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 287 biased orientation of DNA motifs [159 forward-reverse (FR) and 153 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 EPAs.

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. 2). The expression level
96 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 cohesin (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) (Fig. 2). The number of transcription factors
129 showing a significant difference of expression level of putative transcriptional target
130 genes was counted. To examine whether the orientation of DNA motif sequences
131 shortening EPA affected the number of transcription factors showing a significant
132 difference of expression level of putative transcriptional target genes, the number of the
133 transcription factors was compared among forward-reverse, reverse-forward, and any
134 orientation of DNA motif sequences shortening EPA, using chi-square test (p-value <
135 0.05). To avoid missing DNA motif sequences showing a relatively weak statistical
136 significance by multiple testing collection, the above analyses were conducted using
137 monocytes of four people independently, and DNA motif sequences found in monocytes
138 of four people in common were selected. Total 287 of biased (159 forward-reverse and
139 153 reverse-forward) orientation of DNA binding motif sequences of transcription
140 factors 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. 3; 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 SMARCC2, 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 265 of biased (158 forward-reverse and 139 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 four 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 Fig. S1 and Supplemental Table S2). Biased orientation of
157 DNA motif sequences in T cells included CTCF, cohesin (RAD21 and SMC3), ZNF143,
158 YY1, and SMARC. Biased orientation of DNA motif sequences in T cells and
159 monocytes (when using H3K27ac histone modification marks, which is described in the
160 next paragraph) included JUNDM2 (JDP2), which is involved in histone-chaperone
161 activity, promoting nucleosome, and inhibition of histone acetylation (Jin et al. 2006).
162 JDP2 forms a homodimer or heterodimer with various transcription factors
163 (https://en.wikipedia.org/wiki/Jun_dimerization_protein). Among 287, 61 of biased
164 orientation of DNA binding motif sequences of transcription factors were found in both

165 monocytes and T cells in common (Supplemental Table S5). For each orientation, 31
166 forward-reverse and 34 reverse-forward orientation of DNA binding motif sequences of
167 transcription factors were found in both monocytes and T cells in common. Without
168 considering the difference of orientation of DNA binding motif sequences, 84 of biased
169 orientation of DNA binding motif sequences of transcription factors were found in both
170 monocytes and T cells. As a reason for the increase of the number from 61, a
171 transcription factor or an isoform of the same transcription factor may bind to a
172 different DNA binding motif sequence according to cell types and/or in the same cell
173 type. About 50% or more of alternative splicing isoforms are differently expressed
174 among tissues, indicating that most alternative splicing is subject to tissue-specific
175 regulation (Wang et al. 2008) (Chen and Manley 2009) (Das et al. 2007). The same
176 transcription factor has several DNA binding motif sequences and in some cases one of
177 the motif sequence is almost the same as the reverse complement sequence of another
178 motif sequence of the same transcription factor. For example, RAD21 showed both
179 forward-reverse and reverse-forward orientation of the DNA motif sequences, but the
180 number of forward-reverse orientation was relatively small in the genome, and
181 reverse-forward orientation of the DNA motif sequence was frequently observed and
182 co-localized with CTCF. I previously found that a complex of transcription factors
183 would bind to a slightly different DNA binding motif sequence from the combination of
184 DNA binding motif sequences of transcription factors composing the complex in *C.*
185 *elegans* (Tabuchi et al. 2011). From another viewpoint of this study, the expression level
186 of putative transcription target genes of some transcription factors would be different,

187 depending on the genomic locations (enhancers or promoters) of DNA binding motif
188 sequences of the transcription factors in monocytes and T cells of four people.
189 Moreover, using open chromatin regions overlapped with H3K27ac histone
190 modification marks known as enhancer and promoter marks, the same analyses were
191 performed in monocytes and T cells. H3K27ac histone modification marks were used in
192 the analysis of EPI, but were not used in the analysis of TF as insulator like CTCF and
193 cohesin in this study, since new biased orientation of DNA motif sequences was found
194 in this criterion. When H3K27ac histone modification marks were used in the analysis
195 of TF as insulator like CTCF and cohesin, the number of biased orientation of DNA
196 motif sequences was decreased. Total 241 of biased (141 forward-reverse and 120
197 reverse-forward) orientation of DNA binding motif sequences of transcription factors
198 have been found in monocytes of four people in common, whereas only one any
199 orientation (i.e. without considering orientation) of DNA binding motif sequences have
200 been found (Supplemental Table S3). Though the number of biased orientation of DNA
201 motif sequences was reduced, JDP2 was found in monocytes as well as T cells. CTCF,
202 RAD21, SMC3, ZNF143, YY1, and SMARCC2 were also found. For T cells using
203 H3K27ac histone modification marks, total 195 of biased (107 forward-reverse and 110
204 reverse-forward) orientation of DNA binding motif sequences of transcription factors
205 have been found in T cells of four people in common, whereas only six any orientation
206 of DNA binding motif sequences has been found (Supplemental Table S4). Though the
207 number of biased orientation of DNA motif sequences was reduced, CTCF, RAD21,
208 SMC3, YY1, and SMARCC2 were found. Scores of CTCF, RAD21, and SMC3 were

209 increased compared with the result of T cells without using H3K27ac histone
210 modification marks, and they were ranked in the top six. As summary of the results with
211 and without H3K27ac histone modification marks, total 386 of biased (223
212 forward-reverse and 214 reverse-forward) orientation of DNA motif sequences has been
213 found in monocytes of four people in common. Total 348 of biased (202
214 forward-reverse and 198 reverse-forward) orientation of DNA motif sequences has been
215 found in T cells of four people in common. Total of these results of monocytes and T
216 cells was 590 biased (368 forward-reverse and 359 reverse-forward) orientation of DNA
217 motif sequences. Biased orientation of DNA motif sequences found in both monocytes
218 and T cells was listed in Supplemental Table S5.

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

231 transcriptional target genes of each transcription factor show a significant difference
232 between promoters and EPA shortened at the genomic locations of DNA motif
233 sequences. Some experimental data of a cell did not show a significant difference
234 between promoters and EPA (Osato 2018).

235 Instead of DNA binding motif sequences of transcription factors, repeat DNA
236 sequences were also examined. The expression level of transcriptional target genes
237 predicted based on EPA shortened at the genomic locations of repeat DNA sequences
238 was compared with the expression level of transcriptional target genes predicted from
239 promoters. Three reverse-forward orientation of repeat DNA sequences showed a
240 significant difference of expression level of putative transcriptional target genes in
241 monocytes of four people in common (Table 3). Among them, LTR16C repeat DNA
242 sequence was observed in iPS and H1-hESC with enough statistical significance
243 considering multiple tests (p-value $< 10^{-7}$). The same as $CD14^{+}$ monocytes, biased
244 orientation of repeat DNA sequences was examined in $CD4^{+}$ T cells. Three
245 forward-reverse and two reverse-forward orientation of repeat DNA sequences showed
246 a significant difference of expression level of putative transcriptional target genes in T
247 cells of four people in common (Supplemental Table S7). MIRb and MIR3 were also
248 found in the analysis using open chromatin regions overlapped with H3K27ac histone
249 modification marks, which are enhancer and promoter marks. MIR and other transposon
250 sequences are known to be involved in insulators and enhancers (Bejerano et al. 2006;
251 Rebollo et al. 2012; 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 (forward-reverse and reverse-forward)
255 orientation of DNA binding motif sequences of transcription factors, co-location of the
256 DNA binding motif sequences in open chromatin regions in monocytes was analyzed.
257 The number of open chromatin regions with the same pairs of DNA binding motif
258 sequences was counted, and when the pairs of DNA binding motif sequences were
259 enriched with statistical significance (chi-square test, p-value $< 1.0 \times 10^{-10}$), they were
260 listed (Table 4; Supplemental Table S8). Open chromatin regions overlapped with
261 histone modification of an enhancer mark (H3K27ac) (total 26,095 regions) showed a
262 larger number of enriched pairs of DNA motifs than all open chromatin regions (Table
263 4; Supplemental Table S8). H3K27ac is known to be enriched at chromatin interaction
264 anchors (Phanstiel et al. 2017). As already known, CTCF was found with cohesin such
265 as RAD21 and SMC3 (Table 4). Top 30 pairs of FR and RF orientations of DNA motifs
266 co-occupied in the same open chromatin regions were shown (Table 4). Total number of
267 pairs of DNA motifs was 492, consisting of 115 unique DNA motifs, when the pair of
268 DNA motifs were observed in more than 80% of the number of open chromatin regions
269 with the DNA motifs.

270 The same as $CD14^+$ monocytes, to examine the association of 265 biased
271 orientation of DNA binding motif sequences of transcription factors in $CD4^+$ T cells,
272 co-location of the DNA binding motif sequences in open chromatin regions was
273 analyzed. Top 30 pairs of FR and RF orientations of DNA motifs co-occupied in the
274 same open chromatin regions were shown (Supplemental Table S9). Total number of

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

297 biased orientation of DNA motif sequence in their upstream or/and downstream was
298 about 20,000. This is almost the same way of analysis as co-location of biased
299 orientation of DNA motif sequences, and the analysis here used the number of genes
300 instead of open chromatin regions. Fig. 4 and Supplemental Fig. S2 showed that not
301 only a pair of the same transcription factor, but also a pair of different transcription
302 factors were found in upstream and downstream of genes. DNA motif sequences of
303 CTCF and cohesin (RAD21 and SMC3) were enriched near the same genes, but cohesin
304 tended to be found with another transcription factor near the same genes as well. The
305 number of some pairs of biased orientation of DNA motif sequences in genome
306 sequences was small, though they showed statistical significance (Supplemental
307 material 2). The result of the analysis seems to be different between monocytes and T
308 cells and between using H3K27ac histone modification marks and not using them
309 (Supplemental Fig. S2). Some of these transcription factors would form homodimers or
310 heterodimers directory or indirectly composing a complex structure, contributing to
311 make chromatin interactions (Fig. 6).
312
313 **Comparison with chromatin interaction data**
314 To examine whether the biased orientation of DNA motif sequences is associated
315 with chromatin interactions, enhancer-promoter interactions (EPI) predicted based on
316 enhancer-promoter associations (EPA) were compared with chromatin interaction data
317 (Hi-C). Due to the resolution of Hi-C experimental data used in this study (50kb), EPI
318 were adjusted to 50kb resolution. EPI were predicted based on three types of EPA: (i)

319 EPA without being shortened by the genomic locations of DNA motif sequences, (ii)
320 EPA shortened at the genomic locations of DNA motif sequences of transcription factors
321 such as CTCF and cohesin (RAD21, SMC3) without considering their orientation, and
322 (iii) EPA shortened at the genomic locations of forward-reverse or reverse-forward
323 orientation of DNA motif sequences of transcription factors. EPA (iii) showed the
324 highest ratio of EPI overlapped with chromatin interactions (Hi-C) using DNA binding
325 motif sequences of CTCF and cohesin (RAD21 and SMC3), compared with the other
326 two types of EPA (Supplemental Fig. S3). Total 59 biased orientation (40 FR and 23
327 RF) of DNA motif sequences including CTCF, cohesin, and YY1 showed a higher ratio
328 of EPI overlapped with Hi-C chromatin interactions than the other types of EPA in
329 monocytes (Supplemental Table S10). Chromatin interaction data were obtained from
330 different samples from DNase-seq, open chromatin regions, so individual differences
331 seemed to be large from the results of this analysis. Since, for some DNA motif
332 sequences of transcription factors, the number of EPI overlapped with chromatin
333 interactions was small, if higher resolution of chromatin interaction data (such as
334 HiChIP, in situ DNase Hi-C, and in situ Hi-C data, or a tool to improve the resolution
335 such as HiCPlus) is available, the number of EPI overlapped with chromatin
336 interactions would be increased and the difference of the numbers among three types of
337 EPA might be larger and more significant than the current results (Rao et al. 2014;
338 Ramani et al. 2016; Mumbach et al. 2017; Zhang et al. 2018).

339 After the analysis of $CD14^+$ monocytes, to utilize HiChIP chromatin interaction
340 data in $CD4^+$ T cells, the same analysis for $CD14^+$ monocytes was performed using

341 DNase-seq data of four donors in $CD4^+$ T cells (Mumbach et al. 2017). EPI predicted
342 based on EPA were compared with HiChIP chromatin interaction data in $CD4^+$ T cells.
343 The resolutions of HiChIP chromatin interaction data and EPI were adjusted to 5kb. EPI
344 were predicted based on the three types of EPA in the same way as $CD14^+$ monocytes
345 using top 60% of all the transcript excluding transcripts not expressed in T cells. The
346 criteria of the analysis were determined to include known DNA motif sequences
347 involved in chromatin interactions such as CTCF and cohesin in the result and the result
348 was consistent with the result using Hi-C chromatin interaction data. EPA (iii) showed
349 the highest ratio of EPI overlapped with chromatin interactions (HiChIP) using DNA
350 binding motif sequences of CTCF and cohesin (RAD21 and SMC3), compared with the
351 other two types of EPA (Fig. 5). Total 121 biased orientation (73 FR and 58 RF) of DNA
352 motif sequences including CTCF, cohesin, ZNF143, and SMARC showed a higher ratio
353 of EPI overlapped with HiChIP chromatin interactions than the other types of EPA in T
354 cells (Table 5). As expected, the number of EPI overlapped with chromatin interactions
355 (HiChIP) was increased, compared with Hi-C chromatin interactions.

356 Though the ratios of EPI overlapped with chromatin interactions were increased
357 by using many chromatin interaction data including lower score and count of chromatin
358 interactions (a cutoff score of CHiCAGO tool > 1 for Hi-C and more than 1,000 counts
359 for HiChIP), the ratios of EPI overlapped with chromatin interactions showed the same
360 tendency among three types of EPA (FR, any and others). The ratio of EPI overlapped
361 with Hi-C chromatin interactions was increased by using H3K27ac marks in both
362 monocytes and T cells. The ratio of EPI overlapped with HiChIP chromatin interactions

363 was also increased by using H3K27ac marks. The same as $CD14^+$ monocytes,
364 chromatin interaction data were obtained from different samples from DNase-seq, open
365 chromatin regions in $CD4^+$ T cells, so individual differences seemed to be large from
366 the results of this analysis, and (Mumbach et al. 2017) suggested that individual
367 differences of chromatin interactions were larger than those of open chromatin regions.
368 ATAC-seq data, open chromatin regions were available in $CD4^+$ T cells in the paper,
369 however, when using ATAC-seq data, the result of the analysis of biased orientation of
370 DNA motif sequences was different from DNase-seq data, and not included a part of
371 CTCF and cohesin. Thus, DNase-seq data collected from ENCODE and Blueprint
372 projects were employed in this study.
373

374 **Discussion**

375 To find DNA motif sequences of transcription factors and repeat DNA sequences
376 affecting human putative transcriptional target genes, DNA motif sequences were
377 searched from open chromatin regions of monocytes of four people. Total 287 biased
378 (159 forward-reverse and 153 reverse-forward) orientation of DNA motif sequences
379 were found in monocytes of four people in common, whereas only one any orientation
380 (i.e. without considering orientation) of DNA motif sequence was found to affect
381 putative transcriptional target genes, suggesting that EPA shortened at the genomic
382 locations of forward-reverse or reverse-forward orientation of DNA motif sequences of
383 transcription factors is an important character for EPI and transcription of genes.

384 When DNA motif sequences were searched from monocytes of one person, a

385 larger number of biased orientation of DNA motif sequences affecting human putative
386 transcriptional target genes has been found. When the number of people, from which
387 experimental data were obtained, was increased, the number of DNA motif sequences
388 found in all people in common decreased and in some cases, known transcription
389 factors involved in chromatin interactions such as CTCF and cohesin (RAD21 and
390 SMC3) were not identified by statistical tests. This would be caused by individual
391 difference of the same cell type, low quality of experimental data, and experimental
392 errors. Moreover, though forward-reverse orientation of DNA binding motif sequences
393 of CTCF and cohesin is frequently observed at chromatin interaction anchors, the
394 percentage of forward-reverse orientation is not 100, and other orientations of the DNA
395 binding motif sequences are also observed. Though DNA binding motif sequences of
396 CTCF and cohesin are found in various open chromatin regions, DNA binding motif
397 sequences of some transcription factors would be observed less frequently in open
398 chromatin regions. The analyses of cells of a number of people would avoid missing
399 relatively weak statistical significance of DNA motif sequences in cells of each person
400 by multiple testing correction of thousands of statistical tests. A DNA motif sequence
401 was found with p-value < 0.05 in cells of one person and the DNA motif sequence
402 found in the same cell type of four people in common would have p-value $< (0.05)^4 =$
403 0.00000625. Actually, DNA motif sequences with p-value slightly less than 0.05 in
404 monocytes of one person were observed in monocytes of four people in common.

405 EPI were compared with chromatin interactions (Hi-C) in monocytes. EPA
406 shortened at the genomic locations of DNA binding motif sequences of CTCF and

407 cohesin (RAD21 and SMC3) showed a significant difference of the ratios of EPI
408 overlapped with chromatin interactions, according to three types of EPA (FR, any, and
409 others). Using open chromatin regions overlapped with ChIP-seq experimental data of
410 histone modification of an enhancer mark (H3K27ac), the ratio of EPI not overlapped
411 with Hi-C was reduced. (Phanstiel et al. 2017) also reported that there was an especially
412 strong enrichment for loops with H3K27 acetylation peaks at both ends (Fisher's Exact
413 Test, $p = 1.4 \times 10^{-27}$). However, the total number of EPI overlapped with chromatin
414 interactions was also reduced using H3K27ac peaks, so more chromatin interaction data
415 would be needed to obtain reliable results in this study. As an issue of experimental data,
416 chromatin interaction data were obtained using different samples from DNase-seq, open
417 chromatin regions, so individual differences were not reflected to chromatin interaction
418 data. Moreover, the resolution of chromatin interaction data used in monocytes is about
419 50kb, thus the number of chromatin interactions was relatively small (72,284 at 50kb
420 resolution with a cutoff score of CHiCAGO tool > 1 and 16,501 with a cutoff score of
421 CHiCAGO tool > 5). EPA shortened at the genomic locations of DNA binding motif
422 sequences of transcription factors found in various open chromatin regions such as
423 CTCF and cohesin (RAD21 and SMC3) tend to be overlapped with a larger number of
424 chromatin interactions than DNA binding motif sequences of transcription factors less
425 frequently observed in open chromatin regions. Therefore, to examine the difference of
426 the numbers of EPI overlapped with chromatin interactions, according to three types of
427 EPA (FR or RF, any, and others), the number of chromatin interactions should be large
428 enough.

429 As HiChIP chromatin interaction data were available in $CD4^+$ T cells, biased
430 orientations of DNA motif sequences of transcription factors were examined in T cells
431 using DNase-seq data of four people. The resolutions of chromatin interactions and EPI
432 were adjusted to 5kb by fragmentation of genome sequences. In monocytes, the
433 resolution of Hi-C chromatin interaction data was converted by extending anchor
434 regions of chromatin interactions to 50kb length and merging chromatin interactions
435 overlapped with each other. Fragmentation of genome sequences may affect the
436 classification of chromatin interactions of which anchors are located near the border of a
437 fragment, but the number of chromatin interactions would not be decreased, compared
438 with merging chromatin interactions. The number of HiChIP chromatin interactions was
439 19,926,360 at 5kb resolution, 666,149 at 5kb resolution with more than 1,000 counts of
440 chromatin interactions, and 78,209 at 5kb resolution with more than 6,000 counts of
441 chromatin interactions. As expected, the number of EPI overlapped with chromatin
442 interactions was increased, and about half of biased orientation of DNA motif sequences
443 of transcription factors showed a statistical significance in EPI predicted based on EPA
444 shortened at the genomic locations of the DNA motif sequences, compared with the
445 other types of EPA. False positive predictions of EPI would be decreased by using
446 H3K27ac marks and other features. The ratio of EPI overlapped with Hi-C chromatin
447 interactions was increased by using H3K27ac marks in both monocytes and T cells. The
448 ratio of EPI overlapped with HiChIP chromatin interactions was also increased by using
449 H3K27ac marks. However, the number of biased orientation of DNA motif sequences
450 showing higher ratio of EPI overlapped with HiChIP chromatin interactions than the

451 other types of EPA was decreased using H3K27ac marks (Supplemental material 2).

452 Though the number of known DNA binding proteins and transcription factors
453 involved in chromatin interactions would be less than 20 or so, about 300 DNA binding
454 proteins showed enrichment of biased orientation of their DNA binding sequences, and
455 known DNA binding proteins involved in chromatin interactions were found in the
456 biased orientation of DNA binding sequences. JUNDM2 (JDP2) found in T cells and
457 monocytes is not known DNA binding proteins involved in chromatin interactions.
458 JDP2 is involved in histone and nucleosome-related functions such as
459 histone-chaperone activity, promoting nucleosome, and inhibition of histone acetylation
460 (Jin et al. 2006). JDP2 forms a homodimer or heterodimer with various transcription
461 factors (https://en.wikipedia.org/wiki/Jun_dimerization_protein). Thus, the main
462 function of JDP2 may not be associated with chromatin interactions, but it forms a
463 homodimer or heterodimer with transcription factors, potentially forming a chromatin
464 interaction through the binding of JDP2 and another transcription factor, like chromatin
465 interactions formed by YY1 transcription factor (Weintraub et al. 2017). In the analysis
466 of HiChIP chromatin interaction data, JDP2 was shown in the list of biased orientation
467 of DNA motif sequences showing higher ratio of EPI overlapped with HiChIP
468 chromatin interactions than the other types of EPA, using chromatin interactions with
469 more than 6,000 counts (Supplemental material 2). When forming a homodimer or
470 heterodimer with another transcription factor, transcription factors may bind to genome
471 DNA with a specific orientation of their DNA binding sequences. From the analysis of
472 biased orientation of DNA motif sequences of transcription factors, such transcription

473 factors would also be found. If the DNA binding motif sequence of only the mate to a
474 pair of transcription factors was found in EPA, EPA was shortened at one side, which is
475 the genomic location of the DNA binding motif sequence of the mate to the pair, and
476 transcriptional target genes were predicted using the EPA shortened at the side. In this
477 analysis, the mate to both heterodimer and homodimer of transcription factors can be
478 used to examine the effect on the expression level of transcriptional target genes
479 predicted based on the EPA shortened at one side. For heterodimer, biased orientation of
480 DNA motif sequences may also be found in forward-forward or reverse-reverse
481 orientation.

482 I reported that the ratio of the number of functional enrichment of putative
483 transcriptional target genes in the total number of putative transcriptional target genes
484 changed according to the types of EPA, in the same way as gene expression level (Osato
485 2018). JDP2 showed a higher ratio of functional enrichment of putative transcriptional
486 target genes using EPA shortened at the genomic locations of forward-reverse
487 orientation of the DNA binding motif sequence, compared with the other types of EPA
488 (reverse-forward or any orientation) in monocytes and T cells, among monocytes, T
489 cells and $CD20^+$ B cells. While the genome-wide analysis of functional enrichments of
490 putative transcriptional target genes takes computational time, the genome-wide
491 analysis of gene expression level is finished with less computational time. The analysis
492 of gene expression level can be used for a cell type and experimental data where the
493 expression levels of transcriptional target genes predicted based on EPA are
494 significantly different from those predicted from only promoter.

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

499 The analyses in this study revealed novel characters of DNA binding motif
500 sequences of transcription factors and repeat DNA sequences, and would be useful to
501 analyze transcription factors involved in chromatin interactions and/or forming a
502 homodimer, heterodimer or complex with other transcription factors, affecting the
503 expression and transcriptional regulation of genes.

504

505 ## Methods

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

507 To examine transcriptional regulatory target genes, bed files of hg38 of Blueprint
508 DNase-seq data for $CD14^+$ monocytes of four donors (EGAD00001002286; Donor ID:
509 C0010K, C0011I, C001UY, C005PS) were obtained from Blueprint project web site
510 (http://dcc.blueprint-epigenome.eu/#/home), and the bed files of hg38 were converted
511 into those of hg19 using Batch Coordinate Conversion (liftOver) web site
512 (https://genome.ucsc.edu/util.html). Bed files of hg19 of ENCODE H1-hESC
513 (GSM816632; UCSC Accession: wgEncodeEH000556), iPSC (GSM816642; UCSC
514 Accession: wgEncodeEH001110), HUVEC (GSM1014528; UCSC Accession:
515 wgEncodeEH002460), and MCF-7 (GSM816627; UCSC Accession:
516 wgEncodeEH000579) were obtained from the ENCODE websites

517 (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeUwDgf/;
518 http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeUwDnase/).
519 As high resolution of chromatin interaction data using HiChIP became available
520 in $CD4^+$ T cells, to promote the same analysis of $CD14^+$ monocytes in $CD4^+$ T cells,
521 DNase-seq data of four donors were obtained from public databases and Blueprint
522 projects web sites. DNase-seq data of only one donor was available in Blueprint project
523 for $CD4^+$ T cells (Donor ID: S008H1), and three other DNase-seq data were obtained
524 from NCBI Gene Expression Omnibus (GEO) database. Though the peak calling of
525 DNase-seq data available in GEO database was different from other DNase-seq data in
526 ENCODE and Blueprint projects where 150bp length of peaks were usually predicted
527 using HotSpot (John et al. 2011), FASTQ files of DNase-seq data were downloaded
528 from NCBI Sequence Read Archive (SRA) database (SRR097566, SRR097618, and
529 SRR171574). Read sequences of FASTQ files were aligned to the hg19 version of the
530 human genome reference using BWA (Li and Durbin 2009), and the BAM files
531 generated by BWA were converted into SAM files, sorted, and indexed using Samtools
532 (Li et al. 2009). Peaks of the DNase-seq data were predicted using HotSpot-4.1.1.
533 To identify transcription factor binding sites (TFBS) from the DNase-seq data,
534 TRANSFAC (2013.2), JASPAR (2010), UniPROBE, BEEML-PBM, high-throughput
535 SELEX, Human Protein-DNA Interactome, and transcription factor binding sequences
536 of ENCODE ChIP-seq data were used (Wingender et al. 1996; Newburger and Bulyk
537 2009; Portales-Casamar et al. 2010; Xie et al. 2010; Zhao and Stormo 2011; Jolma et al.
538 2013; Kheradpour and Kellis 2014). Position weight matrices of transcription factor

539 binding sequences were transformed into TRANSFAC matrices and then into MEME
540 matrices using in-house scripts and transfac2meme in MEME suite (Bailey et al. 2009).
541 Transcription factor binding sequences of transcription factors derived from vertebrates
542 were used for further analyses. Searches were conducted for transcription factor binding
543 sequences from each narrow peak using FIMO with *p*-value threshold of 10^{-5} (Grant et
544 al. 2011). Transcription factors corresponding to transcription factor binding sequences
545 were searched computationally by comparing their names and gene symbols of HGNC
546 (HUGO Gene Nomenclature Committee) -approved gene nomenclature and 31,848
547 UCSC known canonical transcripts
548 (http://hgdownload.cse.ucsc.edu/goldenpath/hg19/database/knownCanonical.txt.gz), as
549 transcription factor binding sequences were not linked to transcript IDs such as UCSC,
550 RefSeq, and Ensembl transcripts.

551 Target genes of a transcription factor were assigned when its TFBS was found in
552 DNase-seq narrow peaks in promoter or extended regions for enhancer-promoter
553 association of genes (EPA). Promoter and extended regions were defined as follows:
554 promoter regions were those that were within distances of ±5 kb from transcriptional
555 start sites (TSS). Promoter and extended regions were defined as per the following
556 association rule, which is the same as that defined in Figure 3A of a previous study
557 (McLean et al. 2010): the single nearest gene association rule, which extends the
558 regulatory domain to the midpoint between the TSS of the gene and that of the nearest
559 gene upstream and downstream without the limitation of extension length. Extended
560 regions for EPA were shortened at the genomic locations of DNA binding sites of

561 transcription factors that were the closest to a transcriptional start site, and
562 transcriptional target genes were predicted from the shortened enhancer regions using
563 TFBS. Furthermore, promoter and extended regions for EPA were shortened at the
564 genomic locations of forward–reverse orientation of DNA binding sites of transcription
565 factors. When forward or reverse orientation of DNA binding sites were continuously
566 located in genome sequences several times, the most external forward–reverse
567 orientation of DNA binding sites were selected. The genomic positions of genes were
568 identified using 'knownGene.txt.gz' file in UCSC bioinformatics sites (Karolchik et al.
569 2014). The file 'knownCanonical.txt.gz' was also utilized for choosing representative
570 transcripts among various alternate forms for assigning promoter and extended regions
571 for EPA. From the list of transcription factor binding sequences and transcriptional
572 target genes, redundant transcription factor binding sequences were removed by
573 comparing the target genes of a transcription factor binding sequence and its
574 corresponding transcription factor; if identical, one of the transcription factor binding
575 sequences was used. When the number of transcriptional target genes predicted from a
576 transcription factor binding sequence was less than five, the transcription factor binding
577 sequence was omitted.

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

582 For gene expression data, RNA-seq reads mapped onto human hg19 genome

583 sequences were obtained, including ENCODE long RNA-seq reads with poly-A of
584 H1-hESC, iPSC, HUVEC, and MCF-7 (GSM26284, GSM958733, GSM2344099,
585 GSM2344100, GSM958734, and GSM765388), and UCSF-UBC human reference
586 epigenome mapping project RNA-seq reads with poly-A of naive $CD4^+$ T cells
587 (GSM669617). Two replicates were present for H1-hESC, iPSC, HUVEC, and MCF-7,
588 and a single one for $CD4^+$ T cells. RPKMs of the RNA-seq data were calculated using
589 RSeQC (Wang et al. 2012). For monocytes, Blueprint RNA-seq RPKM data
590 (GSE58310, GSE58310_GeneExpression.csv.gz, Monocytes_Day0_RPMI) was used
591 (Saeed et al. 2014). Based on RPKM, UCSC transcripts with expression levels among
592 top 30% of all the transcripts were selected in each cell type.

593 The expression level of transcriptional target genes predicted based on EPA
594 shortened at the genomic locations of DNA motif sequences of transcription factors or
595 repeat DNA sequences was compared with the expression level of transcriptional target
596 genes predicted from promoter. For each DNA motif sequence shortening EPA,
597 transcriptional target genes were predicted using about 3,000 – 5,000 DNA binding
598 motif sequences of transcription factors, and the expression level of putative
599 transcriptional target genes of each transcription factor was compared between EPA and
600 only promoter using Mann-Whitney test (p-value < 0.05). The number of transcription
601 factors showing a significant difference of expression level of putative transcriptional
602 target genes between EPA and promoter was compared among forward-reverse,
603 reverse-forward, and any orientation of DNA motif sequences shortening EPA using
604 chi-square test (p-value < 0.05). When DNA motif sequences of transcription factors or

605 repeat DNA sequences shortening EPA showed a significant difference of expression
606 level of putative transcriptional target genes in forward-reverse, reverse-forward, or any
607 orientation in monocytes of four people in common, the DNA motif sequences were
608 listed.

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

620

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

622 Co-location of biased orientation of DNA binding motif sequences of
623 transcription factors was examined. The number of open chromatin regions with the
624 same pairs of DNA binding motif sequences was counted, and when the pairs of DNA
625 binding motif sequences were enriched with statistical significance (chi-square test,
626 p-value $< 1.0 \times 10^{-10}$), they were listed.

627 For histone modification of an enhancer mark (H3K27ac), bed files of hg38 of
628 Blueprint ChIP-seq data for $CD14^+$ monocytes (EGAD00001001179) and $CD4^+$ T cells
629 (Donor ID: S000RD) was obtained from Blueprint web site
630 (http://dcc.blueprint-epigenome.eu/#/home), and the bed files of hg38 were converted
631 into those of hg19 using Batch Coordinate Conversion (liftOver) web site
632 (https://genome.ucsc.edu/util.html).
633
634 **Comparison with chromatin interaction data**
635 For comparison of EPA in monocytes with chromatin interactions,
636 'PCHiC_peak_matrix_cutoff0.txt.gz' file was downloaded from 'Promoter Capture
637 Hi-C in 17 human primary blood cell types' web site (https://osf.io/u8tzp/files/), and
638 chromatin interactions for Monocytes with scores of CHiCAGO tool > 1 and
639 CHiCAGO tool > 5 were extracted from the file (Javierre et al. 2016). In the same way
640 as monocytes, Hi-C chromatin interaction data of $CD4^+$ T cells (Naive $CD4^+$ T cells,
641 nCD4) were obtained.
642 EPI were predicted using three types of EPA in monocytes: (i) EPA shortened at
643 the genomic locations of forward-reverse or reverse-forward orientation of DNA motif
644 sequences of transcription factors, (ii) EPA shortened at the genomic locations of any
645 orientation (i.e. without considering orientation) of DNA motif sequences of
646 transcription factors, and (iii) EPA without being shortened by DNA motif sequences.
647 EPI predicted using the three types of EPA in common were removed. First, EPI
648 predicted based on EPA (i) were compared with chromatin interactions (Hi-C). The

649 resolution of chromatin interaction data used in this study was 50kb, so EPI were
650 adjusted to 50kb before their comparison. The number and ratio of EPI overlapped with
651 chromatin interactions were counted. Second, EPI were predicted based on EPA (ii), and
652 EPI predicted based on EPA (i) were removed from the EPI. The number and ratio of
653 EPI overlapped with chromatin interactions were counted. Third, EPI were predicted
654 based on EPA (iii), and EPI predicted based on EPA (i) and (ii) were removed from the
655 EPI. The number and ratio of EPI overlapped with chromatin interactions were counted.
656 The number and ratio of the EPI compared two times between EPA (i) and (iii), and
657 EPA (i) and (ii) (binomial distribution, p-value < 0.025 for each test).

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

663

664 **Acknowledgements**

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

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

Figures

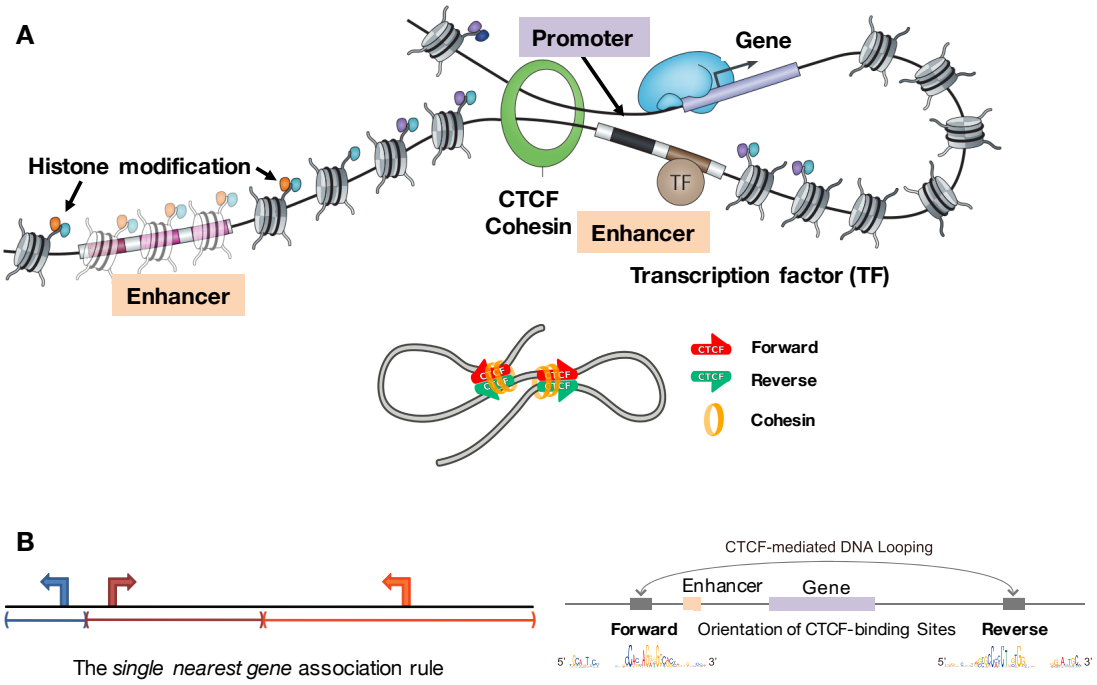
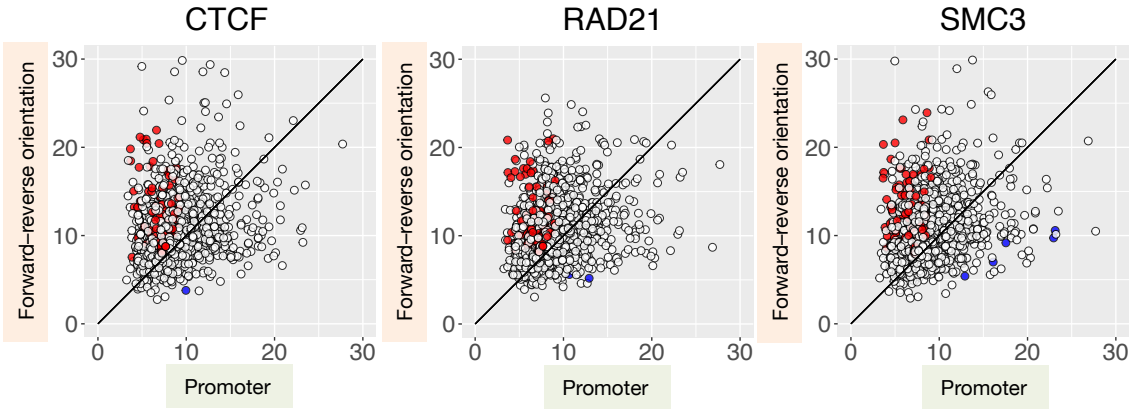


Figure 1. Chromatin interactions and enhancer-promoter association. (A) Forward–reverse orientation of CTCF-binding sites are frequently found in chromatin interaction anchors. CTCF can block the interaction between enhancers and promoters limiting the activity of enhancers to certain functional domains (de Wit et al. 2015; Guo et al. 2015). (B) Computationally-defined regulatory domains for enhancer-promoter association (McLean et al. 2010). The *single nearest gene* association rule extends the regulatory

685 domain to the midpoint between this gene's TSS and the nearest gene's TSS both
686 upstream and downstream. Enhancer-promoter association was shortened at the
687 genomic locations of forward-reverse orientation of DNA binding motif sequences of
688 transcription factors (e.g. CTCF in the figure).
689
690



691
692 **Figure 2.** Comparison of median expression level of putative transcriptional target
693 genes. The median expression levels of target genes of the same transcription factor
694 binding sequences were compared between promoters and EPA shortened at the
695 genomic locations of forward-reverse orientation of DNA motif sequences of CTCF and
696 cohesin (RAD21 and SMC3) in monocytes. Red and blue dots show statistically
697 significant difference (Mann-Whitney test) of the distribution of expression levels of
698 target genes between promoters and EPA. Red dots show the median expression level of
699 target genes was higher based on EPA than promoters, and blue dots show the median
700 expression level of target genes was lower based on EPA than promoters. The
701 expression level of target genes predicted based on the EPA tended to be higher than
702 promoters, implying that transcription factors acted as enhancers of target genes.

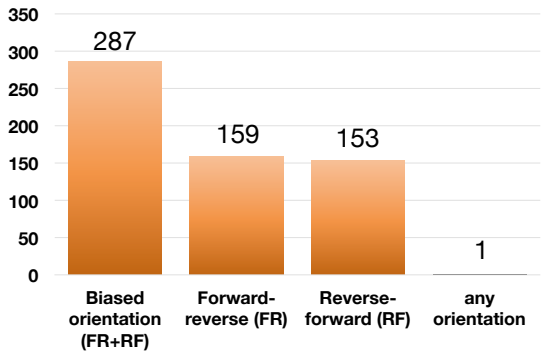


Figure 3. Biased orientation of DNA motif sequences of transcription factors in monocytes. Total 287 of biased orientation of DNA binding motif sequences of transcription factors have been found to affect the expression level of putative transcriptional target genes in monocytes of four people in common, whereas only one any orientation (i.e. without considering orientation) of DNA binding motif sequences have been found to affect the expression level.

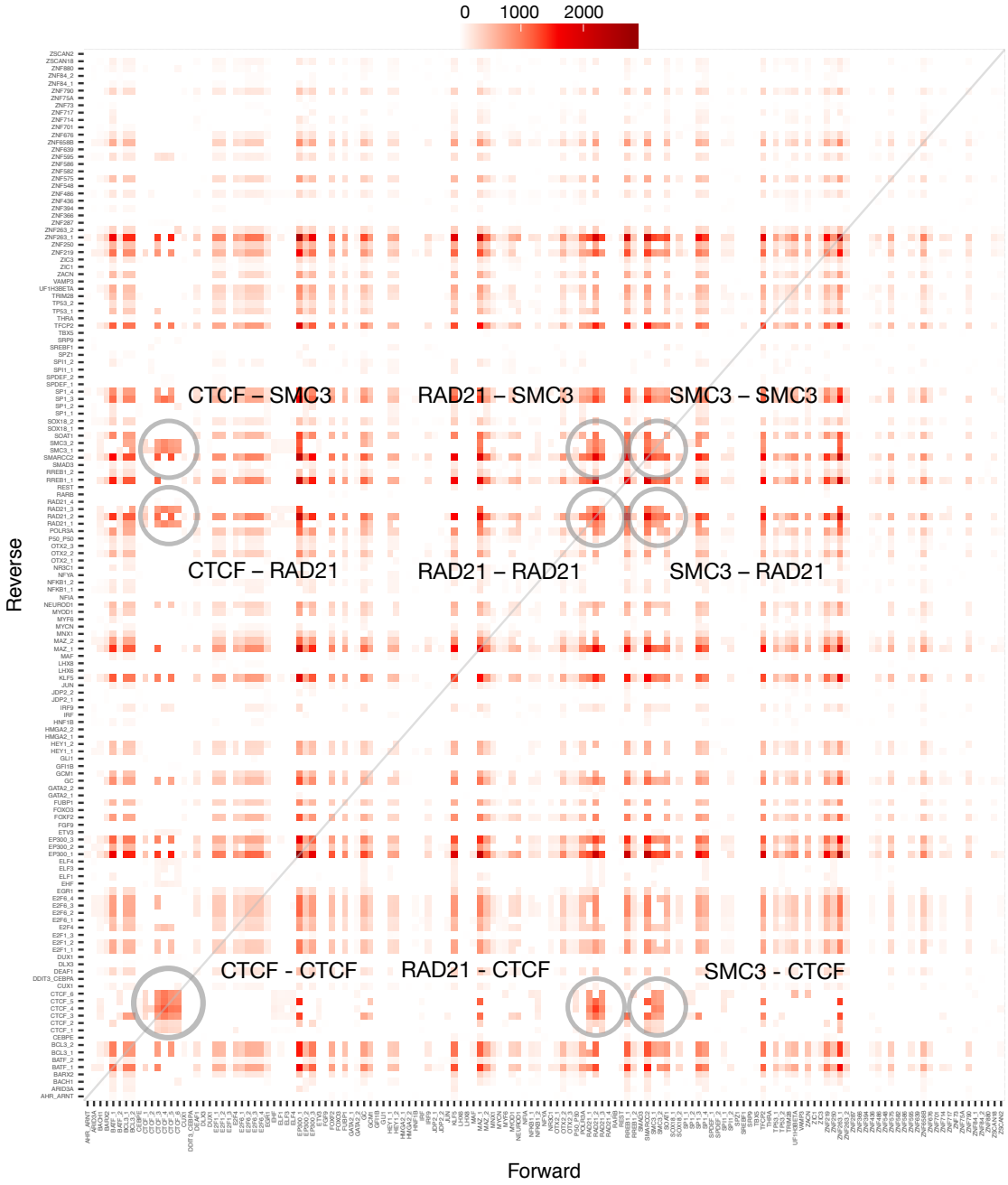


Figure 4. Pairs of biased orientation of DNA binding motif sequences of transcription factors enriched in upstream and downstream of genes. The number of genes with a pair of forward-reverse orientation of DNA motif sequences was counted using all pairs of the DNA motif sequences found in open chromatin regions overlapped with H3K27ac histone modification marks in T cells, and statistical tests (chi-square test) were

720 conducted. The number of genes where a pair of forward-reverse orientation of DNA

721 motif sequences were enriched with p-value < 0.05 was counted and shown as heat map.

722 Gray circles show pairs of DNA motif sequences of CTCF and cohesin (RAD21 and

723 SMC3). CTCF and cohesin are known to be involved in chromatin interactions and

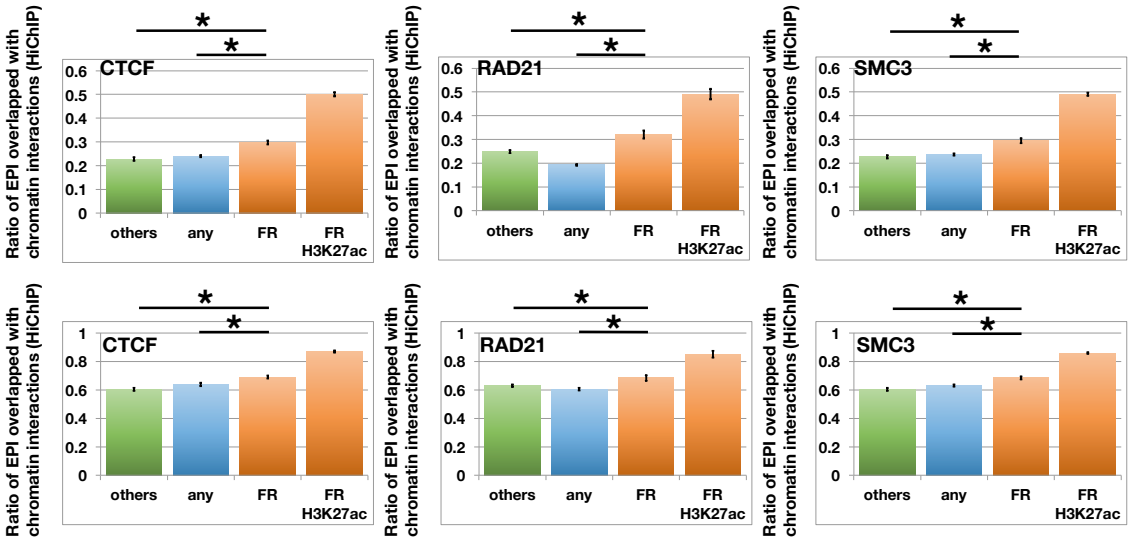
724 have biased orientation of DNA binding motif sequences to form chromatin loops.

725 Other pairs of forward-reverse orientation of DNA motif sequences were also enriched

726 in upstream and downstream of genes.

727

728



729

730 **Figure 5.** Comparison of DNA binding motif sequences of transcription factors with

731 chromatin interactions in T cells. Total 121 biased orientation (73 FR and 58 RF) of

732 DNA motif sequences including CTCF and cohesin showed a higher ratio of EPI

733 overlapped with HiChIP than the other types of EPA in T cells. The upper part of the

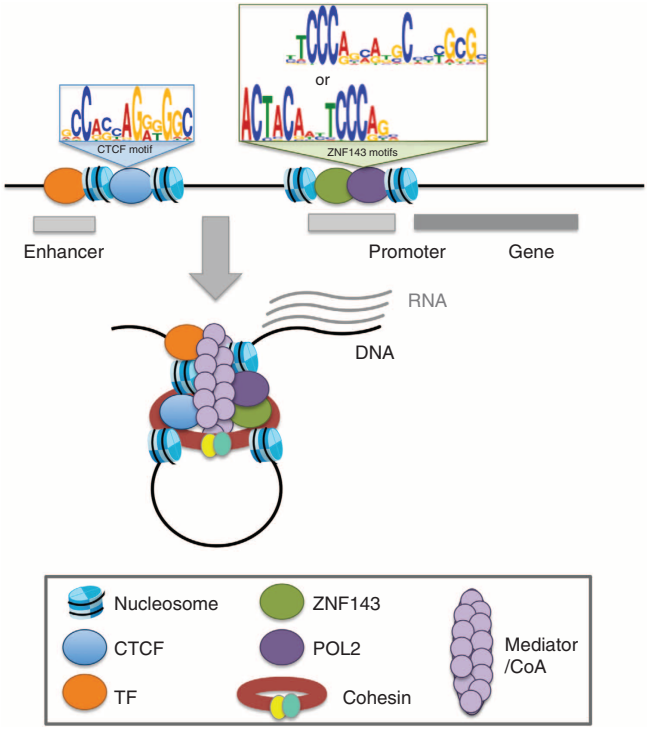
734 figure is a comparison of EPI with HiChIP chromatin interactions with more than 6,000

735 counts, and the lower part of the figure is a comparison of EPI with HiChIP chromatin

736 interactions with more than 1,000 counts. FR: forward-reverse orientation of DNA
737 motifs. any: any orientation of DNA motifs. others: enhancer-promoter association not
738 shortened at the genomic locations of DNA motifs. FR H3K27ac: EPI was predicted
739 from DNA motifs in open chromatin regions overlapped with H3K27ac histone
740 modification marks. * p-value < 10^{-4}.

741

742



743

744 **Figure 6.** Schematic representation of chromatin interactions involving gene promoters.
745 ZNF143 contributes the formation of chromatin interactions by directly binding the
746 promoter of genes establishing looping with distal element bound by CTCF (Bailey et al.
747 2015).

748

749

750 # Tables

751 **Table 1.** Top 90 biased orientation of DNA binding motif sequences of transcription

752 factors in monocytes. TF: DNA binding motif sequences of transcription factors. Score:

753 $-\log_{10}$ (*p*-value).

754 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

755
756
757
758
759
760
761
762
763
764
765

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

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

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

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

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

Repeat DNA seq.	Score
MLT1F1	6.75

Reverse-forward (RF) orientation

Repeat DNA seq.	Score
LTR16C	69.39
L1ME4b	23.90
AluSg	23.61

789 **Table 4.** Top 30 of co-locations of biased orientation of DNA binding motif sequences
790 of transcription factors in monocytes. Co-locations of DNA motif sequences of CTCF
791 with the other biased orientation of DNA motif sequences were shown in a separate
792 table. Motif 1,2: DNA binding motif sequences of transcription factors. # both: the
793 number of open chromatin regions including both Motif 1 and Motif 2. # motif 1: the
794 number of open chromatin regions including Motif 1. # motif 2: the number of open
795 chromatin regions including Motif 2. # others: the number of open chromatin regions
796 not including Motif 1 or Motif 2. Open chromatin regions overlapped with histone
797 modification marks (H3K27ac) were used (Total 26,095 regions).

798

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

799

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

800 **Table 5.** Comparison of biased orientation of DNA binding motif sequences of
801 transcription factors with chromatin interaction data in T cells. TF: DNA binding motif
802 sequences of transcription factors. Score: $-\log_{10}$ (p-value). Inf: p-value = 0. Score 1:
803 Comparison of EPA shortened at the genomic locations of FR or RF orientation of DNA
804 binding sequences with EPA not shortened. Score 2: Comparison of EPA shortened at
805 the genomic locations of FR or RF orientation of DNA binding sequences with EPA
806 shortened at the genomic locations of any orientation of DNA binding sequences.

807

808 Forward-reverse (FR)

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

809

References

Bailey SD, Zhang X, Desai K, Aid M, Corradin O, Cowper-Sal Lari R, Akhtar-Zaidi B, Scacheri PC, Haibe-Kains B, Lupien M. 2015. ZNF143 provides sequence specificity to secure chromatin interactions at gene promoters. *Nat Commun* **2**: 6186.

Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS. 2009. MEME SUITE: tools for motif discovery and searching. *Nucleic acids research* **37**: W202-208.

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

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

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

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

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

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

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

Guo Y, Xu Q, Canzio D, Shou J, Li J, Gorkin DU, Jung I, Wu H, Zhai Y, Tang Y et al. 2015. CRISPR Inversion of CTCF Sites Alters Genome Topology and

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

875 transform. *Bioinformatics (Oxford, England)* **25**: 1754-1760.
876 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G,
877 Durbin R, Genome Project Data Processing S. 2009. The Sequence
878 Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)* **25**:
879 2078-2079.
880 McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM,
881 Bejerano G. 2010. GREAT improves functional interpretation of cis-regulatory
882 regions. *Nature biotechnology* **28**: 495-501.
883 Mumbach MR, Satpathy AT, Boyle EA, Dai C, Gowen BG, Cho SW, Nguyen ML,
884 Rubin AJ, Granja JM, Kazane KR et al. 2017. Enhancer connectome in primary
885 human cells identifies target genes of disease-associated DNA elements. *Nature*
886 *genetics* **49**: 1602-1612.
887 Newburger DE, Bulyk ML. 2009. UniPROBE: an online database of protein binding
888 microarray data on protein-DNA interactions. *Nucleic acids research* **37**:
889 D77-82.
890 Osato N. 2018. Characteristics of functional enrichment and gene expression level of
891 human putative transcriptional target genes. *BMC Genomics* **19**: 957.
892 Phanstiel DH, Van Bortle K, Spacek D, Hess GT, Shamim MS, Machol I, Love MI,
893 Aiden EL, Bassik MC, Snyder MP. 2017. Static and Dynamic DNA Loops form
894 AP-1-Bound Activation Hubs during Macrophage Development. *Molecular cell*
895 **67**: 1037-1048.e1036.
896 Portales-Casamar E, Thongjuea S, Kwon AT, Arenillas D, Zhao X, Valen E, Yusuf D,
897 Lenhard B, Wasserman WW, Sandelin A. 2010. JASPAR 2010: the greatly
898 expanded open-access database of transcription factor binding profiles. *Nucleic*
899 *acids research* **38**: D105-110.
900 Ramani V, Cusanovich DA, Hause RJ, Ma W, Qiu R, Deng X, Blau CA, Disteche CM,
901 Noble WS, Shendure J et al. 2016. Mapping 3D genome architecture through in
902 situ DNase Hi-C. *Nat Protoc* **11**: 2104-2121.
903 Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn
904 AL, Machol I, Omer AD, Lander ES et al. 2014. A 3D map of the human
905 genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**:
906 1665-1680.
907 Rebollo R, Romanish MT, Mager DL. 2012. Transposable elements: an abundant and

908 natural source of regulatory sequences for host genes. *Annu Rev Genet* **46**:
909 21-42.
910 Saeed S, Quintin J, Kerstens HH, Rao NA, Aghajanirefah A, Matarese F, Cheng SC,
911 Ratter J, Berentsen K, van der Ent MA et al. 2014. Epigenetic programming of
912 monocyte-to-macrophage differentiation and trained innate immunity. *Science*
913 *(New York, NY)* **345**: 1251086.
914 Schreiber J, Libbrecht M, Bilmes J, Noble W. 2017. Nucleotide sequence and DNaseI
915 sensitivity are predictive of 3D chromatin architecture. *bioRxiv*.
916 Tabuchi TM, Deplancke B, Osato N, Zhu LJ, Barrasa MI, Harrison MM, Horvitz HR,
917 Walhout AJ, Hagstrom KA. 2011. Chromosome-biased binding and gene
918 regulation by the Caenorhabditis elegans DRM complex. *PLoS Genet* **7**:
919 e1002074.
920 Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF,
921 Schroth GP, Burge CB. 2008. Alternative isoform regulation in human tissue
922 transcriptomes. *Nature* **456**: 470-476.
923 Wang J, Vicente-Garcia C, Seruggia D, Molto E, Fernandez-Minan A, Neto A, Lee E,
924 Gomez-Skarmeta JL, Montoliu L, Lunyak VV et al. 2015. MIR retrotransposon
925 sequences provide insulators to the human genome. *Proc Natl Acad Sci U S A*
926 **112**: E4428-4437.
927 Wang L, Wang S, Li W. 2012. RSeQC: quality control of RNA-seq experiments.
928 *Bioinformatics (Oxford, England)* **28**: 2184-2185.
929 Weintraub AS, Li CH, Zamudio AV, Sigova AA, Hannett NM, Day DS, Abraham BJ,
930 Cohen MA, Nabet B, Buckley DL et al. 2017. YY1 Is a Structural Regulator of
931 Enhancer-Promoter Loops. *Cell* **171**: 1573-1588 e1528.
932 Wingender E, Dietze P, Karas H, Knuppel R. 1996. TRANSFAC: a database on
933 transcription factors and their DNA binding sites. *Nucleic acids research* **24**:
934 238-241.
935 Xie Z, Hu S, Blackshaw S, Zhu H, Qian J. 2010. hPDI: a database of experimental
936 human protein-DNA interactions. *Bioinformatics (Oxford, England)* **26**:
937 287-289.
938 Zhang H, Li F, Jia Y, Xu B, Zhang Y, Li X, Zhang Z. 2017. Characteristic arrangement
939 of nucleosomes is predictive of chromatin interactions at kilobase resolution.
940 *Nucleic acids research* **45**: 12739-12751.

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

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