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 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) (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, cohesion (RAD21 and SMC3),
158 ZNF143, 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. The same transcription factor has several DNA binding motif sequences and in
174 some cases one of the motif sequence is almost the same as the reverse complement
175 sequence of another motif sequence of the same transcription factor. For example,
176 RAD21 showed both forward-reverse and reverse-forward orientation of the DNA motif
177 sequences, but the number of forward-reverse orientation was relatively small in the
178 genome, and reverse-forward orientation of the DNA motif sequence was frequently
179 observed and co-localized with CTCF. I previously found that a complex of
180 transcription factors would bind to a slightly different DNA binding motif sequence
181 from the combination of DNA binding motif sequences of transcription factors
182 consisting of the complex in *C. elegans* (Tabuchi et al. 2011). From another viewpoint
183 of this study, the expression level of putative transcription target genes of some
184 transcription factors would be different, depending on the genomic locations (enhancers
185 or promoters) of DNA binding motif sequences of the transcription factors in monocytes
186 and T cells of four people.

187 Moreover, using open chromatin regions overlapped with H3K27ac histone
188 modification marks known as enhancer and promoter marks, the same analyses were
189 performed in monocytes and T cells. H3K27ac histone modification marks were used in
190 the analysis of EPI, but were not used in the analysis of TF as insulator like CTCF and
191 cohesion, since in this condition, new biased orientation of DNA motif sequences was
192 found. Total 241 of biased (141 forward-reverse and 120 reverse-forward) orientation of
193 DNA binding motif sequences of transcription factors have been found in monocytes of
194 four people in common, whereas only one any orientation (i.e. without considering
195 orientation) of DNA binding motif sequences have been found (Supplemental Table S3).
196 Though the number of biased orientation of DNA motif sequences was reduced, JDP2
197 was found in monocytes as well as T cells. CTCF, RAD21, SMC3, ZNF143, YY1, and
198 SMARCC2 were also found. For T cells using H3K27ac histone modification marks,
199 total 195 of biased (107 forward-reverse and 110 reverse-forward) orientation of DNA
200 binding motif sequences of transcription factors have been found in T cells of four
201 people in common, whereas only six any orientation of DNA binding motif sequences
202 have been found (Supplemental Table S4). Though the number of biased orientation of
203 DNA motif sequences was reduced, CTCF, RAD21, SMC3, YY1, and SMARCC2 were
204 found. Scores of CTCF, RAD21, and SMC3 were increased compared with the result of
205 T cells without using H3K27ac histone modification marks, and they were ranked in the
206 top six. As summary of the results with and without H3K28ac histone modification
207 marks, total 386 of biased (223 forward-reverse and 214 reverse-forward) orientation of
208 DNA motif sequences have been found in monocytes of four people in common. Total

209 348 of biased (202 forward-reverse and 198 reverse-forward) orientation of DNA motif
210 sequences have been found in T cells of four people in common. Total of these results of
211 monocytes and T cells was 590 biased (368 forward-reverse and 359 reverse-forward)
212 orientation of DNA motif sequences. Biased orientation of DNA motif sequences found
213 in both monocytes and T cells was listed in Supplemental Table S5.

214 To examine whether the biased orientation of DNA binding motif sequences of
215 transcription factors was observed in other cell types, the same analyses were conducted
216 in other cell types. However, for other cell types, experimental data of one sample were
217 available in ENCODE database, so the analyses of DNA motif sequences were
218 performed by comparing with the result in monocytes of four people. Among the biased
219 orientation of DNA binding motif sequences found in monocytes, 71, 69, 82, 56, and 53
220 DNA binding motif sequences were also observed in $CD4^+$ T cells, H1-hESC, iPS,
221 Huvec and MCF-7 respectively, including CTCF and cohesin (RAD21 and SMC3)
222 (Table 2; Supplemental Table S6). The scores of DNA binding motif sequences were the
223 highest in monocytes, and the other cell types showed lower scores. The results of the
224 analysis of DNA motif sequences in $CD20^+$ B cells and macrophages did not include
225 CTCF and cohesin, because these analyses can be utilized in cells where the expression
226 level of putative transcriptional target genes of each transcription factor show a
227 significant difference between promoters and EPA shortened at the genomic locations of
228 DNA motif sequences. Some experimental data of a cell did not show a significant
229 difference between promoters and EPA (Osato 2018).

230 Instead of DNA binding motif sequences of transcription factors, repeat DNA

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

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

249 To examine association of 287 biased (forward-reverse and reverse-forward)
250 orientation of DNA binding motif sequences of transcription factors, co-location of the
251 DNA binding motif sequences in open chromatin regions in monocytes was analyzed.
252 The number of open chromatin regions with the same pairs of DNA binding motif

253 sequences was counted, and when the pairs of DNA binding motif sequences were
254 enriched with statistical significance (chi-square test, *p*-value < 1.0×10^{-10}), they were
255 listed (Table 4; Supplemental Table S8). Open chromatin regions overlapped with
256 histone modification of an enhancer mark (H3K27ac) (total 26,095 regions) showed a
257 larger number of enriched pairs of DNA motifs than all open chromatin regions (Table
258 4; Supplemental Table S8). H3K27ac is known to be enriched at chromatin interaction
259 anchors (Phanstiel et al. 2017). As already known, CTCF was found with cohesin such
260 as RAD21 and SMC3 (Table 4). Top 30 pairs of FR and RF orientations of DNA motifs
261 co-occupied in the same open chromatin regions were shown (Table 4). Total number of
262 pairs of DNA motifs was 492, consisting of 115 unique DNA motifs, when the pair of
263 DNA motifs were observed in more than 80% of the number of open chromatin regions
264 with the DNA motifs.

265 The same as $CD14^+$ monocytes, to examine association of 265 biased orientation
266 of DNA binding motif sequences of transcription factors, co-location of the DNA
267 binding motif sequences in open chromatin regions in $CD4^+$ T cells was analyzed. Top
268 30 pairs of FR and RF orientations of DNA motifs co-occupied in the same open
269 chromatin regions were shown (Supplemental Table S9). Total number of pairs of DNA
270 motifs was 332, consisting of 118 unique DNA motifs, when the pair of DNA motifs
271 were observed in more than 60% of the number of open chromatin regions with the
272 DNA motifs (chi-square test, *p*-value < 1.0×10^{-10}). Among them, 12 pairs of DNA
273 motif sequences including a pair of CTCF, SMC3, and RAD21 in T cells were found in
274 monocytes in common (Supplemental Table S9).

275

276 **Pairs of biased orientation of DNA motif sequences surrounding genes**

277 CTCF and cohesin are found in chromatin interaction anchors and are considered
278 to form homodimers. Transcription factors are also known to form heterodimers with
279 another transcription factor. In this study, if the DNA binding motif sequence of only the
280 mate to a pair of transcription factors was found in EPA, EPA was shortened at one side,
281 the genomic location of the DNA binding motif sequence of the mate to the pair, and
282 transcriptional target genes were predicted using the EPA shortened at one side. In this
283 analysis, the mate to both heterodimer and homodimer of transcription factors can be
284 used to examine the effect on the expression level of transcriptional target genes
285 predicted based on the EPA shortened at one side. To examine whether biased
286 orientation of DNA motif sequences tend to be found as a pair of the same transcription
287 factor or different transcription factors in upstream and downstream of genes, the
288 enrichment of pairs of biased orientation of DNA motif sequences was investigated. The
289 number of genes having a pair of forward-reverse or reverse-forward orientation of
290 DNA motif sequences was counted using all pairs of the DNA motif sequences, and
291 statistical tests (chi-square test) were conducted, which is almost the same as the
292 analysis of co-location of biased orientation of DNA motif sequences using the number
293 of genes instead of open chromatin regions in this analysis. Fig. 4 and Supplemental Fig.
294 S2 showed that not only a pair of the same transcription factor, but also a pair of
295 different transcription factors were found in upstream and downstream of genes. DNA
296 motif sequences of CTCF and cohesin (RAD21 and SMC3) were enriched near the

297 same genes, but cohesin tended to be found with another transcription factor near the
298 same genes as well. The result of the analysis seems to be different between monocytes
299 and T cells. Some of these transcription factors would form homodimers or
300 heterodimers directory or indirectly composing a complex structure, contributing to
301 make chromatin interactions.
302
303 **Comparison with chromatin interaction data**
304 To examine whether the biased orientation of DNA motif sequences is associated
305 with chromatin interactions, enhancer-promoter interactions (EPI) predicted based on
306 enhancer-promoter associations (EPA) were compared with chromatin interaction data
307 (Hi-C). Due to the resolution of Hi-C experimental data used in this study (50kb), EPI
308 were adjusted to 50kb resolution. EPI were predicted based on three types of EPA: (i)
309 EPA without being shortened by DNA motif sequences, (ii) EPA shortened at the
310 genomic locations of DNA motif sequences of transcription factors such as CTCF and
311 cohesin (RAD21, SMC3) without considering their orientation, and (iii) EPA shortened
312 at the genomic locations of forward-reverse or reverse-forward orientation of DNA
313 motif sequences of transcription factors. EPA (iii) showed the highest ratio of EPI
314 overlapped with chromatin interactions (Hi-C) using DNA binding motif sequences of
315 CTCF and cohesin (RAD21 and SMC3), compared with the other two types of EPA
316 (Fig. 5). Total 59 biased orientation (40 FR and 23 RF) of DNA motif sequences
317 including CTCF and cohesin showed a higher ratio of EPI overlapped with Hi-C than
318 the other types of EPA in monocytes (Table 5). Chromatin interaction data were

319 obtained using different samples from DNase-seq, open chromatin regions, so
320 individual differences seemed to be large from the results of this analysis. Since, for
321 some DNA motif sequences of transcription factors, the number of EPI overlapped with
322 chromatin interactions was small, if higher resolution of chromatin interaction data
323 (such as HiChIP, in situ DNase Hi-C, and in situ Hi-C data, or a tool to improve the
324 resolution such as HiCPlus) is available, the number of EPI overlapped with chromatin
325 interactions would be increased and the difference of the numbers among three types of
326 EPA might be larger and more significant than the current results (Rao et al. 2014;
327 Ramani et al. 2016; Mumbach et al. 2017; Zhang et al. 2018).

328 After the analysis of $CD14^+$ monocytes, to utilize HiChIP chromatin interaction
329 data in $CD4^+$ T cells, the same analysis for $CD14^+$ monocytes was performed using
330 DNase-seq data of four donors in $CD4^+$ T cells (Mumbach et al. 2017). EPI predicted
331 based on EPA were compared with HiChIP chromatin interaction data in $CD4^+$ T cells.
332 The resolutions of HiChIP chromatin interaction data and EPI were adjusted to 1kb. EPI
333 were predicted based on the three types of EPA in the same way as $CD14^+$ monocytes.
334 EPA (iii) showed the highest ratio of EPI overlapped with chromatin interactions
335 (HiChIP) using DNA binding motif sequences of CTCF and cohesin (RAD21 and
336 SMC3), compared with the other two types of EPA (Supplemental Fig. S3). Total 134
337 biased orientation (88 FR and 54 RF) of DNA motif sequences including CTCF and
338 cohesin showed a higher ratio of EPI overlapped with HiChIP than the other types of
339 EPA in T cells (Supplemental Table S10). As expected, the number of EPI overlapped
340 with chromatin interactions (HiChIP) was increased, compared with Hi-C chromatin

341 interactions. However, the same as $CD14^+$ monocytes, chromatin interaction data were
342 obtained using different samples from DNase-seq, open chromatin regions in $CD4^+$ T
343 cells, so individual differences seemed to be large from the results of this analysis, and
344 (Mumbach et al. 2017) suggested that individual differences of chromatin interactions
345 were larger than those of open chromatin regions. ATAC-seq data, open chromatin
346 regions were available in $CD4^+$ T cells in the paper, however, when using ATAC-seq
347 data, the result of the analysis of biased orientation of DNA motif sequences was
348 different from DNase-seq data, and not included a part of CTCF and cohesin. Thus,
349 DNase-seq data collected from ENCODE and Blueprint projects were employed in this
350 study.
351

352 **Discussion**

353 To find DNA motif sequences of transcription factors and repeat DNA sequences
354 affecting human putative transcriptional target genes, DNA motif sequences were
355 searched from open chromatin regions of monocytes of four people. Total 287 biased
356 (159 forward-reverse and 153 reverse-forward) orientation of DNA motif sequences
357 were found in monocytes of four people in common, whereas only one any orientation
358 (i.e. without considering orientation) of DNA motif sequence was found to affect
359 putative transcriptional target genes, suggesting that EPA shortened at the genomic
360 locations of forward-reverse or reverse-forward orientation of DNA motif sequences of
361 transcription factors is an important character for EPI and transcription of genes.

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

363 larger number of biased orientation of DNA motif sequences affecting human putative
364 transcriptional target genes has been found. When the number of people, from which
365 experimental data were obtained, was increased, the number of DNA motif sequences
366 found in all people in common decreased and in some cases, known transcription
367 factors involved in chromatin interactions such as CTCF and cohesin (RAD21 and
368 SMC3) were not identified by statistical tests. This would be caused by individual
369 difference of the same cell type, low quality of experimental data, and experimental
370 errors. Moreover, though forward-reverse orientation of DNA binding motif sequences
371 of CTCF and cohesin is frequently observed at chromatin interaction anchors, the
372 percentage of forward-reverse orientation is not 100, and other orientations of the DNA
373 binding motif sequences are also observed. Though DNA binding motif sequences of
374 CTCF and cohesin are found in various open chromatin regions, DNA binding motif
375 sequences of some transcription factors would be observed less frequently in open
376 chromatin regions. The analyses of cells of a number of people would avoid missing
377 relatively weak statistical significance of DNA motif sequences in cells of each person
378 by multiple testing correction of thousands of statistical tests. A DNA motif sequence
379 was found with p-value < 0.05 in cells of one person and the DNA motif sequence
380 found in the same cell type of four people in common would have p-value $< (0.05)^4 =$
381 0.00000625. Actually, DNA motif sequences with p-value slightly less than 0.05 in
382 monocytes of one person were observed in monocytes of four people in common.

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

385 cohesin (RAD21 and SMC3) showed a significant difference of the ratios of EPI
386 overlapped with chromatin interactions, according to three types of EPA (FR, any, and
387 others). Using open chromatin regions overlapped with ChIP-seq experimental data of
388 histone modification of an enhancer mark (H3K27ac), the ratio of EPI not overlapped
389 with Hi-C was reduced and the difference of the ratio of EPI overlapped with chromatin
390 interactions among the three types of EPA became larger (data not shown). (Phanstiel et
391 al. 2017) also reported that there was an especially strong enrichment for loops with
392 H3K27 acetylation peaks at both ends (Fisher's Exact Test, $p = 1.4 \times 10^{-27}$). However,
393 the total number of EPI overlapped with chromatin interactions was also reduced using
394 H3K27ac peaks, so more chromatin interaction data would be needed to obtain reliable
395 results in this study. As an issue of experimental data, chromatin interaction data were
396 obtained using different samples from DNase-seq, open chromatin regions, so
397 individual differences were not reflected to chromatin interaction data. Moreover, the
398 resolution of chromatin interaction data used in monocytes is about 50kb, thus the
399 number of chromatin interactions was relatively small (72,284 at 50kb resolution with
400 cutoff score of CHiCAGO tool > 1). EPA shortened at the genomic locations of DNA
401 binding motif sequences of transcription factors found in various open chromatin
402 regions such as CTCF and cohesin (RAD21 and SMC3) tend to be overlapped with a
403 larger number of chromatin interactions than DNA binding motif sequences of
404 transcription factors less frequently observed in open chromatin regions. Therefore, to
405 examine the difference of the numbers of EPI overlapped with chromatin interactions,
406 according to three types of EPA (FR or RF, any, and others), the number of chromatin

407 interactions should be large enough.

408 As HiChIP chromatin interaction data were available in $CD4^+$ T cells, biased
409 orientations of DNA motif sequences of transcription factors were examined in T cells
410 using DNase-seq data of four people. The resolutions of chromatin interactions and EPI
411 were adjusted to 1kb by fragmentation of genome sequences. In monocytes, the
412 resolution of Hi-C chromatin interaction data was converted by extending anchor
413 regions of chromatin interactions to 50kb length and merging chromatin interactions
414 overlapped with each other. Fragmentation of genome sequences may affect the
415 classification of chromatin interactions of which anchors are located near the border of a
416 fragment, but the number of chromatin interactions would not be decreased, compared
417 with merging chromatin interactions. The number of HiChIP chromatin interactions was
418 28,918,687 at 1kb resolution. As expected, the number of EPI overlapped with
419 chromatin interactions was increased, and about half of biased orientation of DNA motif
420 sequences of transcription factors showed a statistical significance in EPI predicted
421 based on EPA shortened at the genomic locations of the DNA motif sequences,
422 compared with the other types of EPA. But the difference of the ratios of EPI
423 overlapped with chromatin interactions, according to three types of EPA (FR or RF, any,
424 and others), became smaller in T cells than monocytes (Supplemental Fig. S3), due to
425 higher resolution of chromatin interactions. False positive predictions of EPI would be
426 decreased by using H3K27ac marks and other features. The ratio of EPI overlapped
427 with Hi-C chromatin interactions was increased by using H3K27ac marks in both
428 monocytes and T cells. The ratio of EPI overlapped with HiChIP chromatin interactions

429 was also increased by using H3K27ac marks. However, the number of biased
430 orientation of DNA motif sequences showing higher ratio of EPI overlapped with
431 HiChIP chromatin interactions than the other types of EPA was slightly decreased using
432 H3K27ac marks (data not shown).

433 Though the number of known DNA binding proteins and transcription factors
434 involved in chromatin interactions would be less than 20 or so, about 300 DNA binding
435 proteins showed enrichment of biased orientation of their DNA binding sequences, and
436 known DNA binding proteins involved in chromatin interactions were found in the
437 biased orientation of DNA binding sequences. JUNDM2 (JDP2) found in T cells and
438 monocytes is not known DNA binding proteins involved in chromatin interactions.
439 JDP2 is involved in histone and nucleosome-related functions such as
440 histone-chaperone activity, promoting nucleosome, and inhibition of histone acetylation
441 (Jin et al. 2006). JDP2 forms a homodimer or heterodimer with various transcription
442 factors (https://en.wikipedia.org/wiki/Jun_dimerization_protein). Thus, the main
443 function of JDP2 may not be associated with chromatin interactions, but it forms a
444 homodimer or heterodimer with transcription factors, potentially forming a chromatin
445 interaction through the binding of JDP2 and another transcription factor, like chromatin
446 interactions formed by YY1 transcription factor (Weintraub et al. 2017). When forming
447 a homodimer or heterodimer with another transcription factor, transcription factors may
448 bind to genome DNA with a specific orientation of their DNA binding sequences. From
449 the analysis of biased orientation of DNA motif sequences of transcription factors, such
450 transcription factors would also be found. If the DNA binding motif sequence of only

451 the mate to a pair of transcription factors was found in EPA, EPA was shortened at one
452 side, the genomic location of the DNA binding motif sequence of the mate to the pair,
453 and transcriptional target genes were predicted using the EPA shortened at one side. In
454 this analysis, the mate to both heterodimer and homodimer of transcription factors can
455 be used to examine the effect on the expression level of transcriptional target genes
456 predicted based on the EPA shortened at one side. For heterodimer, biased orientation of
457 DNA motif sequences may be found in forward-forward or reverse-reverse orientation.

458 I reported that the ratio of the number of functional enrichment of putative
459 transcriptional target genes in the total number putative transcriptional target genes
460 changed according to the types of EPA, in the same way as gene expression level (Osato
461 2018). JDP2 showed a higher ratio of functional enrichment of putative transcriptional
462 target genes using EPA shortened at genomic locations of forward-reverse orientation of
463 the DNA binding motif sequence, compared with the other types of EPA
464 (reverse-forward or any orientation) in monocytes and T cells, among monocytes, T
465 cells and $CD20^+$ B cells. While the genome-wide analysis of functional enrichments of
466 putative transcriptional target genes takes computational time, the genome-wide
467 analysis of gene expression level is finished with less computational time. The analysis
468 of gene expression level can be used for a cell type and experimental data where the
469 expression levels of transcriptional target genes predicted from EPA are significantly
470 different from those predicted from only promoter.

471 It has been reported that CTCF and cohesion-binding sites are frequently mutated
472 in cancer (Katainen et al. 2015). Some biased orientation of DNA motif sequences

473 would be associated with chromatin interactions and might be associated with diseases
474 including cancer.

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

480

481 ## Methods

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

483 To examine transcriptional regulatory target genes, bed files of hg38 of Blueprint
484 DNase-seq data for $CD14^+$ monocytes of four donors (EGAD00001002286; Donor ID:
485 C0010K, C0011I, C001UY, C005PS) were obtained from Blueprint project web site
486 (http://dcc.blueprint-epigenome.eu/#/home), and the bed files of hg38 were converted
487 into those of hg19 using Batch Coordinate Conversion (liftOver) web site
488 (https://genome.ucsc.edu/util.html). Bed files of hg19 of ENCODE
489 $CD4^+$_Naive_Wb11970640 (GSM1014537; UCSC Accession: wgEncodeEH003156),
490 H1-hESC (GSM816632; UCSC Accession: wgEncodeEH000556), iPSC (GSM816642;
491 UCSC Accession: wgEncodeEH001110), HUVEC (GSM1014528; UCSC Accession:
492 wgEncodeEH002460), and MCF-7 (GSM816627; UCSC Accession:
493 wgEncodeEH000579) were obtained from the ENCODE websites
494 (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeUwDgf/;

495 http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeUwDnase/).

496 As high resolution of chromatin interaction data using HiChIP became available
497 in $CD4^+$ T cells, to promote the same analysis of $CD14^+$ monocytes in $CD4^+$ T cells,
498 DNase-seq data of four donors were obtained from ENCODE and Blueprint projects
499 web sites. DNase-seq data of only one donor was available in Blueprint project for
500 $CD4^+$ T cells (Donor ID: S008H1), and three other DNase-seq data were obtained from
501 ENCODE project. Though the peak calling of DNase-seq data available in ENCODE
502 database was different from other DNase-seq data in ENCODE and Blueprint projects
503 where 150bp length of peaks were usually predicted using HotSpot (John et al. 2011),
504 FASTQ files of DNase-seq data were downloaded from NCBI Sequence Read Archive
505 (SRA) database (SRR097566, SRR097618, and SRR171574). Read sequences of
506 FASTQ files were aligned to the hg19 version of the human genome reference using
507 BWA (Li and Durbin 2009), and the BAM files generated by BWA were converted into
508 SAM files, sorted, and indexed using Samtools (Li et al. 2009). Peaks of the DNase-seq
509 data were predicted using HotSpot-4.1.1.

510 To identify transcription factor binding sites (TFBS) from the DNase-seq data,
511 TRANSFAC (2013.2), JASPAR (2010), UniPROBE, BEEML-PBM, high-throughput
512 SELEX, Human Protein-DNA Interactome, and transcription factor binding sequences
513 of ENCODE ChIP-seq data were used (Wingender et al. 1996; Newburger and Bulyk
514 2009; Portales-Casamar et al. 2010; Xie et al. 2010; Zhao and Stormo 2011; Jolma et al.
515 2013; Kheradpour and Kellis 2014). Position weight matrices of transcription factor
516 binding sequences were transformed into TRANSFAC matrices and then into MEME

517 matrices using in-house scripts and transfac2meme in MEME suite (Bailey et al. 2009).
518 Transcription factor binding sequences of transcription factors derived from vertebrates
519 were used for further analyses. Searches were conducted for transcription factor binding
520 sequences from each narrow peak using FIMO with *p*-value threshold of 10^{-5} (Grant et
521 al. 2011). Transcription factors corresponding to transcription factor binding sequences
522 were searched computationally by comparing their names and gene symbols of HGNC
523 (HUGO Gene Nomenclature Committee) -approved gene nomenclature and 31,848
524 UCSC known canonical transcripts
525 (http://hgdownload.cse.ucsc.edu/goldenpath/hg19/database/knownCanonical.txt.gz), as
526 transcription factor binding sequences were not linked to transcript IDs such as UCSC,
527 RefSeq, and Ensembl transcripts.

528 Target genes of a transcription factor were assigned when its TFBS was found in
529 DNase-seq narrow peaks in promoter or extended regions for enhancer-promoter
530 association of genes (EPA). Promoter and extended regions were defined as follows:
531 promoter regions were those that were within distances of ±5 kb from transcriptional
532 start sites (TSS). Promoter and extended regions were defined as per the following
533 association rule, which is the same as that defined in Figure 3A of a previous study
534 (McLean et al. 2010): the single nearest gene association rule, which extends the
535 regulatory domain to the midpoint between the TSS of the gene and that of the nearest
536 gene upstream and downstream without the limitation of extension length. Extended
537 regions for EPA were shortened at the genomic locations of DNA binding sites of
538 transcription factors that were the closest to a transcriptional start site, and

539 transcriptional target genes were predicted from the shortened enhancer regions using
540 TFBS. Furthermore, promoter and extended regions for EPA were shortened at the
541 genomic locations of forward–reverse orientation of DNA binding sites of transcription
542 factors. When forward or reverse orientation of DNA binding sites were continuously
543 located in genome sequences several times, the most external forward–reverse
544 orientation of DNA binding sites were selected. The genomic positions of genes were
545 identified using 'knownGene.txt.gz' file in UCSC bioinformatics sites (Karolchik et al.
546 2014). The file 'knownCanonical.txt.gz' was also utilized for choosing representative
547 transcripts among various alternate forms for assigning promoter and extended regions
548 for EPA. From the list of transcription factor binding sequences and transcriptional
549 target genes, redundant transcription factor binding sequences were removed by
550 comparing the target genes of a transcription factor binding sequence and its
551 corresponding transcription factor; if identical, one of the transcription factor binding
552 sequences was used. When the number of transcriptional target genes predicted from a
553 transcription factor binding sequence was less than five, the transcription factor binding
554 sequence was omitted.

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

559 For gene expression data, RNA-seq reads mapped onto human hg19 genome
560 sequences were obtained, including ENCODE long RNA-seq reads with poly-A of

561 H1-hESC, iPSC, HUVEC, and MCF-7 (GSM26284, GSM958733, GSM2344099,
562 GSM2344100, GSM958734, and GSM765388), and UCSF-UBC human reference
563 epigenome mapping project RNA-seq reads with poly-A of naive $CD4^+$ T cells
564 (GSM669617). Two replicates were present for H1-hESC, iPSC, HUVEC, and MCF-7,
565 and a single one for $CD4^+$ T cells. RPKMs of the RNA-seq data were calculated using
566 RSeQC (Wang et al. 2012). For monocytes, Blueprint RNA-seq RPKM data
567 (GSE58310, GSE58310_GeneExpression.csv.gz, Monocytes_Day0_RPMI) was used
568 (Saeed et al. 2014). Based on RPKM, UCSC transcripts with expression levels among
569 top 30% of all the transcripts were selected in each cell type.

570 The expression level of transcriptional target genes predicted based on EPA
571 shortened at the genomic locations of DNA motif sequences of transcription factors or
572 repeat DNA sequences was compared with the expression level of transcriptional target
573 genes predicted from promoter. For each DNA motif sequence shortening EPA,
574 transcriptional target genes were predicted using about 3,000 – 5,000 DNA binding
575 motif sequences of transcription factors, and the expression level of putative
576 transcriptional target genes of each transcription factor was compared between EPA and
577 only promoter using Mann-Whitney test (p-value < 0.05). The number of transcription
578 factors showing a significant difference of expression level of putative transcriptional
579 target genes between EPA and promoter was compared among forward-reverse,
580 reverse-forward, and any orientation of DNA motif sequences shortening EPA using
581 chi-square test (p-value < 0.05). When DNA motif sequences of transcription factors or
582 repeat DNA sequences shortening EPA showed a significant difference of expression

583 level of putative transcriptional target genes in forward-reverse, reverse-forward, or any
584 orientation in monocytes of four people in common, the DNA motif sequences were
585 listed.

586 Though forward-reverse orientation of DNA binding motif sequences of CTCF
587 and cohesin is frequently observed at chromatin interaction anchors, the percentage of
588 forward-reverse orientation is not 100, and other orientations of the DNA binding motif
589 sequences are also observed. Though DNA binding motif sequences of CTCF and
590 cohesin are found in various open chromatin regions, DNA binding motif sequences of
591 some transcription factors would be observed less frequently in open chromatin regions.
592 The analyses of cells of a number of people would avoid missing relatively weak
593 statistical significance of DNA motif sequences in cells of each person by multiple
594 testing correction of thousands of statistical tests. A DNA motif sequence was found
595 with p-value < 0.05 in cells of one person and the DNA motif sequence found in the
596 same cell type of four people in common would have p-value $< 0.05^4 = 0.00000625$.

597

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

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

604 For histone modification of an enhancer mark (H3K27ac), bed files of hg38 of

605 Blueprint ChIP-seq data for $CD14^+$ monocytes (EGAD00001001179) and $CD4^+$ T cells
606 (Donor ID: S000RD) was obtained from Blueprint web site
607 (http://dcc.blueprint-epigenome.eu/#/home), and the bed files of hg38 were converted
608 into those of hg19 using Batch Coordinate Conversion (liftOver) web site
609 (https://genome.ucsc.edu/util.html).
610
611 **Comparison with chromatin interaction data**
612 For comparison of EPA in monocytes with chromatin interactions,
613 'PCHiC_peak_matrix_cutoff0.txt.gz' file was downloaded from 'Promoter Capture
614 Hi-C in 17 human primary blood cell types' web site (https://osf.io/u8tzp/files/), and
615 chromatin interactions for Monocytes with scores of CHiCAGO tool > 1 were extracted
616 from the file (Javierre et al. 2016).
617 EPI were predicted using three types of EPA in monocytes: (i) EPA shortened at
618 the genomic locations of forward-reverse or reverse-forward orientation of DNA motif
619 sequences of transcription factors, (ii) EPA shortened at the genomic locations of any
620 orientation (i.e. without considering orientation) of DNA motif sequences of
621 transcription factors, and (iii) EPA without being shortened by DNA motif sequences.
622 EPI predicted using the three types of EPA in common were removed. First, EPI
623 predicted based on EPA (i) were compared with chromatin interactions (Hi-C). The
624 resolution of chromatin interaction data used in this study was 50kb, so EPI were
625 adjusted to 50kb before their comparison. The number and ratio of EPI overlapped with
626 chromatin interactions were counted. Second, EPI were predicted based on EPA (ii), and

627 EPI predicted based on EPA (i) were removed from the EPI. The number and ratio of
628 EPI overlapped with chromatin interactions were counted. Third, EPI were predicted
629 based on EPA (iii), and EPI predicted based on EPA (i) and (ii) were removed from the
630 EPI. The number and ratio of EPI overlapped with chromatin interactions were counted.
631 The number and ratio of the EPI compared two times between EPA (i) and (iii), and
632 EPA (i) and (ii) (binomial distribution, p-value < 0.025 for each test).

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

637

638 **Acknowledgements**

639 The supercomputing resource was provided by Human Genome Center of the Institute
640 of Medical Science at the University of Tokyo. Computations were partially performed
641 on the NIG supercomputer at ROIS National Institute of Genetics. Publication charges
642 for this article were funded by JSPS KAKENHI Grant Number 16K00387. This
643 research was partially supported by the Platform Project for Supporting in Drug
644 Discovery and Life Science Research（Platform for Dynamic Approaches to Living
645 System）from Japan Agency for Medical Research and Development (AMED). This
646 research was partially supported by Development of Fundamental Technologies for
647 Diagnosis and Therapy Based upon Epigenome Analysis from Japan Agency for
648 Medical Research and Development (AMED). This work was partially supported by

JST CREST Grant Number JPMJCR15G1, Japan.

Figures

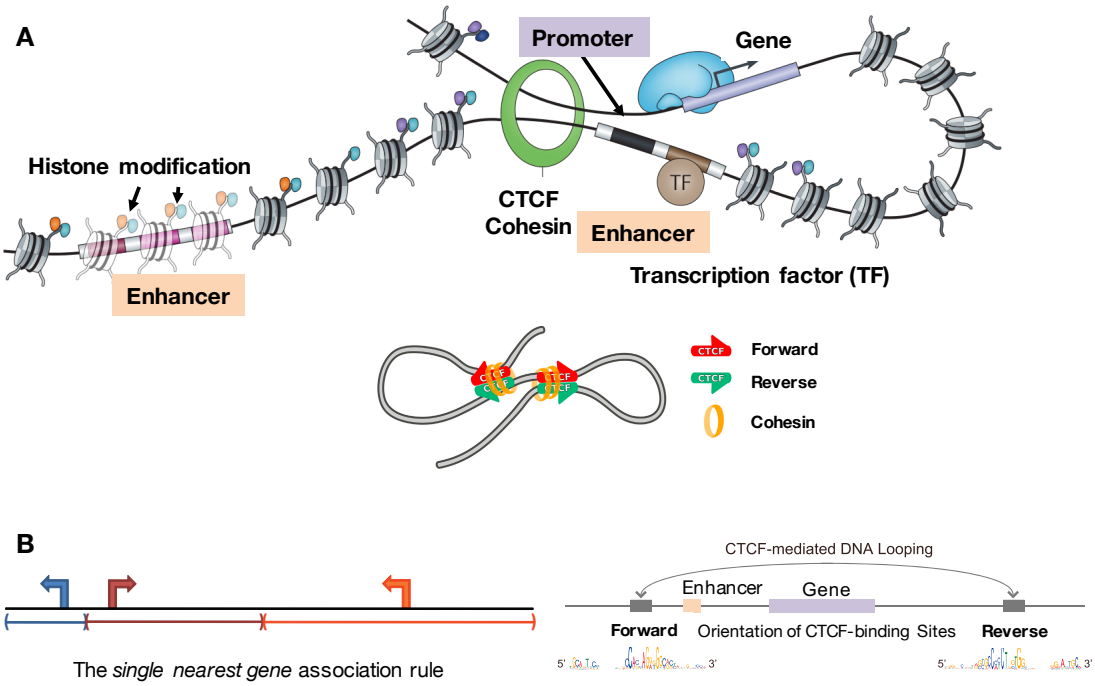

Figure 1. Chromatin interactions and enhancer-promoter association. (A) Forward–reverse orientation of CTCF-binding sites are frequently found in chromatin interaction anchors. CTCF can block the interaction between enhancers and promoters limiting the activity of enhancers to certain functional domains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

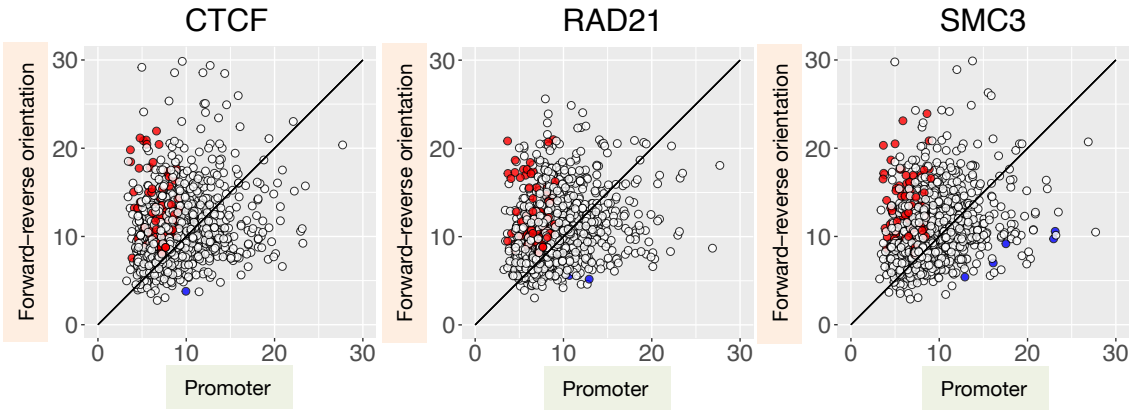

Figure 2. Comparison of median expression level of putative transcriptional target genes. The median expression levels of target genes of the same transcription factor binding sequences were compared between promoters and EPA shortened at the genomic locations of forward-reverse orientation of DNA motif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

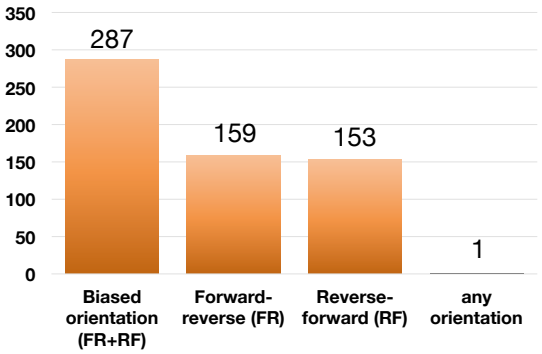

Figure 3. Biased orientation of DNA motif sequences of transcription factors in monocytes. Total 287 of biased orientation of DNA binding motif sequences of

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.

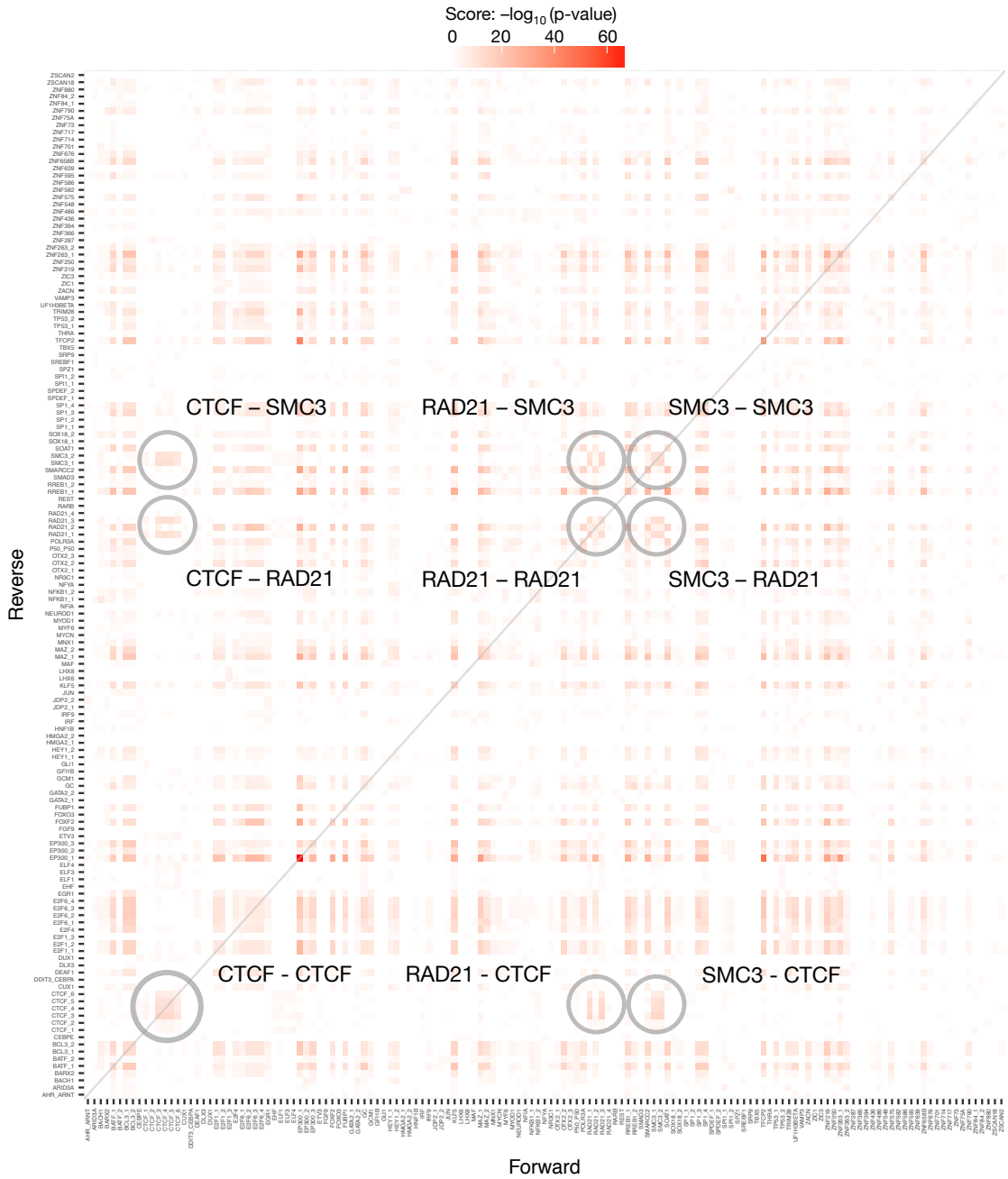

Figure 4. Pairs of biased orientation of DNA binding motif sequences of transcription factors enriched in upstream and downstream of some genes. The number of genes

having a pair of forward-reverse orientation of DNA motif sequences was counted using all pairs of the DNA motif sequences, and statistical tests (chi-square test) were conducted in T cells. Gray circles show pairs of DNA motif sequences of CTCF and cohesin (RAD21 and SMC3). CTCF and cohesin are known to be involved in chromatin interactions and tend to show biased orientation of DNA binding motif sequences to form chromatin loops. Other pairs of forward-reverse orientation of DNA motif sequences were also enriched near genes.

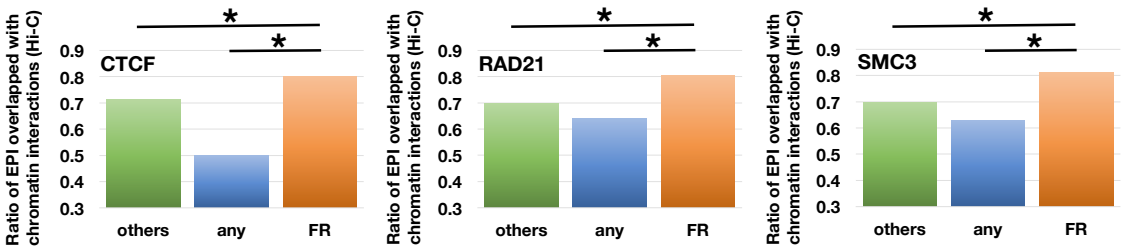

Figure 5. Comparison of DNA binding motif sequences of transcription factors with chromatin interactions in monocytes. Total 59 biased orientation (40 FR and 23 RF) of DNA motif sequences including CTCF and cohesin showed a higher ratio of EPI overlapped with Hi-C than the other types of EPA in monocytes. 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

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

Reverse-forward orientation

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

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).

Monocytes H3K27ac_EP

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

DNA repeat seq.	Score
MLT1F1	6.75

Reverse-forward (RF) orientation

DNA repeat seq.	Score
LTR16C	69.39
L1ME4b	23.90
AluSg	23.61

760 **Table 4.** Top 30 of co-locations of biased orientation of DNA binding motif sequences
761 of transcription factors in monocytes. Co-locations of DNA motif sequences of CTCF
762 with other biased orientation of DNA motif sequences were shown in a separate table.
763 Motif 1,2: DNA binding motif sequences of transcription factors. # both: the number of
764 open chromatin regions including both Motif 1 and Motif 2. # motif 1: the number of
765 open chromatin regions including Motif 1. # motif 2: the number of open chromatin
766 regions including Motif 2. # others: the number of open chromatin regions not including
767 Motif 1 or Motif 2. Open chromatin regions overlapped with histone modification
768 (H3K27ac) were used (Total 26,095 regions).

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

769

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

770

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
BRF1	3.21	5.75
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
FOS	28.64	3.75
FOSB	7.98	1.99
FOXO4	7.98	5.84
GCM1	2.51	3.47
GTF3C2	20.26	4.66
IRF7	16.36	2.03
KLF14	18.52	8.08
KLF3	6.42	6.43
KLF5	8.64	2.69
MAZ	1.82	4.55
NR3C1	6.39	4.50
PARG	5.38	4.21
PITX3	10.31	5.62
PLAGL1	11.50	5.42
POU6F1	31.28	7.60
PTF1A	8.91	4.50
RAD21	9.28	5.16
RFTN1	42.64	6.22
RREB1	7.47	6.40
RUNX2	2.97	1.68
SETDB1	1.63	2.08
SMAD2	4.10	3.48
SMC3	10.24	6.21
SP1	20.98	4.64
SPI1	3.86	1.71
TFAP2A	5.10	1.63
TFE3	94.90	3.64
YY1	13.88	7.04
ZBTB6	15.08	3.93
ZIC1	5.37	1.63
ZNF148	14.51	4.47
ZNF219	15.44	7.99
ZNF28	11.88	3.36
ZNF460	14.14	1.63

Reverse-forward orientation

TF	Score 1	Score 2
ATF3	29.41	5.12
BDP1	13.04	6.13
CTCF	11.59	8.71
DNAAF2	19.53	1.95
E2F2	3.90	1.74
EBF1	2.45	5.28
FOXP2	3.11	3.04
IRF4	9.16	1.71
IRF8	3.67	3.29
KLF16	6.94	2.53
PAX5	47.18	3.82
POLR3A	24.37	7.71
RAD21	4.14	2.23
RXRA	14.78	5.39
SMAD4	19.98	2.50
SP1	8.87	5.20
SP3	5.83	2.84
SP4	12.56	3.60
SP8	7.80	2.75
TFAP2A	4.62	3.59
TFAP2B	6.97	2.24
TRIM28	1.74	2.43
ZNF202	18.81	2.19

793 **References**

794

795 Bailey SD, Zhang X, Desai K, Aid M, Corradin O, Cowper-Sal Lari R, Akhtar-Zaidi B,
796 Scacheri PC, Haibe-Kains B, Lupien M. 2015. ZNF143 provides sequence
797 specificity to secure chromatin interactions at gene promoters. *Nat Commun* **2**:
798 6186.
799 Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble
800 WS. 2009. MEME SUITE: tools for motif discovery and searching. *Nucleic*
801 *acids research* **37**: W202-208.
802 Barutcu AR, Lajoie BR, Fritz AJ, McCord RP, Nickerson JA, van Wijnen AJ, Lian JB,
803 Stein JL, Dekker J, Stein GS et al. 2016. SMARCA4 regulates gene expression
804 and higher-order chromatin structure in proliferating mammary epithelial cells.
805 *Genome research* **26**: 1188-1201.
806 Bejerano G, Lowe CB, Ahituv N, King B, Siepel A, Salama SR, Rubin EM, Kent WJ,
807 Haussler D. 2006. A distal enhancer and an ultraconserved exon are derived
808 from a novel retroposon. *Nature* **441**: 87-90.
809 de Souza FS, Franchini LF, Rubinstein M. 2013. Exaptation of transposable elements
810 into novel cis-regulatory elements: is the evidence always strong? *Mol Biol Evol*
811 **30**: 1239-1251.
812 de Wit E, Vos ES, Holwerda SJ, Valdes-Quezada C, Verstegen MJ, Teunissen H,
813 Splinter E, Wijchers PJ, Krijger PH, de Laat W. 2015. CTCF Binding Polarity
814 Determines Chromatin Looping. *Molecular cell* **60**: 676-684.
815 Grant CE, Bailey TL, Noble WS. 2011. FIMO: scanning for occurrences of a given
816 motif. *Bioinformatics (Oxford, England)* **27**: 1017-1018.
817 Guo Y, Xu Q, Canzio D, Shou J, Li J, Gorkin DU, Jung I, Wu H, Zhai Y, Tang Y et al.
818 2015. CRISPR Inversion of CTCF Sites Alters Genome Topology and
819 Enhancer/Promoter Function. *Cell* **162**: 900-910.
820 Javierre BM, Burren OS, Wilder SP, Kreuzhuber R, Hill SM, Sewitz S, Cairns J,
821 Wingett SW, Varnai C, Thiecke MJ et al. 2016. Lineage-Specific Genome
822 Architecture Links Enhancers and Non-coding Disease Variants to Target Gene
823 Promoters. *Cell* **167**: 1369-1384 e1319.
824 Jeong M, Huang X, Zhang X, Su J, Shamim M, Bochkov I, Reyes J, Jung H, Heikamp

825 E, Presser Aiden A et al. 2017. A Cell Type-Specific Class of Chromatin Loops
826 Anchored at Large DNA Methylation Nadirs. *bioRxiv*.
827 Ji X, Dadon DB, Abraham BJ, Lee TI, Jaenisch R, Bradner JE, Young RA. 2015.
828 Chromatin proteomic profiling reveals novel proteins associated with
829 histone-marked genomic regions. *Proc Natl Acad Sci U S A* **112**: 3841-3846.
830 Jin C, Kato K, Chimura T, Yamasaki T, Nakade K, Murata T, Li H, Pan J, Zhao M, Sun
831 K et al. 2006. Regulation of histone acetylation and nucleosome assembly by
832 transcription factor JDP2. *Nat Struct Mol Biol* **13**: 331-338.
833 Jjingo D, Conley AB, Wang J, Marino-Ramirez L, Lunyak VV, Jordan IK. 2014.
834 Mammalian-wide interspersed repeat (MIR)-derived enhancers and the
835 regulation of human gene expression. *Mob DNA* **5**: 14.
836 John S, Sabo PJ, Thurman RE, Sung MH, Biddie SC, Johnson TA, Hager GL,
837 Stamatoyannopoulos JA. 2011. Chromatin accessibility pre-determines
838 glucocorticoid receptor binding patterns. *Nature genetics* **43**: 264-268.
839 Jolma A, Yan J, Whitington T, Toivonen J, Nitta KR, Rastas P, Morgunova E, Enge M,
840 Taipale M, Wei G et al. 2013. DNA-binding specificities of human transcription
841 factors. *Cell* **152**: 327-339.
842 Karolchik D, Barber GP, Casper J, Clawson H, Cline MS, Diekhans M, Dreszer TR,
843 Fujita PA, Guruvadoo L, Haeussler M et al. 2014. The UCSC Genome Browser
844 database: 2014 update. *Nucleic acids research* **42**: D764-770.
845 Katainen R, Dave K, Pitkanen E, Palin K, Kivioja T, Valimaki N, Gylfe AE, Ristolainen
846 H, Hanninen UA, Cajuso T et al. 2015. CTCF/cohesin-binding sites are
847 frequently mutated in cancer. *Nature genetics* **47**: 818-821.
848 Kheradpour P, Kellis M. 2014. Systematic discovery and characterization of regulatory
849 motifs in ENCODE TF binding experiments. *Nucleic acids research* **42**:
850 2976-2987.
851 Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler
852 transform. *Bioinformatics (Oxford, England)* **25**: 1754-1760.
853 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G,
854 Durbin R, Genome Project Data Processing S. 2009. The Sequence
855 Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)* **25**:
856 2078-2079.
857 McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM,

858 Bejerano G. 2010. GREAT improves functional interpretation of cis-regulatory
859 regions. *Nature biotechnology* **28**: 495-501.
860 Mumbach MR, Satpathy AT, Boyle EA, Dai C, Gowen BG, Cho SW, Nguyen ML,
861 Rubin AJ, Granja JM, Kazane KR et al. 2017. Enhancer connectome in primary
862 human cells identifies target genes of disease-associated DNA elements. *Nature*
863 *genetics* **49**: 1602-1612.
864 Newburger DE, Bulyk ML. 2009. UniPROBE: an online database of protein binding
865 microarray data on protein-DNA interactions. *Nucleic acids research* **37**:
866 D77-82.
867 Osato N. 2018. Characteristics of functional enrichment and gene expression level of
868 human putative transcriptional target genes. *BMC Genomics* **19**: 957.
869 Phanstiel DH, Van Bortle K, Spacek D, Hess GT, Shamim MS, Machol I, Love MI,
870 Aiden EL, Bassik MC, Snyder MP. 2017. Static and Dynamic DNA Loops form
871 AP-1-Bound Activation Hubs during Macrophage Development. *Molecular cell*
872 **67**: 1037-1048.e1036.
873 Portales-Casamar E, Thongjuea S, Kwon AT, Arenillas D, Zhao X, Valen E, Yusuf D,
874 Lenhard B, Wasserman WW, Sandelin A. 2010. JASPAR 2010: the greatly
875 expanded open-access database of transcription factor binding profiles. *Nucleic*
876 *acids research* **38**: D105-110.
877 Ramani V, Cusanovich DA, Hause RJ, Ma W, Qiu R, Deng X, Blau CA, Disteche CM,
878 Noble WS, Shendure J et al. 2016. Mapping 3D genome architecture through in
879 situ DNase Hi-C. *Nat Protoc* **11**: 2104-2121.
880 Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn
881 AL, Machol I, Omer AD, Lander ES et al. 2014. A 3D map of the human
882 genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**:
883 1665-1680.
884 Rebollo R, Romanish MT, Mager DL. 2012. Transposable elements: an abundant and
885 natural source of regulatory sequences for host genes. *Annu Rev Genet* **46**:
886 21-42.
887 Saeed S, Quintin J, Kerstens HH, Rao NA, Aghajanirefah A, Matarese F, Cheng SC,
888 Ratter J, Berentsen K, van der Ent MA et al. 2014. Epigenetic programming of
889 monocyte-to-macrophage differentiation and trained innate immunity. *Science*
890 *(New York, NY)* **345**: 1251086.

891 Schreiber J, Libbrecht M, Bilmes J, Noble W. 2017. Nucleotide sequence and DNaseI
892 sensitivity are predictive of 3D chromatin architecture. *bioRxiv*.
893 Tabuchi TM, Deplancke B, Osato N, Zhu LJ, Barrasa MI, Harrison MM, Horvitz HR,
894 Walhout AJ, Hagstrom KA. 2011. Chromosome-biased binding and gene
895 regulation by the Caenorhabditis elegans DRM complex. *PLoS Genet* **7**:
896 e1002074.
897 Wang J, Vicente-Garcia C, Seruggia D, Molto E, Fernandez-Minan A, Neto A, Lee E,
898 Gomez-Skarmeta JL, Montoliu L, Lunyak VV et al. 2015. MIR retrotransposon
899 sequences provide insulators to the human genome. *Proc Natl Acad Sci U S A*
900 **112**: E4428-4437.
901 Wang L, Wang S, Li W. 2012. RSeQC: quality control of RNA-seq experiments.
902 *Bioinformatics (Oxford, England)* **28**: 2184-2185.
903 Weintraub AS, Li CH, Zamudio AV, Sigova AA, Hannett NM, Day DS, Abraham BJ,
904 Cohen MA, Nabet B, Buckley DL et al. 2017. YY1 Is a Structural Regulator of
905 Enhancer-Promoter Loops. *Cell* **171**: 1573-1588 e1528.
906 Wingender E, Dietze P, Karas H, Knuppel R. 1996. TRANSFAC: a database on
907 transcription factors and their DNA binding sites. *Nucleic acids research* **24**:
908 238-241.
909 Xie Z, Hu S, Blackshaw S, Zhu H, Qian J. 2010. hPDI: a database of experimental
910 human protein-DNA interactions. *Bioinformatics (Oxford, England)* **26**:
911 287-289.
912 Zhang H, Li F, Jia Y, Xu B, Zhang Y, Li X, Zhang Z. 2017. Characteristic arrangement
913 of nucleosomes is predictive of chromatin interactions at kilobase resolution.
914 *Nucleic acids research* **45**: 12739-12751.
915 Zhang Y, An L, Xu J, Zhang B, Zheng WJ, Hu M, Tang J, Yue F. 2018. Enhancing Hi-C
916 data resolution with deep convolutional neural network HiCPlus. *Nat Commun*
917 **9**: 750.
918 Zhao Y, Stormo GD. 2011. Quantitative analysis demonstrates most transcription factors
919 require only simple models of specificity. *Nature biotechnology* **29**: 480-483.

920