1-hESC, iPSC, HUVEC, and MCF-7,
493 and a single one for $CD4^+$ T cells. RPKMs of the RNA-seq data were calculated using
494 RSeQC (Wang et al. 2012). For monocytes, Blueprint RNA-seq RPKM data

495 (GSE58310, GSE58310_GeneExpression.csv.gz, Monocytes_Day0_RPMI) was used
496 (Saeed et al. 2014). Based on RPKM, UCSC transcripts with expression levels among
497 top 30% of all the transcripts were selected in each cell type.

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

514 Though forward-reverse orientation of DNA binding motif sequences of CTCF
515 and cohesin is frequently observed at chromatin interaction anchors, the percentage of
516 forward-reverse orientation is not 100, and other orientations of the DNA binding motif

517 sequences are also observed. Though DNA binding motif sequences of CTCF and
518 cohesin are found in various open chromatin regions, DNA binding motif sequences of
519 some transcription factors would be observed less frequently in open chromatin regions.
520 The analyses of cells of a number of people would avoid missing relatively weak
521 statistical significance of DNA motif sequences in cells of each person by multiple
522 testing correction of thousands of statistical tests. A DNA motif sequence was found
523 with p-value < 0.05 in cells of one person and the DNA motif sequence found in the
524 same cell type of four people in common would have p-value < $0.05^4 = 0.00000625$.
525

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

527 Co-location of biased orientation of DNA binding motif sequences of
528 transcription factors was examined. The number of open chromatin regions with the
529 same pairs of DNA binding motif sequences was counted, and when the pairs of DNA
530 binding motif sequences were enriched with statistical significance (chi-square test,
531 p-value < 1.0×10^{-10}), they were listed.

532 For histone modification of an enhancer mark (H3K27ac), bed files of hg38 of
533 Blueprint ChIP-seq data for $CD14^+$ monocytes (EGAD00001001179) and $CD4^+$ T cells
534 (Donor ID: S000RD) was obtained from Blueprint web site
535 (http://dcc.blueprint-epigenome.eu/#/home), and the bed files of hg38 were converted
536 into those of hg19 using Batch Coordinate Conversion (liftOver) web site
537 (https://genome.ucsc.edu/util.html).
538

539 **Comparison with chromatin interaction data**

540 For comparison of EPA in monocytes with chromatin interactions,
541 'PCHiC_peak_matrix_cutoff0.txt.gz' file was downloaded from 'Promoter Capture
542 Hi-C in 17 human primary blood cell types' web site (https://osf.io/u8tzp/files/), and
543 chromatin interactions for Monocytes with scores of CHiCAGO tool > 1 were extracted
544 from the file (Javierre et al. 2016).

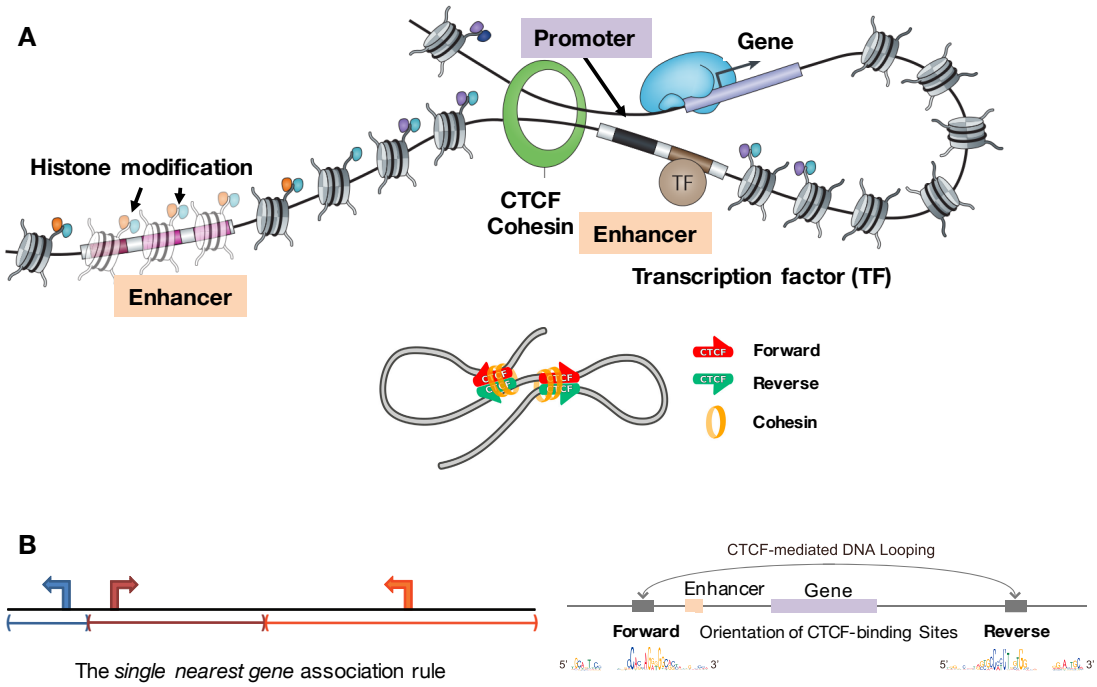
545 EPI were predicted using three types of EPA in monocytes: (i) EPA shortened at
546 the genomic locations of forward-reverse or reverse-forward orientation of DNA motif
547 sequences of transcription factors, (ii) EPA shortened at the genomic locations of any
548 orientation (i.e. without considering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

561 For comparison of EPA with chromatin interactions (HiChIP) in $CD4^+$ T cells,
562 'GSM2705049_Naive_HiChIP_H3K27ac_B2T1_allValidPairs.txt' file was downloaded
563 from Gene Expression Omnibus (GEO) database (GSM2705049). The resolutions of
564 chromatin interaction data and EPI were adjusted to 1kb before their comparison.
565

566 **Acknowledgements**

567 The supercomputing resource was provided by Human Genome Center of the Institute
568 of Medical Science at the University of Tokyo. Computations were partially performed
569 on the NIG supercomputer at ROIS National Institute of Genetics. Publication charges
570 for this article were funded by JSPS KAKENHI Grant Number 16K00387. This
571 research was partially supported by the Platform Project for Supporting in Drug
572 Discovery and Life Science Research（Platform for Dynamic Approaches to Living
573 System）from Japan Agency for Medical Research and Development (AMED). This
574 research was partially supported by Development of Fundamental Technologies for
575 Diagnosis and Therapy Based upon Epigenome Analysis from Japan Agency for
576 Medical Research and Development (AMED). This work was partially supported by
577 JST CREST Grant Number JPMJCR15G1, Japan.
578
579
580
581
582
583
584

585 **Figures**



586

587 **Figure 1.** Chromatin interactions and enhancer-promoter association. (A) Forward–
588 reverse orientation of CTCF-binding sites are frequently found in chromatin interaction
589 anchors. CTCF can block the interaction between enhancers and promoters limiting the
590 activity of enhancers to certain functional domains (de Wit et al. 2015; Guo et al. 2015).
591 (B) Computationally-defined regulatory domains for enhancer-promoter association
592 (McLean et al. 2010). The *single nearest gene* association rule extends the regulatory
593 domain to the midpoint between this gene's TSS and the nearest gene's TSS both
594 upstream and downstream. Enhancer-promoter association was shortened at the
595 genomic locations of forward-reverse orientation of DNA binding motif sequences of
596 transcription factors (e.g. CTCF in the figure).

597

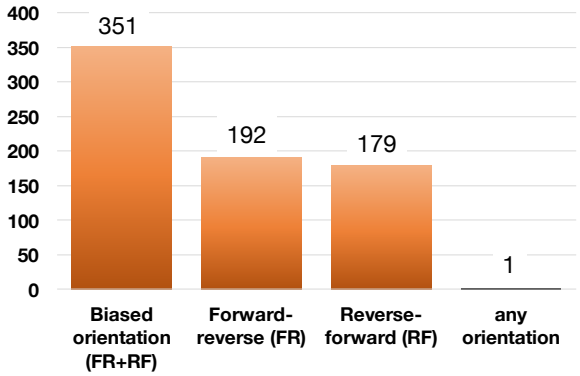


Figure 2. Biased-orientation of DNA motif sequences of transcription factors in monocytes. Total 351 of biased orientation of DNA binding motif sequences of transcription factors have been found to affect the expression level of putative transcriptional target genes in monocytes of four people in common, whereas only one any orientation (i.e. without considering orientation) of DNA binding motif sequences have been found to affect the expression level.

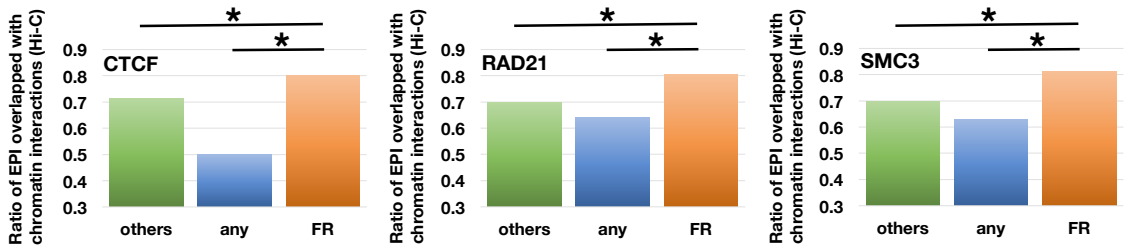


Figure 3. Comparison of DNA binding motif sequences of transcription factors with chromatin interactions in monocytes. Total 62 biased orientation (41 FR and 24 RF) of DNA motif sequences including CTCF and cohesin showed a higher ratio of EPI overlapped with Hi-C than the other types of EPA in monocytes.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

Tables

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

Forward-reverse orientation

TF	Score	TF	Score	TF	Score
SRF	136.39	EBF	55.63	EGR2	37.31
IRF7	119.13	P53	54.99	YLL054C	37.20
CLAMP	107.65	ZNF92	54.69	ZNF316	37.15
STAT1	105.40	E47	52.16	RSRFC4	36.96
TFE3	94.78	REST	52.05	PLAGL1	36.91
ZNF623	90.72	EIF5A2	51.51	PACC	36.76
CEBPG	90.30	KLF14	51.36	YBR033W	36.32
ZNF317	88.08	TP63	50.43	STA5A	35.28
ZNF93	87.64	MAFA	50.06	TCF12	35.23
SMAD2_SMAD3_SMAD4	85.65	IKZF1	48.58	RAD21	35.02
ZNF836	81.83	RDR1	48.04	ZNF573	34.69
E2F	81.67	SMC3	47.07	EHF	34.43
FOXA1	75.73	ESR2	47.03	ZIC1	34.07
MINI20	75.11	YB1	46.87	GCM1	34.00
HMBOX1	72.11	HIC1	46.40	GLI2	33.53
USF	70.65	NR1I2	46.29	RXRA_VDR	33.46
ZNF660	68.80	ZNF648	45.77	GATA2	32.95
STAT4	66.19	ZN219	43.55	TCFAP2C	32.89
HXB1	63.60	HME1	43.47	HNF4	32.83
ZNF709	63.52	NR2F6	42.34	DEC	32.83
CTCF	62.26	ZNF33B	41.81	PF13	32.82
FOXN2	61.95	DREB2E	41.72	BTEB2	32.60
D	59.93	NR2F2	41.24	E2F8	32.38
STA5B	58.88	ZNF214	41.22	AT3G16280	31.88
CMYB	58.66	CEBPZ	40.56	GLI1	31.75
ZNF646	57.77	ZNF695	39.37	ZNF460	31.58
MYB	57.17	AT3G04100	39.18	GAL4P	31.56
ZNF716	56.47	PTF1A	38.20	CPRF1	31.25
ZNF676	56.30	SMRC2	37.70	STAT3	31.17
KR	55.78	GCMA	37.46	WT1	31.03

Reverse-forward orientation

TF	Score	TF	Score	TF	Score
RFX5	155.33	NR2F1	54.18	TCFAP2E	42.44
ZNF286B	100.66	ISL1	53.90	BDP1	42.37
ZNF225	97.70	NFAC1	53.35	AHR	42.36
CEBPB	86.76	RDS2	52.88	TATA	42.27
SRF	85.93	NRF2	52.31	ZNF449	42.05
ZFP579	80.23	RFX3	51.72	SP8	41.63
TCF1	79.47	TCFAP2B	51.52	ZIC2	41.57
SOX9	76.41	ETV7	51.20	ZFP202	41.27
IRF3	75.35	KLF16	50.51	CRZ1P	40.83
ZNF28	71.26	NFKB2	50.12	VND	40.72
NFY	69.96	ZNF195	49.93	ZNF652	40.55
PAX5	69.88	HAND1_E47	49.09	CRZ1	40.18
WT1	69.28	ETV6	47.96	MYOG	39.93
STAT1	69.06	MAFK	47.86	ASCL2	39.92
TOE3	68.44	TBP	47.63	HSF2	39.69
ZNF670	65.07	NR5A2	47.55	DOBOX5	39.39
NANOG	64.97	LEU3	47.42	SP3	38.95
SP1	64.05	SMAD3	47.22	PO3F2	38.74
TCF2	63.84	ZNF331	46.93	FLI1	38.72
TEAD1	62.71	ZNF238	46.86	RXRA	38.27
NRF3	62.47	ZSCAN2	46.77	GABP1_GABP2	38.20
NKX3	62.30	CMYC_MAX	45.48	ZNF711	37.70
ZBRK1	62.26	HSF1	44.61	SOX2	37.44
STAT5B	61.91	ZBTB2	44.47	RPH1	37.12
CTCF	60.90	RAP1	44.47	RBAK	37.10
PARP	57.56	CREM	44.31	ZN337	36.91
STAT5A	55.66	ZNF343	44.03	ZBTB7B	36.73
ZNF30	55.64	ZBTB18	42.76	ETV3	36.60
E74A	55.05	OBOX2	42.53	PO2F2	36.45
ZNF687	54.64	RNF96	42.48	FOX	36.01

648 **Table 2.** Top 30 forward-reverse orientation of DNA binding motif sequences of

649 transcription factors in MCF-7 and iPS cells. TF: DNA binding motif sequences of

650 transcription factors. Score: $-\log_{10}$ (*p*-value).

651 MCF-7 iPS

TF	Score
CTCF	43.69
RAD21	27.00
SMC3	26.64
HOXA7	14.14
E2F	10.69
CAC	9.91
DREB2E	9.79
BM9	9.67
REST	9.22
STAT1	8.12
SIRT6	8.03
STA5B	7.60
KLF3	7.56
PLAL1	6.88
YY1	6.06
CLAMP	5.69
RREB1	5.64
ZN219	5.57
PACC	5.28
SNT2	5.27
ZBTB6	5.18
MAZ	4.77
ZNF143	4.70
E2F5	4.56
ETS	4.32
STAT3	4.17
KR	3.70
XRCC4	3.52
SETB1	3.44
DBP	3.39

TF	Score
STA5B	34.45
ZNF521	33.95
E47	31.42
E2F5	30.89
ZBTB24	28.96
CTCF	26.94
TFE3	26.55
ZNF317	25.95
ZNF143	24.73
PAZ5	23.07
BBX	21.72
HES1	21.17
ZN148	21.13
ZIC3	21.04
STAT3	20.96
STAT1	20.28
AP1	20.09
KLF3	18.15
PARG	17.78
WT1	17.51
XRCC4	17.08
NKX25	16.99
YLL054C	16.85
SIRT6	16.85
MAZ	16.46
ETS	15.72
RUNX2	13.89
SMC3	13.84
RAD21	13.31
EGR2	13.26

652

653

654

655 **Table 3.** Biased orientation of repeat DNA sequences in monocytes. Score: $-\log_{10}$

656 (*p*-value).

Reverse-forward (RF) orientation	
DNA repeat seq.	Score
LTR16C	81.65
L1ME4b	36.78
AluSg	32.26

657

658

659

660 **Table 4.** Top 30 of co-locations of biased orientation of DNA binding motif sequences
661 of transcription factors in monocytes. Co-locations of DNA motif sequences of CTCF
662 with other biased orientation of DNA motif sequences were shown in a separate table.
663 Motif 1,2: DNA binding motif sequences of transcription factors. # both: the number of
664 open chromatin regions including both Motif 1 and Motif 2. # motif 1: the number of
665 open chromatin regions including Motif 1. # motif 2: the number of open chromatin
666 regions including Motif 2. # others: the number of open chromatin regions not including
667 Motif 1 or Motif 2. Open chromatin regions overlapped with histone modification
668 (H3K27ac) were used (Total 26,095 regions).

669

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	ZN263	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

670

Motif 1	Motif 2	# both	# motif 1	# motif 2	# others	*p*-value
SMAD2	ZN263	7427	1578	1805	15285	0
SMRC2	ZN263	7354	1192	1893	15656	0
RREB1	SP1	7208	1957	1165	15765	0
SIRT6	ZN263	7126	1154	2121	15694	0
MAZ	SP1	7052	909	1321	16813	0
MAZ	ZN263	7022	999	2225	15849	0
SP1	ZN263	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	SMRC2	6668	1612	1878	15937	0
PLAL1	SMAD2	6637	1074	2368	16016	0
MAZ	SMAD2	6627	1334	2378	15756	0
RARG	SMAD2	6602	1012	2403	16078	0
PLAL1	ZN263	6571	1140	2661	15723	0
EGR2	RREB1	6565	787	2600	16143	0
KLF6	RREB1	6563	932	2573	16027	0
MAZ	SMRC2	6562	1459	1984	16090	0
RREB1	ZN263	6551	1607	2681	15256	0
KLF6	SP1	6541	954	1832	16768	0
PLAL1	RREB1	6496	1215	2640	15744	0
EGR2	SP1	6492	860	1881	16862	0
KLF6	SMAD2	6491	1004	2514	16086	0
RARG	ZN263	6390	1224	2842	15639	0
PLAL1	SP1	6318	1393	2055	16329	0
KLF6	ZN263	6304	1191	2928	15672	0
EGR2	SMAD2	6270	1082	2735	16008	0
EGR2	ZN263	6253	1099	2979	15764	0
RARG	RREB1	6219	1395	2917	15564	0
EGR2	MAZ	6219	1133	1742	17001	0

Table 5. Comparison of biased orientation of DNA binding motif sequences of transcription factors with chromatin interaction data in monocytes. TF: DNA binding motif sequences of transcription factors. Score: $-\log_{10}$ (p-value). Score 1: Comparison of EPA shortened at genomic locations of FR or RF orientation of DNA binding sequences with EPA not shortened. Score 2: Comparison of EPA shortened at genomic locations of FR or RF orientation of DNA binding sequences with EPA shortened at genomic locations of any orientation of DNA binding sequences.

Forward-reverse

TF	Score 1	Score 2
AP1	28.64	3.75
AT3G04100	24.51	7.07
BTEB2	8.64	2.69
CTCF	6.09	6.80
E2F	19.43	3.24
E2F5	10.58	2.02
EGR2	4.28	10.11
ETS	7.66	5.74
FOSB	7.98	1.99
FOXO4	7.98	5.84
GCM1	2.51	3.47
GCR	6.39	4.50
IRF7	16.36	2.03
KLF14	18.52	8.08
KLF3	6.42	6.43
MAZ	1.82	4.55
MIG2	42.64	6.22
PARG	5.38	4.21
PITX3	10.31	5.62
PLAL1	11.50	5.42
PO6F1	31.28	7.60
PTF1A	8.91	4.50
PU1	3.86	1.71
RAD21	9.28	5.16
RREB1	7.47	6.40
RUNX2	2.97	1.68
SETB1	1.63	2.08
SMAD2	4.10	3.48
SMC3	10.24	6.21
SP1	20.98	4.64
TCFAP2A	5.10	1.63
TF3B	3.21	5.75
TF3CB	20.26	4.66
TFE3	94.90	3.64
YY1	13.88	7.04
ZBTB6	15.08	3.93
ZIC1	5.37	1.63
ZN148	14.51	4.47
ZN219	15.44	7.99
ZNF28	11.88	3.36
ZNF460	14.14	1.63

Reverse-forward orientation

TF	Score 1	Score 2
ASCL2	6.23	3.80
ATF3	17.21	2.23
BC11A	1.65	4.07
BDP1	23.49	6.85
CTCF	12.32	11.74
E2F2	23.33	1.95
EBF1	14.83	7.11
GABP1_GABP2	6.43	2.80
IRF3	2.38	1.98
IRF8	4.30	4.62
NRF3	17.95	1.78
PAX5	48.74	2.74
POLR3A	41.85	9.98
RAD21	1.79	2.26
RNF96	7.69	1.93
RPN4	21.32	4.48
RXRA	29.89	11.87
SMAD4	41.27	2.44
SP2	2.09	11.15
SP4	12.74	1.84
SP8	20.03	2.17
TCFAP2A	8.90	3.53
TCFAP2E	5.72	1.84
ZFP202	23.15	4.85

693 **References**

694

695 Bailey SD, Zhang X, Desai K, Aid M, Corradin O, Cowper-Sal Lari R, Akhtar-Zaidi B,
696 Scacheri PC, Haibe-Kains B, Lupien M. 2015. ZNF143 provides sequence
697 specificity to secure chromatin interactions at gene promoters. *Nat Commun* **2**:
698 6186.
699 Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble
700 WS. 2009. MEME SUITE: tools for motif discovery and searching. *Nucleic*
701 *acids research* **37**: W202-208.
702 Barutcu AR, Lajoie BR, Fritz AJ, McCord RP, Nickerson JA, van Wijnen AJ, Lian JB,
703 Stein JL, Dekker J, Stein GS et al. 2016. SMARCA4 regulates gene expression
704 and higher-order chromatin structure in proliferating mammary epithelial cells.
705 *Genome research* **26**: 1188-1201.
706 Bejerano G, Lowe CB, Ahituv N, King B, Siepel A, Salama SR, Rubin EM, Kent WJ,
707 Haussler D. 2006. A distal enhancer and an ultraconserved exon are derived
708 from a novel retroposon. *Nature* **441**: 87-90.
709 de Souza FS, Franchini LF, Rubinstein M. 2013. Exaptation of transposable elements
710 into novel cis-regulatory elements: is the evidence always strong? *Mol Biol Evol*
711 **30**: 1239-1251.
712 de Wit E, Vos ES, Holwerda SJ, Valdes-Quezada C, Verstegen MJ, Teunissen H,
713 Splinter E, Wijchers PJ, Krijger PH, de Laat W. 2015. CTCF Binding Polarity
714 Determines Chromatin Looping. *Molecular cell* **60**: 676-684.
715 Grant CE, Bailey TL, Noble WS. 2011. FIMO: scanning for occurrences of a given
716 motif. *Bioinformatics (Oxford, England)* **27**: 1017-1018.
717 Guo Y, Xu Q, Canzio D, Shou J, Li J, Gorkin DU, Jung I, Wu H, Zhai Y, Tang Y et al.
718 2015. CRISPR Inversion of CTCF Sites Alters Genome Topology and
719 Enhancer/Promoter Function. *Cell* **162**: 900-910.
720 Javierre BM, Burren OS, Wilder SP, Kreuzhuber R, Hill SM, Sewitz S, Cairns J,
721 Wingett SW, Varnai C, Thiecke MJ et al. 2016. Lineage-Specific Genome
722 Architecture Links Enhancers and Non-coding Disease Variants to Target Gene
723 Promoters. *Cell* **167**: 1369-1384 e1319.
724 Jeong M, Huang X, Zhang X, Su J, Shamim M, Bochkov I, Reyes J, Jung H, Heikamp

725 E, Presser Aiden A et al. 2017. A Cell Type-Specific Class of Chromatin Loops
726 Anchored at Large DNA Methylation Nadirs. *bioRxiv*.
727 Ji X, Dadon DB, Abraham BJ, Lee TI, Jaenisch R, Bradner JE, Young RA. 2015.
728 Chromatin proteomic profiling reveals novel proteins associated with
729 histone-marked genomic regions. *Proc Natl Acad Sci U S A* **112**: 3841-3846.
730 Jin C, Kato K, Chimura T, Yamasaki T, Nakade K, Murata T, Li H, Pan J, Zhao M, Sun
731 K et al. 2006. Regulation of histone acetylation and nucleosome assembly by
732 transcription factor JDP2. *Nat Struct Mol Biol* **13**: 331-338.
733 Jjingo D, Conley AB, Wang J, Marino-Ramirez L, Lunyak VV, Jordan IK. 2014.
734 Mammalian-wide interspersed repeat (MIR)-derived enhancers and the
735 regulation of human gene expression. *Mob DNA* **5**: 14.
736 John S, Sabo PJ, Thurman RE, Sung MH, Biddie SC, Johnson TA, Hager GL,
737 Stamatoyannopoulos JA. 2011. Chromatin accessibility pre-determines
738 glucocorticoid receptor binding patterns. *Nature genetics* **43**: 264-268.
739 Jolma A, Yan J, Whitington T, Toivonen J, Nitta KR, Rastas P, Morgunova E, Enge M,
740 Taipale M, Wei G et al. 2013. DNA-binding specificities of human transcription
741 factors. *Cell* **152**: 327-339.
742 Karolchik D, Barber GP, Casper J, Clawson H, Cline MS, Diekhans M, Dreszer TR,
743 Fujita PA, Guruvadoo L, Haeussler M et al. 2014. The UCSC Genome Browser
744 database: 2014 update. *Nucleic acids research* **42**: D764-770.
745 Kheradpour P, Kellis M. 2014. Systematic discovery and characterization of regulatory
746 motifs in ENCODE TF binding experiments. *Nucleic acids research* **42**:
747 2976-2987.
748 Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler
749 transform. *Bioinformatics (Oxford, England)* **25**: 1754-1760.
750 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G,
751 Durbin R, Genome Project Data Processing S. 2009. The Sequence
752 Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)* **25**:
753 2078-2079.
754 McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM,
755 Bejerano G. 2010. GREAT improves functional interpretation of cis-regulatory
756 regions. *Nature biotechnology* **28**: 495-501.
757 Mumbach MR, Satpathy AT, Boyle EA, Dai C, Gowen BG, Cho SW, Nguyen ML,

758 Rubin AJ, Granja JM, Kazane KR et al. 2017. Enhancer connectome in primary
759 human cells identifies target genes of disease-associated DNA elements. *Nature*
760 *genetics* **49**: 1602-1612.
761 Newburger DE, Bulyk ML. 2009. UniPROBE: an online database of protein binding
762 microarray data on protein-DNA interactions. *Nucleic acids research* **37**:
763 D77-82.
764 Osato N. 2018. Characteristics of functional enrichment and gene expression level of
765 human putative transcriptional target genes. *BMC Genomics* **19**: 957.
766 Phanstiel DH, Van Bortle K, Spacek D, Hess GT, Shamim MS, Machol I, Love MI,
767 Aiden EL, Bassik MC, Snyder MP. 2017. Static and Dynamic DNA Loops form
768 AP-1-Bound Activation Hubs during Macrophage Development. *Molecular cell*
769 **67**: 1037-1048.e1036.
770 Portales-Casamar E, Thongjuea S, Kwon AT, Arenillas D, Zhao X, Valen E, Yusuf D,
771 Lenhard B, Wasserman WW, Sandelin A. 2010. JASPAR 2010: the greatly
772 expanded open-access database of transcription factor binding profiles. *Nucleic*
773 *acids research* **38**: D105-110.
774 Ramani V, Cusanovich DA, Hause RJ, Ma W, Qiu R, Deng X, Blau CA, Disteche CM,
775 Noble WS, Shendure J et al. 2016. Mapping 3D genome architecture through in
776 situ DNase Hi-C. *Nat Protoc* **11**: 2104-2121.
777 Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn
778 AL, Machol I, Omer AD, Lander ES et al. 2014. A 3D map of the human
779 genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**:
780 1665-1680.
781 Rebollo R, Romanish MT, Mager DL. 2012. Transposable elements: an abundant and
782 natural source of regulatory sequences for host genes. *Annu Rev Genet* **46**:
783 21-42.
784 Saeed S, Quintin J, Kerstens HH, Rao NA, Aghajanirefah A, Matarese F, Cheng SC,
785 Ratter J, Berentsen K, van der Ent MA et al. 2014. Epigenetic programming of
786 monocyte-to-macrophage differentiation and trained innate immunity. *Science*
787 *(New York, NY)* **345**: 1251086.
788 Schreiber J, Libbrecht M, Bilmes J, Noble W. 2017. Nucleotide sequence and DNaseI
789 sensitivity are predictive of 3D chromatin architecture. *bioRxiv*.
790 Tabuchi TM, Deplancke B, Osato N, Zhu LJ, Barrasa MI, Harrison MM, Horvitz HR,

791 Walhout AJ, Hagstrom KA. 2011. Chromosome-biased binding and gene
792 regulation by the Caenorhabditis elegans DRM complex. *PLoS Genet* **7**:
793 e1002074.
794 Wang J, Vicente-Garcia C, Seruggia D, Molto E, Fernandez-Minan A, Neto A, Lee E,
795 Gomez-Skarmeta JL, Montoliu L, Lunyak VV et al. 2015. MIR retrotransposon
796 sequences provide insulators to the human genome. *Proc Natl Acad Sci U S A*
797 **112**: E4428-4437.
798 Wang L, Wang S, Li W. 2012. RSeQC: quality control of RNA-seq experiments.
799 *Bioinformatics (Oxford, England)* **28**: 2184-2185.
800 Weintraub AS, Li CH, Zamudio AV, Sigova AA, Hannett NM, Day DS, Abraham BJ,
801 Cohen MA, Nabet B, Buckley DL et al. 2017. YY1 Is a Structural Regulator of
802 Enhancer-Promoter Loops. *Cell* **171**: 1573-1588 e1528.
803 Wingender E, Dietze P, Karas H, Knuppel R. 1996. TRANSFAC: a database on
804 transcription factors and their DNA binding sites. *Nucleic acids research* **24**:
805 238-241.
806 Xie Z, Hu S, Blackshaw S, Zhu H, Qian J. 2010. hPDI: a database of experimental
807 human protein-DNA interactions. *Bioinformatics (Oxford, England)* **26**:
808 287-289.
809 Zhang H, Li F, Jia Y, Xu B, Zhang Y, Li X, Zhang Z. 2017. Characteristic arrangement
810 of nucleosomes is predictive of chromatin interactions at kilobase resolution.
811 *Nucleic acids research* **45**: 12739-12751.
812 Zhang Y, An L, Xu J, Zhang B, Zheng WJ, Hu M, Tang J, Yue F. 2018. Enhancing Hi-C
813 data resolution with deep convolutional neural network HiCPlus. *Nat Commun*
814 **9**: 750.
815 Zhao Y, Stormo GD. 2011. Quantitative analysis demonstrates most transcription factors
816 require only simple models of specificity. *Nature biotechnology* **29**: 480-483.
817