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. Though the number of chromatin interactions was about seventy thousand in Hi-C at 50kb resolution, about twenty millions of chromatin interactions were recently identified by HiChIP at 5kb resolution. It has been reported that long range of chromatin interactions tends to include less CTCF at their anchors. It is still unclear what proteins are associated with chromatin interactions.

To find DNA binding motif sequences of transcription factors (TF), such as CTCF, and repeat DNA sequences affecting the interaction between enhancers and promoters of genes and their expression, first I predicted TF bound in enhancers and promoters using DNA motif sequences of TF and experimental data of open chromatin regions in monocytes and other cell types, which were obtained from public and commercial databases. Second, transcriptional target genes of each TF were predicted based on enhancer-promoter association (EPA). EPA was shortened at the genomic locations of FR or reverse-forward (RF) orientation of DNA motif sequence of a TF,

which were supposed to be at chromatin interaction anchors and acted as insulator sites like CTCF. Then, the expression levels of the transcriptional target genes predicted based on the EPA were compared with those predicted from only promoters.

Total 287 biased orientation of DNA motifs (159 FR and 153 RF orientation, the reverse complement sequences of some DNA motifs were also registered in databases, so the total number was smaller than the number of FR and RF) affected the expression level of putative transcriptional target genes significantly in $CD14^+$ monocytes of four people in common. The same analysis was conducted in $CD4^+$ T cells of four people. DNA motif sequences of CTCF, cohesin and other transcription factors involved in chromatin interactions were found to be a biased orientation. Transposon sequences, which are known to be involved in insulators and enhancers, showed a biased orientation. The biased orientation of DNA motif sequences tended to be co-localized in the same open chromatin regions. Moreover, for ~86% of FR and RF orientations of DNA motif sequences, EPI predicted from EPA that were shortened at the genomic locations of the biased orientation of DNA motif sequence were overlapped with chromatin interaction data (Hi-C and HiChIP) significantly more than other types of EPAs.

Keywords: transcriptional target genes, gene expression, transcription factors, enhancer, enhancer-promoter interactions, chromatin interactions, CTCF, cohesin, open chromatin regions, co-location of transcription factors, homodimer, heterodimer, complex

Background

Chromatin interactions have important roles for enhancer-promoter interactions (EPI) and regulating the transcription of genes. CTCF and cohesin proteins are located 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 Hi-C experiments include DNA binding motif sequences of CTCF. Though the number of chromatin interactions was about seventy thousand in Hi-C at 50kb resolution (Javierre et al. 2016), about twenty millions of chromatin interactions were recently identified by HiChIP at 5kb resolution (Mumbach et al. 2017). However, it has been reported that long range of chromatin interactions tends to include less CTCF at their anchors (Jeong et al. 2017). Other DNA binding proteins such as ZNF143, YY1, and SMARCA4 (BRG1) are found to be associated with chromatin interactions and EPI (Bailey et al. 2015; Barutcu et al. 2016; Weintraub et al. 2017). CTCF, cohesin, ZNF143, YY1 and SMARCA4 have other biological functions as well as chromatin interactions and EPI. The DNA binding motif sequences of the transcription factors (TF) are found in open chromatin regions near transcriptional start sites (TSS) as well as chromatin interaction anchors.

DNA binding motif sequence of ZNF143 was enriched at both chromatin interaction anchors. ZNF143's correlation with the CTCF-cohesin cluster relies on its

77 weakest binding sites, found primarily at distal regulatory elements defined by the
78 'CTCF-rich' chromatin state. The strongest ZNF143-binding sites map to promoters
79 bound by RNA polymerase II (POL2) and other promoter-associated factors, such as the
80 TATA-binding protein (TBP) and the TBP-associated protein, together forming a
81 'promoter' cluster (Bailey et al. 2015).

82 DNA binding motif sequence of YY1 does not seem to be enriched at both
83 chromatin interaction anchors (Z-score < 2), whereas DNA binding motif sequence of
84 ZNF143 is significantly enriched (Z-score > 7; Bailey et al. 2015 Figure 2a). In the
85 analysis of YY1, to identify a protein factor that might contribute to EPI, (Ji et al. 2015)
86 performed chromatin immune precipitation with mass spectrometry (ChIP-MS), using
87 antibodies directed toward histones with modifications characteristic of enhancer and
88 promoter chromatin (H3K27ac and H3K4me3, respectively). Of 26 transcription factors
89 that occupy both enhancers and promoters, four are essential based on a CRISPR
90 cell-essentiality screen and two (CTCF, YY1) are expressed in >90% of tissues
91 examined (Weintraub et al. 2017). These analyses started from the analysis of histone
92 modifications of enhancer and promoter marks rather than chromatin interactions. Other
93 protein factors associated with chromatin interactions may be found from other
94 researches.

95 As computational approaches, machine-learning analyses to predict chromatin
96 interactions have been proposed (Schreiber et al. 2017; Zhang et al. 2017). However,
97 they were not intended to find DNA motif sequences of TF affecting chromatin
98 interactions, EPI, and the expression level of transcriptional target genes, which were

99 examined in this study.

100 DNA binding proteins involved in chromatin interactions are supposed to affect
101 the transcription of genes in the loops formed by chromatin interactions. In my previous
102 analysis, the expression level of human putative transcriptional target genes was
103 affected, according to the criteria of enhancer-promoter association (EPA) (Fig. 1B;
104 (Osato 2018)). EPI were predicted based on EPA shortened at the genomic locations of
105 FR orientation of CTCF binding sites, and transcriptional target genes of each TF bound
106 in enhancers and promoters were predicted based on the EPI. The EPA affected the
107 expression levels of putative transcriptional target genes the most among three types of
108 EPAs, compared with the expression levels of transcriptional target genes predicted
109 from only promoters (Fig. 2). The expression levels tended to be increased in
110 monocytes and $CD4^+$ T cells, implying that enhancers activated the transcription of
111 genes, and decreased in ES and iPS cells, implying that enhancers repressed the
112 transcription of genes. These analyses suggested that enhancers affected the
113 transcription of genes significantly, when EPI were predicted properly. Other DNA
114 binding proteins involved in chromatin interactions, as well as CTCF, may locate at
115 chromatin interaction anchors with a pair of biased orientation of DNA binding motif
116 sequences, affecting the expression level of putative transcriptional target genes in the
117 loops formed by the chromatin interactions. As experimental issues of the analyses of
118 chromatin interactions, chromatin interaction data are changed, according to
119 experimental techniques, depth of DNA sequencing, and even replication sets of the
120 same cell type. Chromatin interaction data may not be saturated enough to cover all

121 chromatin interactions. Supposing these characters of DNA binding proteins associated
122 with chromatin interactions and avoiding experimental issues of the analyses of
123 chromatin interactions, here I searched for DNA motif sequences of TF and repeat DNA
124 sequences, affecting EPI and the expression level of putative transcriptional target genes
125 in $CD14^+$ monocytes and $CD4^+$ T cells of four people and other cell types without using
126 chromatin interaction data. Then, putative EPI were compared with chromatin
127 interaction data.
128

129 ## Results

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

131 Transcription factor binding sites (TFBS) were predicted using open chromatin
132 regions and DNA motif sequences of transcription factors (TF) collected from various
133 databases and journal papers (see Methods). Transcriptional target genes were predicted
134 using TFBS in promoters and enhancer-promoter association (EPA) shortened at the
135 genomic locations of DNA binding motif sequence of a TF acting as insulator such as
136 CTCF and cohesin (RAD21 and SMC3) (Fig. 1B). To find DNA motif sequences of TF
137 acting as insulator, other than CTCF and cohesin, and repeat DNA sequences affecting
138 the expression level of genes, EPI were predicted based on EPA shortened at the
139 genomic locations of the DNA motif sequence of a TF or a repeat DNA sequence, and
140 transcriptional target genes of each TF bound in enhancers and promoters were
141 predicted based on the EPI. The expression levels of the putative transcriptional target
142 genes were compared with those predicted from promoters using Mann Whiteney U test,

143 two-sided (p-value < 0.05) (Fig. 2A). The number of TF showing a significant
144 difference of expression level of their putative transcriptional target genes was counted.
145 To examine whether the orientation of a DNA motif sequence, which is supposed to act
146 as insulator sites and shorten EPA, affected the number of TF showing a significant
147 difference of expression level of their putative transcriptional target genes, the number
148 of the TF was compared among forward-reverse (FR), reverse-forward (RF), and any
149 orientation (i.e. without considering orientation) of a DNA motif sequence shortening
150 EPA, using chi-square test (p-value < 0.05). To avoid missing DNA motif sequences
151 showing a relatively weak statistical significance by multiple testing collection, the
152 above analyses were conducted using monocytes of four people independently, and
153 DNA motif sequences found in monocytes of four people in common were selected.
154 Total 287 of biased (159 FR and 153 RF) orientation of DNA binding motif sequences
155 of TF were found in monocytes of four people in common, whereas only one any
156 orientation of DNA binding motif sequence was found in monocytes of four people in
157 common (Fig. 2B; Table 1; Supplemental Table S1). FR orientation of DNA motif
158 sequences included CTCF, cohesin (RAD21 and SMC3), ZNF143 and YY1, which are
159 associated with chromatin interactions and EPI. SMARCA4 (BRG1) is associated with
160 topologically associated domain (TAD), which is higher-order chromatin organization.
161 The DNA binding motif sequence of SMARCA4 was not registered in the databases
162 used in this study. FR orientation of DNA motif sequences included SMARCC2, a
163 member of the same SWI/SNF family of proteins as SMARCA4.

164 The same analysis was conducted using DNase-seq data of $CD4^+$ T cells of four

165 people. Total 265 of biased (158 FR and 139 RF) orientation of DNA binding motif
166 sequences of TF were found in T cells of four people in common, whereas only four
167 any orientation (i.e. without considering orientation) of DNA binding motif sequences
168 were found in T cells of four people in common (Supplemental Fig. S1 and
169 Supplemental Table S2). Biased orientation of DNA motif sequences in T cells included
170 CTCF, cohesin (RAD21 and SMC3), ZNF143, YY1, and SMARC. Biased orientation
171 of DNA motif sequences in T cells and monocytes (using H3K27ac histone
172 modification marks, which is described in the next paragraph) included JUNDM2
173 (JDP2), which is involved in histone-chaperone activity, promoting nucleosome, and
174 inhibition of histone acetylation (Jin et al. 2006). JDP2 forms a homodimer or
175 heterodimer with various TF (https://en.wikipedia.org/wiki/Jun_dimerization_protein).
176 Among 287, 61 of biased orientation of DNA binding motif sequences of TF were
177 found in both monocytes and T cells in common (Supplemental Table S5). For each
178 orientation, 31 FR and 34 RF orientation of DNA binding motif sequences of TF were
179 found in both monocytes and T cells in common. Without considering the difference of
180 orientation of DNA binding motif sequences, 84 of biased orientation of DNA binding
181 motif sequences of TF were found in both monocytes and T cells. As a reason for the
182 increase of the number (84) from 61, a TF or an isoform of the same TF may bind to a
183 different DNA binding motif sequence according to cell types and/or in the same cell
184 type. About 50% or more of alternative splicing isoforms are differently expressed
185 among tissues, indicating that most alternative splicing is subject to tissue-specific
186 regulation (Wang et al. 2008) (Chen and Manley 2009) (Das et al. 2007). The same TF

187 has several DNA binding motif sequences and in some cases one of the motif sequences
188 is almost the same as the reverse complement sequence of another motif sequence of the
189 same TF. For example, RAD21 had both FR and RF orientation of DNA motif
190 sequences, but the number of the FR orientation of DNA motif sequence was relatively
191 small in the genome, and the RF orientation of DNA motif sequence was frequently
192 observed and co-localized with CTCF. I previously found that a complex of TF would
193 bind to a slightly different DNA binding motif sequence from the combination of DNA
194 binding motif sequences of TF composing the complex in *C. elegans* (Tabuchi et al.
195 2011). From another viewpoint of this study, the expression level of putative
196 transcription target genes of some TF would be different, depending on the genomic
197 locations (enhancers or promoters) of DNA binding motif sequences of the TF in
198 monocytes and T cells of four people.

199 Moreover, using open chromatin regions overlapped with H3K27ac histone
200 modification marks known as enhancer and promoter marks, the same analyses were
201 performed in monocytes and T cells. H3K27ac histone modification marks were used in
202 the analysis of EPI, but were not used in the analysis of TF as insulator like CTCF and
203 cohesin in this study, since new biased orientation of DNA motif sequences were found
204 in this criterion. When H3K27ac histone modification marks were used in the analysis
205 of TF as insulator like CTCF and cohesin, the number of biased orientation of DNA
206 motif sequences was decreased. Total 241 of biased (141 FR and 120 RF) orientation of
207 DNA binding motif sequences of TF were found in monocytes of four people in
208 common, whereas only one any orientation of DNA binding motif sequence was found

209 (Supplemental Table S3). Though the number of biased orientation of DNA motif
210 sequences was reduced, JDP2 was found in monocytes as well as T cells. CTCF,
211 RAD21, SMC3, ZNF143, YY1, and SMARCC2 were also found. For T cells using
212 H3K27ac histone modification marks, total 195 of biased (107 FR and 110 RF)
213 orientation of DNA binding motif sequences of TF were found in T cells of four people
214 in common, whereas only six any orientation of DNA binding motif sequences were
215 found (Supplemental Table S4). Though the number of biased orientation of DNA motif
216 sequences was reduced, CTCF, RAD21, SMC3, YY1, and SMARCC2 were found.
217 Scores of CTCF, RAD21, and SMC3 were increased compared with the result of T cells
218 without using H3K27ac histone modification marks, and they were ranked in the top six.
219 As summary of the results with and without H3K27ac histone modification marks, total
220 386 of biased (223 FR and 214 RF) orientation of DNA motif sequences were found in
221 monocytes of four people in common. Total 348 of biased (202 FR and 198 RF)
222 orientation of DNA motif sequences were found in T cells of four people in common.
223 Total number of these results in monocytes and T cells was 590 biased (368 FR and 359
224 RF) orientation of DNA motif sequences. Biased orientation of DNA motif sequences
225 found in both monocytes and T cells were listed in Supplemental Table S5.

226 To examine whether the biased orientation of DNA binding motif sequences of
227 TF were observed in other cell types, the same analyses were conducted in other cell
228 types. However, for other cell types, experimental data of one sample were available in
229 ENCODE database, so the analyses of DNA motif sequences were performed by
230 comparing with the result in monocytes of four people. Among the biased orientation of

231 DNA binding motif sequences found in monocytes, 69, 82, 56, and 53 DNA binding
232 motif sequences were also observed in H1-hESC, iPS, Huvec and MCF-7 respectively,
233 including CTCF and cohesin (RAD21 and SMC3) (Table 2; Supplemental Table S6).
234 The scores of DNA binding motif sequences were the highest in monocytes, and the
235 other cell types showed lower scores. The results of the analysis of DNA motif
236 sequences in $CD20^+$ B cells and macrophages did not include CTCF and cohesin,
237 because these analyses can be utilized in cells where the expression level of putative
238 transcriptional target genes of each TF show a significant difference between promoters
239 and EPA shortened at the genomic locations of a DNA motif sequence acting as
240 insulator sites. Some experimental data of a cell did not show a significant difference
241 between promoters and the EPA (Osato 2018).

242 Instead of DNA binding motif sequences of TF, repeat DNA sequences were also
243 examined. The expression levels of transcriptional target genes of each TF predicted
244 based on EPA that were shortened at the genomic locations of a repeat DNA sequence
245 were compared with those predicted from promoters. Three RF orientation of repeat
246 DNA sequences showed a significant difference of expression level of putative
247 transcriptional target genes in monocytes of four people in common (Table 3). Among
248 them, LTR16C repeat DNA sequence was observed in iPS and H1-hESC with enough
249 statistical significance considering multiple tests (p-value $< 10^{-7}$). The same as $CD14^+$
250 monocytes, biased orientation of repeat DNA sequences were examined in $CD4^+$ T cells.
251 Three FR and two RF orientation of repeat DNA sequences showed a significant
252 difference of expression level of putative transcriptional target genes in T cells of four

253 people in common (Supplemental Table S7). MIRb and MIR3 were also found in the
254 analysis using open chromatin regions overlapped with H3K27ac histone modification
255 marks, which are enhancer and promoter marks. MIR and other transposon sequences
256 are known to act as insulators and enhancers (Bejerano et al. 2006; Rebollo et al. 2012;
257 de Souza et al. 2013; Jjingo et al. 2014; Wang et al. 2015).
258
259 **Co-location of biased orientation of DNA motif sequences**
260 To examine the association of 287 biased (FR and RF) orientation of DNA
261 binding motif sequences of TF, co-location of the DNA binding motif sequences in
262 open chromatin regions was analyzed in monocytes. The number of open chromatin
263 regions with the same pairs of DNA binding motif sequences was counted, and when
264 the pairs of DNA binding motif sequences were enriched with statistical significance
265 (chi-square test, p-value $< 1.0 \times 10^{-10}$), they were listed (Table 4; Supplemental Table
266 S8; Fig. 3A). Open chromatin regions overlapped with histone modification of enhancer
267 and promoter marks (H3K27ac) (total 26,095 regions) showed a larger number of
268 enriched pairs of DNA motifs than all open chromatin regions (Table 4; Supplemental
269 Table S8; Fig. 3A). H3K27ac is known to be enriched at chromatin interaction anchors
270 (Phanstiel et al. 2017). As already known, CTCF was found with cohesin such as
271 RAD21 and SMC3 (Table 4; Fig. 3A). Top 30 pairs of FR and RF orientations of DNA
272 motifs co-occupied in the same open chromatin regions were shown (Table 4). Total
273 number of pairs of DNA motifs was 492, consisting of 115 unique DNA motifs, when
274 the pair of DNA motifs were observed in more than 80% of the number of open

275 chromatin regions with the DNA motifs. Biased orientation of DNA binding motif
276 sequences of TF tended to be co-localized in the same open chromatin regions.

277 To examine the association of 265 biased orientation of DNA binding motif
278 sequences of TF in $CD4^+$ T cells, co-location of the DNA binding motif sequences in
279 open chromatin regions was analyzed. Top 30 pairs of FR and RF orientations of DNA
280 motifs co-occupied in the same open chromatin regions were shown (Supplemental
281 Table S9; Supplemental Fig. S2). Total number of pairs of DNA motifs was 332,
282 consisting of 118 unique DNA motifs, when the pair of DNA motifs were observed in
283 more than 60% of the number of open chromatin regions with the DNA motifs
284 (chi-square test, p-value $< 1.0 \times 10^{-10}$). Among them, 12 pairs of DNA motif
285 sequences including a pair of CTCF, SMC3, and RAD21 in T cells were found in
286 monocytes in common (Supplemental Table S9).

287

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

289 CTCF and cohesin are found in chromatin interaction anchors and are considered
290 to form homodimers. TF are also known to form heterodimers with another TF. In this
291 study, if the DNA binding motif sequence of only the mate to a pair of TF was found in
292 EPA, EPA was shortened at one side, which is the genomic location of the DNA
293 binding motif sequence of the mate to the pair, and transcriptional target genes were
294 predicted using the EPA shortened at the side. In this analysis, the mate to both
295 heterodimer and homodimer of TF can be used to examine the effect on the expression
296 level of transcriptional target genes predicted based on the EPA shortened at one side.

297 To examine whether biased orientation of DNA motif sequences tend to be found as a
298 pair of the same TF or different TF in upstream and downstream of genes, the
299 enrichment of pairs of biased orientation of DNA motif sequences was investigated. The
300 number of genes with a pair of FR or RF orientation of DNA motif sequences was
301 counted using all pairs of the DNA motif sequences, and statistical tests (chi-square
302 test) were conducted. The total number of genes (transcripts) with a biased orientation
303 of DNA motif sequence in their upstream or/and downstream was about 20,000. This is
304 almost the same way of analysis as co-location of biased orientation of DNA motif
305 sequences, and the analysis here used the number of genes instead of open chromatin
306 regions. Fig. 3B and Supplemental Fig. S3 showed that not only pairs of the same TF,
307 but also pairs of different TF were found in upstream and downstream of genes. DNA
308 motif sequences of CTCF and cohesin (RAD21 and SMC3) were enriched near the
309 same genes, but cohesin tended to be found with another TF near the same genes as well.
310 The number of some pairs of biased orientation of DNA motif sequences in genome
311 sequences was small, though they showed statistical significance (Supplemental
312 material 2). The result of the analysis seems to be different between monocytes and T
313 cells and between using H3K27ac histone modification marks and not using them
314 (Supplemental Fig. S3). Some of these TF would form homodimers or heterodimers
315 directory or indirectly composing a complex structure, contributing to make chromatin
316 interactions (Fig. 5).
317
318

319 **Comparison with chromatin interaction data**

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

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

354 After the analysis of $CD14^+$ monocytes, to utilize HiChIP chromatin interaction
355 data in $CD4^+$ T cells, the same analysis for $CD14^+$ monocytes was performed using
356 DNase-seq data of four donors in $CD4^+$ T cells (Mumbach et al. 2017). EPI predicted
357 based on EPA were compared with three replications (B2T1, B2T2, and B3T1) of
358 HiChIP chromatin interaction data in $CD4^+$ T cells respectively. The resolutions of
359 HiChIP chromatin interaction data and EPI were adjusted to 5kb. EPI were predicted
360 based on the three types of EPA in the same way as $CD14^+$ monocytes using top 60%
361 expression level of all transcripts (genes) excluding transcripts not expressed in T cells.
362 The criteria of the analysis were determined to include known DNA motif sequences

363 involved in chromatin interactions such as CTCF and cohesin in the result, and the
364 result was consistent with that using Hi-C chromatin interaction data. EPA (iii) showed
365 the highest ratio of EPI overlapped with chromatin interactions (HiChIP) using DNA
366 binding motif sequences of CTCF and cohesin (RAD21 and SMC3), compared with the
367 other two types of EPA (i) and (ii) (n = 4, binomial test, two-sided, 95% confidence
368 interval) (Fig. 4). Total 88 biased orientation (53 FR and 42 RF) of DNA motif
369 sequences, which included CTCF, cohesin (RAD21 and SMC3), and SMARC in three
370 replications (B2T1, B2T2, and B3T1) and ZNF143 in two replications (B2T2 and
371 B3T1), showed a significantly higher ratio of EPI overlapped with HiChIP chromatin
372 interactions (more than 1,000 counts for each interaction) than the other types of EPAs
373 in T cells (Table 5). When comparing EPI predicted based on only EPA (i) and (iii) with
374 the chromatin interactions, total 247 biased orientation (148 FR and 130 RF) of DNA
375 motif sequences showed a significantly higher ratio of EPI predicted based on EPA (iii)
376 overlapped with the chromatin interactions than EPI predicted based on EPA (i) (Table
377 5; Supplemental material 2). As expected, the number of EPI overlapped with
378 chromatin interactions (HiChIP) was increased, compared with Hi-C chromatin
379 interactions. Most of biased orientation of DNA motif sequences (86%) were found to
380 be correlated with chromatin interactions, when comparing EPI predicted based on EPA
381 (i) and (iii) with HiChIP chromatin interactions.

382 Moreover, to examine the enhancer activity of EPI, the distribution of expression
383 level of putative target genes of EPI was compared between EPI overlapped with
384 HiChIP chromatin interactions and EPI not overlapped with them. Though the target

385 genes of EPI were selected from top 60% expression level of all transcripts (genes)
386 excluding transcripts not expressed in T cells, target genes of EPI overlapped with
387 chromatin interactions showed a significantly higher expression level than EPI not
388 overlapped with them, implying that EPI overlapped with chromatin interactions
389 activated the expression of target genes in T cells. Almost all (99.5%) FR and RF
390 orientation of DNA motifs showed a significantly higher expression level of putative
391 target genes of EPI overlapped with chromatin interactions than EPI not overlapped.
392 When a biased orientation of DNA motif showed a significantly higher expression level
393 of putative target genes of EPI overlapped with chromatin interactions than EPI not
394 overlapped, '1' was marked with in the tables of the comparison of EPI and HiChIP
395 chromatin interactions in Supplemental material 2. When a DNA motif showed a
396 significantly lower expression level, '-1' was marked with, however, it was not
397 observed in this analysis. When there was not significant difference of expression level,
398 '0' was marked with.

399 Though the ratios of EPI overlapped with chromatin interactions were increased
400 by using many chromatin interaction data including lower score and count of chromatin
401 interactions (a cutoff score of CHiCAGO tool > 1 for Hi-C and more than 1,000 counts
402 for HiChIP), the ratios of EPI overlapped with chromatin interactions showed the same
403 tendency among the three types of EPAs. The ratio of EPI overlapped with Hi-C
404 chromatin interactions was increased using H3K27ac marks in both monocytes and T
405 cells. The ratio of EPI overlapped with HiChIP chromatin interactions was also
406 increased using H3K27ac marks. Chromatin interaction data were obtained from

407 different samples from DNase-seq, open chromatin regions in $CD4^+$ T cells, so
408 individual differences seemed to be large from the results of this analysis, and
409 (Mumbach et al. 2017) suggested that individual differences of chromatin interactions
410 were larger than those of open chromatin regions. ATAC-seq data, open chromatin
411 regions were available in $CD4^+$ T cells in the paper, however, when using ATAC-seq
412 data, the result of the analysis of biased orientation of DNA motif sequences was
413 different from DNase-seq data, and not included a part of CTCF and cohesin. Thus,
414 DNase-seq data collected from ENCODE and Blueprint projects were employed in this
415 study.
416

417 **Discussion**

418 To find DNA motif sequences of transcription factors (TF) and repeat DNA
419 sequences affecting the expression level of human putative transcriptional target genes,
420 the DNA motif sequences were searched from open chromatin regions of monocytes of
421 four people. Total 287 biased [159 forward-reverse (FR) and 153 reverse-forward (RF)]
422 orientation of DNA motif sequences of TF were found in monocytes of four people in
423 common, whereas only one any orientation (i.e. without considering orientation) of
424 DNA motif sequence of TF was found to affect the expression level of putative
425 transcriptional target genes, suggesting that enhancer-promoter association (EPA)
426 shortened at the genomic locations of FR or RF orientation of the DNA motif sequence
427 of a TF or a repeat DNA sequence is an important character for the prediction of
428 enhancer-promoter interactions (EPI) and the transcriptional regulation of genes.

429 When DNA motif sequences were searched from monocytes of one person, a
430 larger number of biased orientation of DNA motif sequences affecting the expression
431 level of human putative transcriptional target genes were found. When the number of
432 donors, from which experimental data were obtained, was increased, the number of
433 DNA motif sequences found in all people in common decreased and in some cases,
434 known transcription factors involved in chromatin interactions such as CTCF and
435 cohesin (RAD21 and SMC3) were not identified by statistical tests. This would be
436 caused by individual difference of the same cell type, low quality of experimental data,
437 and experimental errors. Moreover, though FR orientation of DNA binding motif
438 sequences of CTCF and cohesin is frequently observed at chromatin interaction anchors,
439 the percentage of FR orientation is not 100, and other orientations of the DNA binding
440 motif sequences are also observed. Though DNA binding motif sequences of CTCF and
441 cohesin are found in various open chromatin regions, DNA binding motif sequences of
442 some TF would be observed less frequently in open chromatin regions. The analyses of
443 experimental data of a number of people would avoid missing relatively weak statistical
444 significance of DNA motif sequences of TF in experimental data of each person by
445 multiple testing correction of thousands of statistical tests. A DNA motif sequence was
446 found with p-value < 0.05 in experimental data of one person and the DNA motif
447 sequence found in the same cell type of four people in common would have p-value $<$
448 $0.05^4 = 6.25 \times 10^{-6}$. Actually, DNA motif sequences with p-value slightly less than
449 0.05 in monocytes of one person were observed in monocytes of four people in
450 common.

451 EPI were compared with chromatin interactions (Hi-C) in monocytes. EPAs
452 shortened at the genomic locations of DNA binding motif sequences of CTCF and
453 cohesin (RAD21 and SMC3) showed a significant difference of the ratios of EPI
454 overlapped with chromatin interactions, according to three types of EPAs (see Methods).
455 Using open chromatin regions overlapped with ChIP-seq experimental data of histone
456 modification of an enhancer mark (H3K27ac), the ratio of EPI not overlapped with
457 Hi-C was reduced. (Phanstiel et al. 2017) also reported that there was an especially
458 strong enrichment for loops with H3K27 acetylation peaks at both ends (Fisher's Exact
459 Test, $p = 1.4 \times 10^{-27}$). However, the total number of EPI overlapped with chromatin
460 interactions was also reduced using H3K27ac peaks, so more chromatin interaction data
461 would be needed to obtain reliable results in this analysis. As an issue of experimental
462 data, data for chromatin interactions and open chromatin regions were came from
463 different samples and donors, so individual differences would exist in the data.
464 Moreover, the resolution of chromatin interaction data used in monocytes was about
465 50kb, thus the number of chromatin interactions was relatively small (72,284 at 50kb
466 resolution with a cutoff score of CHiCAGO tool > 1 and 16,501 with a cutoff score of
467 CHiCAGO tool > 5). EPI predicted based on EPA shortened at the genomic locations of
468 DNA binding motif sequence of TF that were found in various open chromatin regions
469 such as CTCF and cohesin (RAD21 and SMC3) tended to be overlapped with a larger
470 number of chromatin interactions than TF less frequently observed in open chromatin
471 regions. Therefore, to examine the difference of the numbers of EPI overlapped with
472 chromatin interactions, according to the three types of EPAs, the number of chromatin

473 interactions should be large enough.

474 As HiChIP chromatin interaction data were available in $CD4^+$ T cells, biased
475 orientation of DNA motif sequences of TF were examined in T cells using DNase-seq
476 data of four people. The resolutions of chromatin interactions and EPI were adjusted to
477 5kb by fragmentation of genome sequences. In monocytes, the resolution of Hi-C
478 chromatin interaction data was converted by extending anchor regions of chromatin
479 interactions to 50kb length and merging the chromatin interactions overlapped with
480 each other. Fragmentation of genome sequences may affect the classification of
481 chromatin interactions of which anchors are located near the border of a fragment, but
482 the number of chromatin interactions would not be decreased, compared with merging
483 chromatin interactions. The number of HiChIP chromatin interactions was 19,926,360
484 at 5kb resolution, 666,149 at 5kb resolution with chromatin interactions (more than
485 1,000 counts for each interaction), and 78,209 at 5kb resolution with chromatin
486 interactions (more than 6,000 counts for each interaction). As expected, the number of
487 EPI overlapped with chromatin interactions was increased, and 33 – 86% of biased
488 orientation of DNA motif sequences of TF showed a statistical significance in EPI
489 predicted based on EPA shortened at the genomic locations of the DNA motif sequence,
490 compared with the other types of EPAs or EPA not shortened. False positive predictions
491 of EPI would be decreased by using H3K27ac marks and other features. The ratio of
492 EPI overlapped with Hi-C chromatin interactions was increased using H3K27ac marks
493 in both monocytes and T cells. The ratio of EPI overlapped with HiChIP chromatin
494 interactions was also increased using H3K27ac marks. However, the number of biased

495 orientation of DNA motif sequences showing a higher ratio of EPI overlapped with
496 HiChIP chromatin interactions than the other types of EPAs was decreased using
497 H3K27ac marks (Supplemental material 2).

498 Though the number of known DNA binding proteins and TF involved in
499 chromatin interactions would be less than 30 or so, about 300 DNA binding proteins
500 showed enrichment of biased orientation of their DNA binding sequences, and known
501 DNA binding proteins involved in chromatin interactions were found in the biased
502 orientation of DNA binding sequences. JUNDM2 (JDP2) found in T cells and
503 monocytes is not known DNA binding proteins involved in chromatin interactions.
504 JDP2 is involved in histone and nucleosome-related functions such as
505 histone-chaperone activity, promoting nucleosome, and inhibition of histone acetylation
506 (Jin et al. 2006). JDP2 forms a homodimer or heterodimer with various TF
507 (https://en.wikipedia.org/wiki/Jun_dimerization_protein). Thus, the main function of
508 JDP2 may not be associated with chromatin interactions, but it forms a homodimer or
509 heterodimer with TF, potentially forming a chromatin interaction through the binding of
510 JDP2 and another TF, like chromatin interactions formed by YY1 (Weintraub et al.
511 2017). In the analysis of HiChIP chromatin interaction data, JDP2 was in the list of
512 biased orientation of DNA motif sequences showing a higher ratio of EPI overlapped
513 with HiChIP chromatin interactions than the other types of EPAs, using chromatin
514 interactions with more than 6,000 counts for each interaction (Supplemental material 2).
515 When forming a homodimer or heterodimer with another TF, TF may bind to genome
516 DNA with a specific orientation of their DNA binding sequences. From the analysis of

517 biased orientation of DNA motif sequences of TF, TF forming heterodimer would also
518 be found. If the DNA binding motif sequence of only the mate to a pair of TF was
519 found in EPA, EPA was shortened at one side, which is the genomic location of the
520 DNA binding motif sequence of the mate to the pair, and transcriptional target genes
521 were predicted using the EPA shortened at the side. In this analysis, the mate to both
522 heterodimer and homodimer of TF can be used to examine the effect on the expression
523 level of transcriptional target genes predicted based on the EPA shortened at one side.
524 For heterodimer, biased orientation of DNA motif sequences may also be found in
525 forward-forward or reverse-reverse orientation.

526 I reported that the ratio of the number of functional enrichment of putative
527 transcriptional target genes in the total number of putative transcriptional target genes
528 changed according to the types of EPAs, in the same way as gene expression level
529 (Osato 2018). JDP2 showed a higher ratio of functional enrichment of putative
530 transcriptional target genes using EPA shortened at the genomic locations of FR
531 orientation of the DNA binding motif sequence, compared with the other types of EPAs
532 (RF or any orientation) in monocytes and T cells, among monocytes, T cells and $CD20^+$
533 B cells. While the genome-wide analysis of functional enrichment of putative
534 transcriptional target genes takes computational time, the genome-wide analysis of gene
535 expression level is finished with less computational time. The analysis of gene
536 expression level can be used for a cell type and experimental data where the expression
537 levels of transcriptional target genes predicted based on EPA are significantly different
538 from those predicted from only promoter.

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

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

547

548 ## Methods

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

550 To examine transcriptional regulatory target genes of transcription factors (TF),
551 bed files of hg38 of Blueprint DNase-seq data for $CD14^+$ monocytes of four donors
552 (EGAD00001002286; Donor ID: C0010K, C0011I, C001UY, C005PS) were obtained
553 from Blueprint project web site (http://dcc.blueprint-epigenome.eu/#/home), and the bed
554 files of hg38 were converted into those of hg19 using Batch Coordinate Conversion
555 (liftOver) web site (https://genome.ucsc.edu/util.html). Bed files of hg19 of ENCODE
556 H1-hESC (GSM816632; UCSC Accession: wgEncodeEH000556), iPSC (GSM816642;
557 UCSC Accession: wgEncodeEH001110), HUVEC (GSM1014528; UCSC Accession:
558 wgEncodeEH002460), and MCF-7 (GSM816627; UCSC Accession:
559 wgEncodeEH000579) were obtained from the ENCODE websites
560 (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeUwDgf/;

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

562 As high resolution of chromatin interaction data using HiChIP became available
563 in $CD4^+$ T cells, to promote the same analysis of $CD14^+$ monocytes in $CD4^+$ T cells,
564 DNase-seq data of four donors were obtained from a public database and Blueprint
565 projects web sites. DNase-seq data of only one donor was available in Blueprint project
566 for $CD4^+$ T cells (Donor ID: S008H1), and three other DNase-seq data were obtained
567 from NCBI Gene Expression Omnibus (GEO) database. Though the peak calling of
568 DNase-seq data available in GEO database was different from other DNase-seq data in
569 ENCODE and Blueprint projects where 150bp length of peaks were usually predicted
570 using HotSpot (John et al. 2011), FASTQ files of DNase-seq data were downloaded
571 from NCBI Sequence Read Archive (SRA) database (SRR097566, SRR097618, and
572 SRR171574). Read sequences of the FASTQ files were aligned to the hg19 version of
573 the human genome reference using BWA (Li and Durbin 2009), and the BAM files
574 generated by BWA were converted into SAM files, sorted, and indexed using Samtools
575 (Li et al. 2009). Peaks of the DNase-seq data were predicted using HotSpot-4.1.1.

576 To identify transcription factor binding sites (TFBS) from the DNase-seq data,
577 TRANSFAC (2013.2), JASPAR (2010), UniPROBE, BEEML-PBM, high-throughput
578 SELEX, Human Protein-DNA Interactome, and transcription factor binding sequences
579 of ENCODE ChIP-seq data were used (Wingender et al. 1996; Newburger and Bulyk
580 2009; Portales-Casamar et al. 2010; Xie et al. 2010; Zhao and Stormo 2011; Jolma et al.
581 2013; Kheradpour and Kellis 2014). Position weight matrices of transcription factor
582 binding sequences were transformed into TRANSFAC matrices and then into MEME

583 matrices using in-house scripts and transfac2meme in MEME suite (Bailey et al. 2009).
584 Transcription factor binding sequences of TF derived from vertebrates were used for
585 further analyses. Transcription factor binding sequences were searched from each
586 narrow peak of DNase-seq data using FIMO with *p*-value threshold of 10^{-5} (Grant et al.
587 2011). TF corresponding to transcription factor binding sequences were searched
588 computationally by comparing their names and gene symbols of HGNC (HUGO Gene
589 Nomenclature Committee) -approved gene nomenclature and 31,848 UCSC known
590 canonical transcripts
591 (http://hgdownload.cse.ucsc.edu/goldenpath/hg19/database/knownCanonical.txt.gz), as
592 transcription factor binding sequences were not linked to transcript IDs such as UCSC,
593 RefSeq, and Ensembl transcripts.
594 Target genes of a TF were assigned when its TFBS was found in DNase-seq
595 narrow peaks in promoter or extended regions for enhancer-promoter association of
596 genes (EPA). Promoter and extended regions were defined as follows: promoter regions
597 were those that were within distance of ±5 kb from transcriptional start sites (TSS).
598 Promoter and extended regions were defined as per the following association rule,
599 which is the same as that defined in Figure 3A of a previous study (McLean et al.
600 2010): the single nearest gene association rule, which extends the regulatory domain to
601 the midpoint between the TSS of the gene and that of the nearest gene upstream and
602 downstream without the limitation of extension length. Extended regions for EPA were
603 shortened at the genomic locations of DNA binding sites of a TF that was the closest to
604 a transcriptional start site, and transcriptional target genes were predicted from the

605 shortened enhancer regions using TFBS. Furthermore, promoter and extended regions
606 for EPA were shortened at the genomic locations of forward–reverse (FR) orientation of
607 DNA binding sites of a TF. When forward or reverse orientation of DNA binding sites
608 were continuously located in genome sequences several times, the most external
609 forward–reverse orientation of DNA binding sites were selected. The genomic positions
610 of genes were identified using 'knownGene.txt.gz' file in UCSC bioinformatics sites
611 (Karolchik et al. 2014). The file 'knownCanonical.txt.gz' was also utilized for choosing
612 representative transcripts among various alternate forms for assigning promoter and
613 extended regions for EPA. From the list of transcription factor binding sequences and
614 transcriptional target genes, redundant transcription factor binding sequences were
615 removed by comparing the target genes of a transcription factor binding sequence and
616 its corresponding TF; if identical, one of the transcription factor binding sequences was
617 used. When the number of transcriptional target genes predicted from a transcription
618 factor binding sequence was less than five, the transcription factor binding sequence
619 was omitted.

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

624 For gene expression data, RNA-seq reads mapped onto human hg19 genome
625 sequences were obtained, including ENCODE long RNA-seq reads with poly-A of
626 H1-hESC, iPSC, HUVEC, and MCF-7 (GSM26284, GSM958733, GSM2344099,

627 GSM2344100, GSM958734, and GSM765388), and UCSF-UBC human reference
628 epigenome mapping project RNA-seq reads with poly-A of naive $CD4^+$ T cells
629 (GSM669617). Two replicates were present for H1-hESC, iPSC, HUVEC, and MCF-7,
630 and a single one for $CD4^+$ T cells. RPKMs of the RNA-seq data were calculated using
631 RSeQC (Wang et al. 2012). For monocytes, Blueprint RNA-seq RPKM data
632 (GSE58310, GSE58310_GeneExpression.csv.gz, Monocytes_Day0_RPMI) were used
633 (Saeed et al. 2014). Based on RPKM, UCSC transcripts with expression level among
634 top 30% of all the transcripts were selected in each cell type.

635 The expression level of transcriptional target genes predicted based on EPA
636 shortened at the genomic locations of DNA motif sequence of a TF or a repeat DNA
637 sequence was compared with the expression level of transcriptional target genes
638 predicted from promoter. For each DNA motif sequence shortening EPA,
639 transcriptional target genes were predicted using about 3,000 – 5,000 DNA binding
640 motif sequences of TF, and the distribution of expression level of putative
641 transcriptional target genes of each TF was compared between EPA and only promoter
642 using Mann-Whitney test, two-sided (p-value < 0.05). The number of TF showing a
643 significant difference of expression level of putative transcriptional target genes
644 between EPA and promoter was compared among forward-reverse (FR),
645 reverse-forward (RF), and any orientation (i.e. without considering orientation) of a
646 DNA motif sequence shortening EPA using chi-square test (p-value < 0.05). When a
647 DNA motif sequence of a TF or a repeat DNA sequence shortening EPA showed a
648 significant difference of expression level of putative transcriptional target genes among

649 FR, RF, or any orientation in monocytes of four people in common, the DNA motif
650 sequence was listed.

651 Though forward-reverse orientation of DNA binding motif sequences of CTCF
652 and cohesin are frequently observed at chromatin interaction anchors, the percentage of
653 forward-reverse orientation is not 100, and other orientations of the DNA binding motif
654 sequences are also observed. Though DNA binding motif sequences of CTCF and
655 cohesin are found in various open chromatin regions, DNA binding motif sequences of
656 some transcription factors would be observed less frequently in open chromatin regions.
657 The analyses of experimental data of a number of people would avoid missing relatively
658 weak statistical significance of DNA motif sequences in experimental data of each
659 person by multiple testing correction of thousands of statistical tests. A DNA motif
660 sequence was found with p-value < 0.05 in experimental data of one person and the
661 DNA motif sequence found in the same cell type of four people in common would have
662 p-value $< 0.05^4 = 6.25 \times 10^{-6}$.

663

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

665 Co-location of biased orientation of DNA binding motif sequences of TF was
666 examined. The number of open chromatin regions with the same pair of DNA binding
667 motif sequences was counted, and when the pair of DNA binding motif sequences were
668 enriched with statistical significance (chi-square test, p-value $< 1.0 \times 10^{-10}$), they were
669 listed. For histone modification of an enhancer mark (H3K27ac), bed files of hg38 of
670 Blueprint ChIP-seq data for $CD14^+$ monocytes (EGAD00001001179) and $CD4^+$ T cells

671 (Donor ID: S000RD) were obtained from Blueprint web site
672 (http://dcc.blueprint-epigenome.eu/#/home), and the bed files of hg38 were converted
673 into those of hg19 using Batch Coordinate Conversion (liftOver) web site
674 (https://genome.ucsc.edu/util.html). Networks of co-locations of biased orientation of
675 DNA motif sequences were plotted using Cytoscape v3.71 with yFiles Layout
676 Algorithms for Cytoscape (Shannon et al. 2003).
677
678 **Comparison with chromatin interaction data**
679 For comparison of EPA in monocytes with chromatin interactions,
680 'PCHiC_peak_matrix_cutoff0.txt.gz' file was downloaded from 'Promoter Capture
681 Hi-C in 17 human primary blood cell types' web site (https://osf.io/u8tzp/files/), and
682 chromatin interactions for Monocytes with scores of CHiCAGO tool > 1 and
683 CHiCAGO tool > 5 were extracted from the file (Javierre et al. 2016). In the same way
684 as monocytes, Hi-C chromatin interaction data of $CD4^+$ T cells (Naive $CD4^+$ T cells,
685 nCD4) were obtained.
686 Enhancer-promoter interactions (EPI) were predicted using three types of EPAs
687 in monocytes: (i) EPA shortened at the genomic locations of FR or RF orientation of
688 DNA motif sequence of a TF, (ii) EPA shortened at the genomic locations of any
689 orientation (i.e. without considering orientation) of DNA motif sequence of a TF, and
690 (iii) EPA without being shortened by a DNA motif sequence. EPI predicted using the
691 three types of EPAs in common were removed. First, EPI predicted based on EPA (i)
692 were compared with chromatin interactions (Hi-C). The resolution of chromatin

693 interaction data used in this study was 50kb, so EPI were adjusted to 50kb before their
694 comparison. The number and ratio of EPI overlapped with chromatin interactions were
695 counted. Second, EPI were predicted based on EPA (ii), and EPI predicted based on
696 EPA (i) were removed from the EPI. The number and ratio of EPI overlapped with
697 chromatin interactions were counted. Third, EPI were predicted based on EPA (iii), and
698 EPI predicted based on EPA (i) and (ii) were removed from the EPI. The number and
699 ratio of EPI overlapped with chromatin interactions were counted. The number and ratio
700 of the EPI were compared two times between EPA (i) and (iii), and EPA (i) and (ii)
701 (binomial distribution, p-value < 0.025 for each test, two-sided, 95% confidence
702 interval).

703 For comparison of EPA with chromatin interactions (HiChIP) in $CD4^+$ T cells,
704 'GSM2705049_Naive_HiChIP_H3K27ac_B2T1_allValidPairs.txt',
705 'GSM2705050_Naive_HiChIP_H3K27ac_B2T2_allValidPairs.txt', and
706 'GSM2705051_Naive_HiChIP_H3K27ac_B3T1_allValidPairs.txt' files were
707 downloaded from Gene Expression Omnibus (GEO) database (GSM2705049,
708 GSM2705050 and GSM2705051). The resolutions of chromatin interaction data and
709 EPI were adjusted to 5kb before their comparison. Chromatin interactions with more
710 than 6,000 and 1,000 counts for each interaction were used in this study.

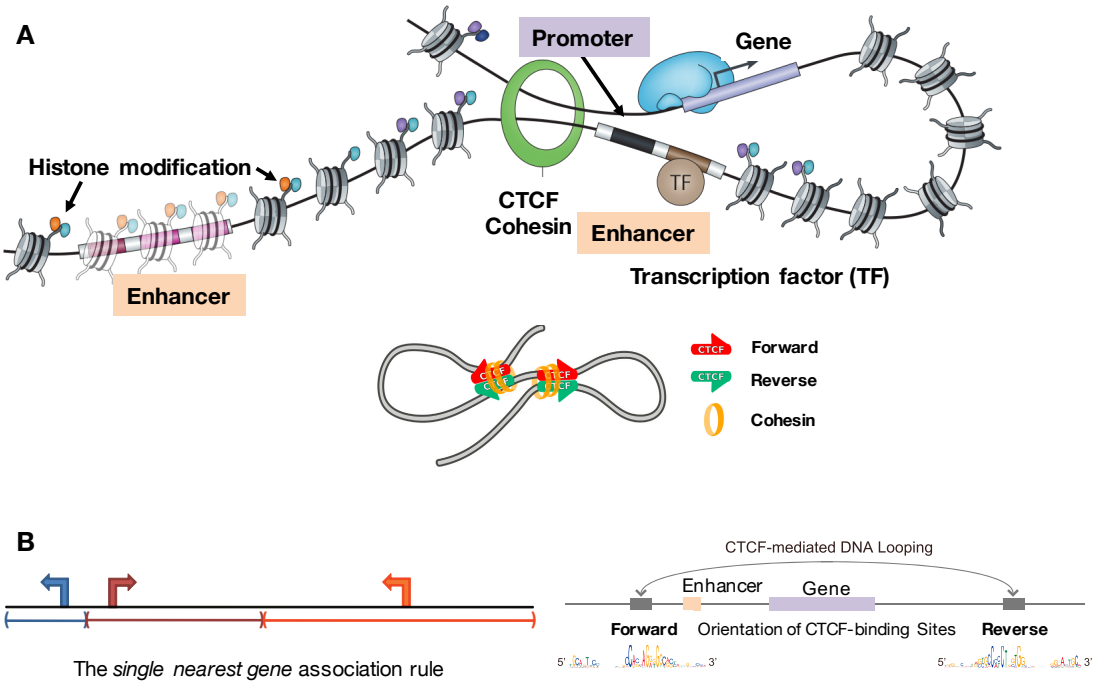
711 The expression level of putative target genes of EPI overlapped with HiChIP
712 chromatin interactions was compared with EPI not overlapped with them. For each FR
713 or RF orientation of DNA motif, EPI were predicted based on EPA and the overlap of
714 EPI with chromatin interactions was examined. When a putative transcriptional target

715 gene of a TF in an enhancer was found in both EPI overlapped with a chromatin
716 interaction and EPI not overlapped with, the target gene was removed. The distribution
717 of expression level of putative target genes was compared using Mann-Whitney test,
718 two-sided (p-value < 0.05).
719

720 **Acknowledgements**

721 The supercomputing resource was provided by Human Genome Center of the Institute
722 of Medical Science at the University of Tokyo. Computations were partially performed
723 on the NIG supercomputer at ROIS National Institute of Genetics. Publication charges
724 for this article were funded by JSPS KAKENHI Grant Number 16K00387. This
725 research was partially supported by the Platform Project for Supporting in Drug
726 Discovery and Life Science Research（Platform for Dynamic Approaches to Living
727 System）from Japan Agency for Medical Research and Development (AMED). This
728 research was partially supported by Development of Fundamental Technologies for
729 Diagnosis and Therapy Based upon Epigenome Analysis from Japan Agency for
730 Medical Research and Development (AMED). This work was partially supported by
731 JST CREST Grant Number JPMJCR15G1, Japan.
732

733 **Figures**



734

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

745

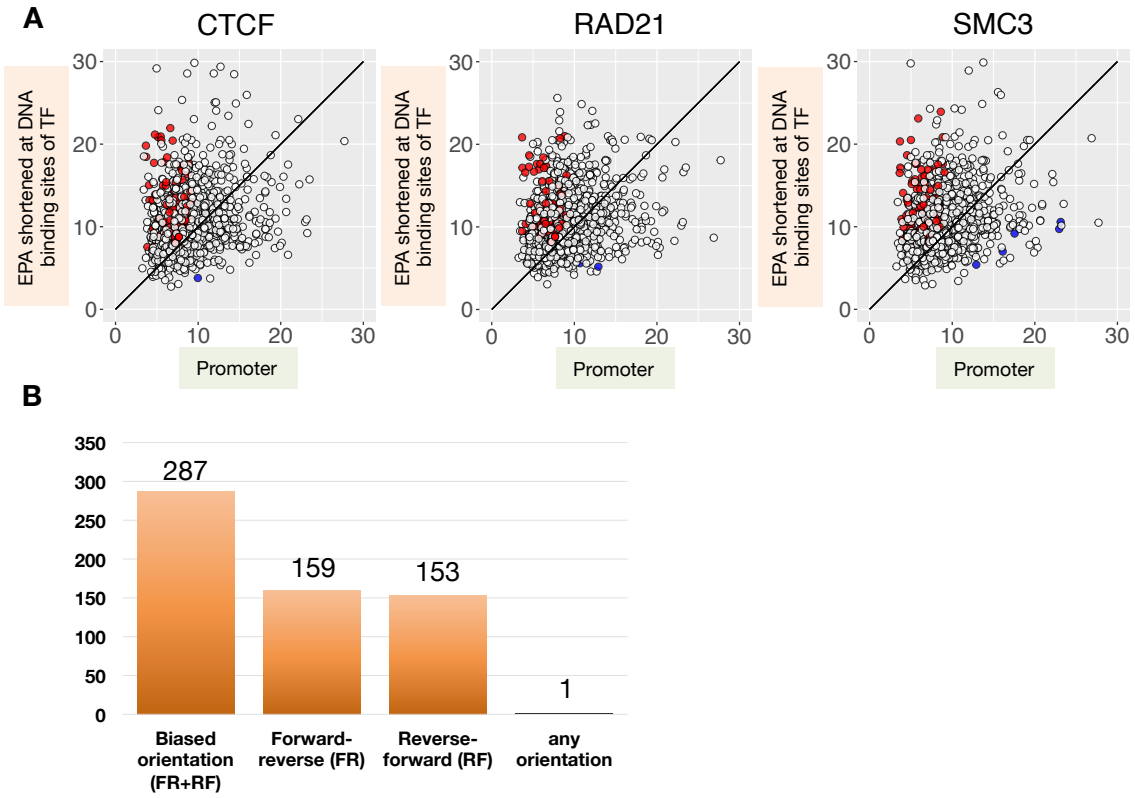
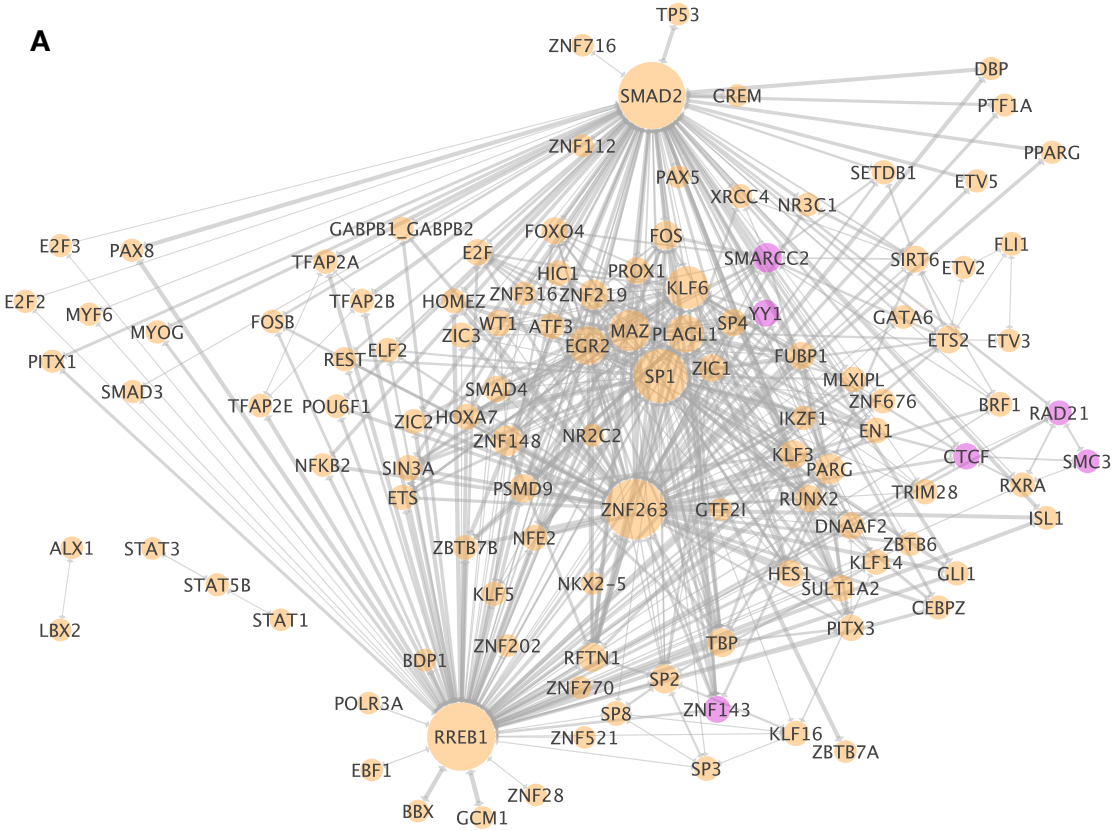


Figure 2. Biased orientation of DNA motif sequences of transcription factors affecting the transcription of genes. (A) Comparison of expression level of putative transcriptional target genes. The median expression levels of target genes of the same transcription factor binding sequences were compared between promoters and enhancer-promoter association (EPA) shortened at the genomic locations of forward-reverse orientation of DNA motif sequence of CTCF and cohesin (RAD21 and SMC3) respectively in monocytes. Red and blue dots show statistically significant difference (Mann-Whitney test) of the distribution of expression levels of target genes between promoters and the EPA. Red dots show the median expression level of target genes was higher based on the EPA than promoters, and blue dots show the median expression level of target genes was lower based on the EPA than promoters. The

758 expression level of target genes predicted based on the EPA tended to be higher than
759 promoters, implying that transcription factors acted as activators of target genes. (B)
760 Biased orientation of DNA motif sequences of transcription factors in monocytes. Total
761 287 of biased orientation of DNA binding motif sequences of transcription factors were
762 found to affect the expression level of putative transcriptional target genes in monocytes
763 of four people in common, whereas only one any orientation (i.e. without considering
764 orientation) of DNA binding motif sequence was found to affect the expression level.
765



766
767

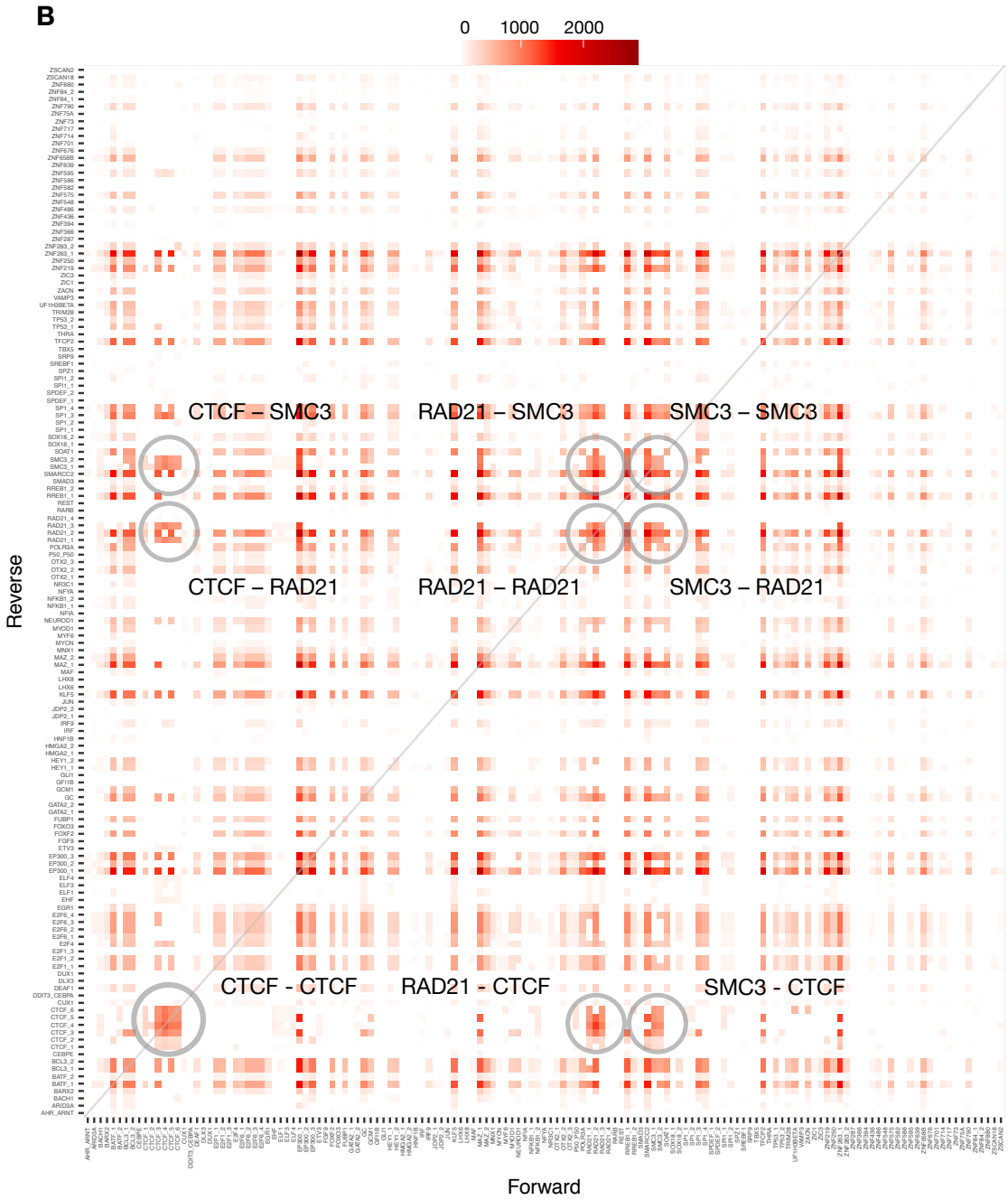


Figure 3. Association of biased orientation of DNA binding motif sequences of transcription factors. (A) Co-location of biased orientation of DNA binding motif sequences of transcription factors. Co-locations of the DNA motif sequences in open chromatin regions were examined in monocytes. When the same DNA motif sequence

773 of a TF was co-localized with DNA motif sequences of more than one TF in the same or
774 different open chromatin regions, these co-locations of DNA motif sequences were
775 connected to each other. Purple circles of TFs are known to be involved in chromatin
776 interactions. The line thickness is correlated with the number of open chromatin regions
777 with co-location of the same pair of DNA motif sequences. The size of a circle is
778 correlated with the number of unique TF co-localized with the TF of the circle, which is
779 the number of edges of the node. (B) Pairs of biased orientation of DNA binding motif
780 sequences of transcription factors enriched in upstream and downstream of genes. The
781 number of genes with a pair of forward-reverse orientation of DNA motif sequences
782 was counted using all pairs of the DNA motif sequences found in open chromatin
783 regions overlapped with H3K27ac histone modification marks in T cells, and statistical
784 tests (chi-square test) were conducted. The number of genes where a pair of
785 forward-reverse orientation of DNA motif sequences were enriched with p-value < 0.05
786 was counted and shown as heat map. Gray circles show pairs of DNA motif sequences
787 of CTCF and cohesin (RAD21 and SMC3). CTCF and cohesin are known to be
788 involved in chromatin interactions and have biased orientation of DNA binding motif
789 sequences to form chromatin loops. Other pairs of forward-reverse orientation of DNA
790 motif sequences were also enriched in upstream and downstream of genes.
791

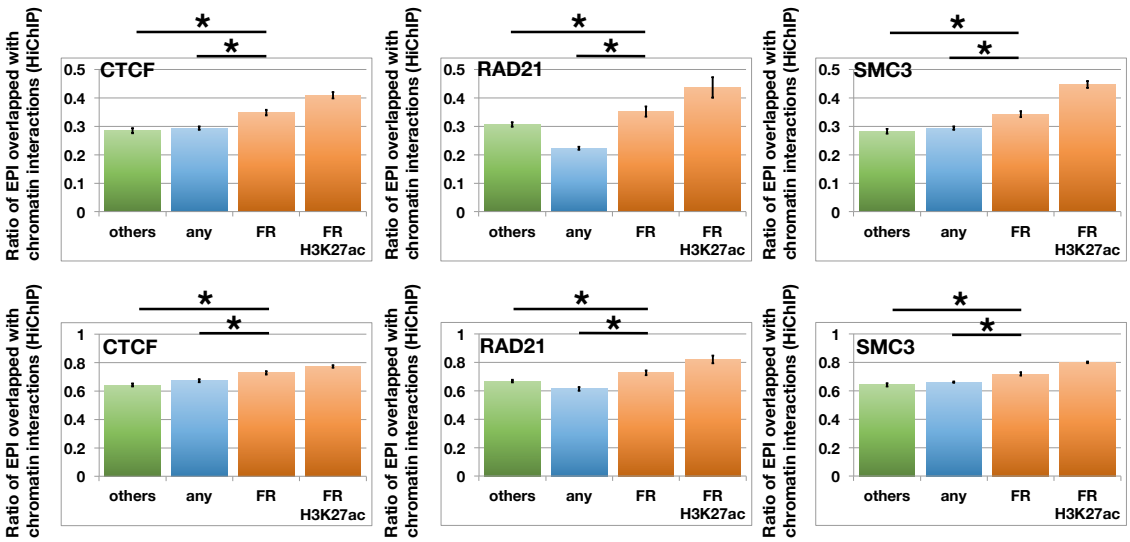
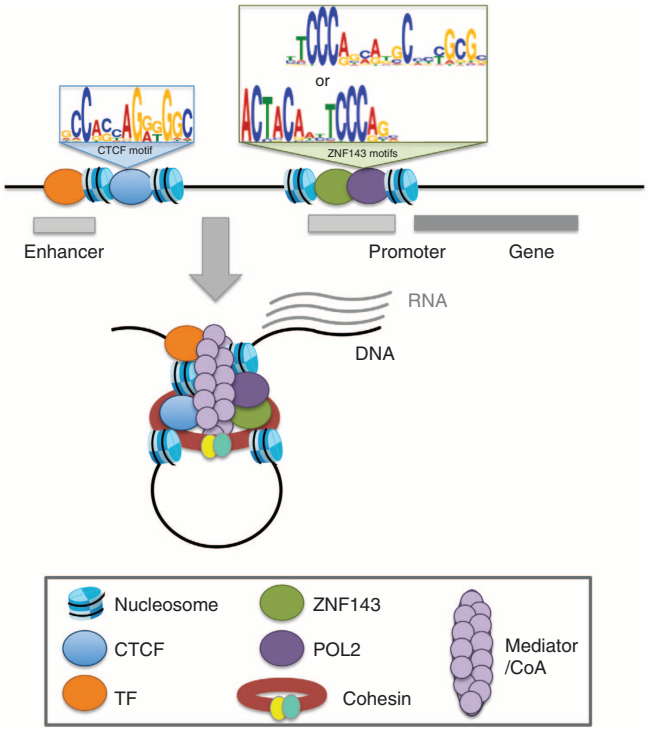


Figure 4. Comparison of enhancer-promoter interactions (EPI) with chromatin interactions in T cells. EPI were predicted based on enhancer-promoter association (EPA) shortened at the genomic locations of biased orientation of DNA binding motif sequence of a transcription factor. Total 88 biased orientation (53 FR and 42 RF) of DNA motif sequences including CTCF and cohesin showed a significantly higher ratio of EPI overlapped with three replications of HiChIP chromatin interactions respectively than the other types of EPAs in T cells. The upper part of the figure is a comparison of EPI with HiChIP chromatin interactions with more than 6,000 counts for each interaction, and the lower part of the figure is a comparison of EPI with HiChIP chromatin interactions with more than 1,000 counts for each interaction. FR: forward-reverse orientation of DNA motif. any: any orientation (i.e. without considering orientation) of DNA motif. others: enhancer-promoter association not shortened at the genomic locations of DNA motif. FR H3K27ac: EPI were predicted using DNA motif in open chromatin regions overlapped with H3K27ac histone modification marks. *

807 p-value < 10^{-9}.



808

809 **Figure 5.** Schematic representation of chromatin interactions involving gene promoters.

810 ZNF143 contributes the formation of chromatin interactions by directly binding the

811 promoter of genes establishing looping with distal element bound by CTCF (Bailey et al.

812 2015).

813

814 **Tables**

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

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

817 $-\log_{10}$ (*p*-value).

818 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

819

820

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

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

827 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

828
829
830
831 **Table 3.** Biased orientation of repeat DNA sequences in monocytes. Score: $-\log_{10}$
832 (*p*-value).

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

Repeat DNA seq.	Score
MLT1F1	6.75

833

Reverse-forward (RF) orientation

Repeat DNA seq.	Score
LTR16C	69.39
L1ME4b	23.90
AluSg	23.61

834

835 **Table 4.** Top 30 of co-locations of biased orientation of DNA binding motif sequences
836 of transcription factors in monocytes. Co-locations of DNA motif sequence of CTCF
837 with another biased orientation of DNA motif sequence were shown in a separate table.
838 Motif 1,2: DNA binding motif sequences of transcription factors. # both: the number of
839 open chromatin regions including both Motif 1 and Motif 2. # motif 1: the number of
840 open chromatin regions including Motif 1. # motif 2: the number of open chromatin
841 regions including Motif 2. # others: the number of open chromatin regions not including
842 Motif 1 and Motif 2. Open chromatin regions overlapped with histone modification
843 marks (H3K27ac) were used (Total 26,095 regions).

844

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

845

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

846 **Table 5.** Comparison of enhancer-promoter interactions (EPI) with chromatin
847 interactions in T cells. The upper tables show the number of biased orientation of DNA
848 motifs, where a significantly higher ratio of EPI, which were predicted based on
849 enhancer-promoter association (EPA) (iii), overlapped with HiChIP chromatin
850 interaction data than the other types of EPA (i) and (ii). The lower tables show the 53
851 FR and 42 RF orientations of DNA motifs found in common among B2T1, B2T2 and
852 B3T1 replications in the upper table. TF: DNA binding motif sequence of a
853 transcription factor. Score: $-\log_{10}$ (p-value). Inf: p-value = 0. Score 1: Comparison of
854 EPA shortened at the genomic locations of FR or RF orientation of DNA motif
855 sequence [EPA (iii)] with EPA not shortened [EPA (i)]. Score 2: Comparison of EPA
856 shortened at the genomic locations of FR or RF orientation of DNA motif sequence
857 [EPA (iii)] with EPA shortened at the genomic locations of any orientation of DNA
858 motif sequence [EPA (ii)].

1. Comparison of EPI with HiChIP data among three EPA (i), (ii) and (iii)

Replication of HiChIP data	Total no. (FR + RF) of DNA motifs	No. of FR DNA motifs	No. of RF DNA motifs
B2T1	114	67	55
B2T2	109	64	55
B3T1	129	77	62
DNA motifs found in common among replications			
B2T1 and B2T2	96	56	48
B2T1 and B3T1	100	59	48
B2T2 and B3T1	98	61	46
B2T1, B2T2 and B3T1	88	53	42

2. Comparison of EPI with HiChIP data among two EPA (i) and (iii)

Replication of HiChIP data	Total no. (FR + RF) of DNA motifs	No. of FR DNA motifs	No. of RF DNA motifs
B2T1	251	150	132
B2T2	252	151	132
B3T1	256	154	134
DNA motifs found in common among replications			
B2T1 and B2T2	249	149	131
B2T1 and B3T1	249	149	131
B2T2 and B3T1	250	150	131
B2T1, B2T2 and B3T1	247	148	130

859
860

Forward-reverse (FR) orientation

TF	Score 1	Score 2
APEX1	11.61	4.20
AR	313.05	3.18
CDC5L	74.26	10.85
CTCF	127.57	22.15
DDIT3_CEBPA	323.31	6.55
EGR1	106.24	4.65
EGR3	128.66	3.46
ELF5	138.35	3.04
EP300	48.37	9.07
FOXF2	58.93	5.61
FOXN4	149.19	8.18
FOXO3	323.31	5.43
GATA5	56.91	2.46
GFI1B	184.50	8.50
HBP1	323.31	4.00
HOXA2	Inf	7.95
INSM2	323.31	2.28
IRF	316.95	3.24
IRF5	54.29	2.01
KLF1	78.84	7.86
LHX8	323.31	3.55
MAF	69.75	4.04
MINI20	107.00	2.39
MYC	44.64	1.78
NFKB1	184.16	9.48
NR2C2	120.01	6.47
OTX2	45.97	1.86

TF	Score 1	Score 2
P50_P50	158.48	7.98
PITX1	54.06	3.71
RAD21	81.14	9.49
RARB	222.14	5.78
REL	134.25	9.66
RREB	114.09	3.51
SMC3	101.43	22.77
SOX18	94.79	4.19
SRP9	138.72	2.19
TCF7L2	118.73	4.88
TFCP2	276.08	4.14
TP53	232.21	4.45
TP63	153.80	3.33
YBX1	184.46	3.25
ZBTB7A	48.34	5.97
ZNF219	45.97	8.82
ZNF253	Inf	20.00
ZNF585A	166.21	3.15
ZNF616	99.63	6.31
ZNF681	323.31	35.37
ZNF721	124.07	3.71
ZNF75A	133.89	6.49
ZNF770	90.87	3.91
ZNF786	323.31	12.81
ZNF790	85.57	9.62
ZNF846	323.31	10.80

Reverse-forward (RF) orientation

TF	Score 1	Score 2
BCL11B	91.43	5.00
CEBPA	Inf	1.83
CREB3	141.55	15.24
CTCF	105.84	18.34
DBP	195.89	2.48
EP300	163.02	5.08
ESR1	73.03	4.44
ETV1	219.09	3.31
FOS	73.22	5.67
FOXA	323.31	6.39
FOXA1	323.31	7.12
FOXN4	88.04	6.94
FOXO3	323.31	4.61
HOXB6	254.32	3.05
IRF5	Inf	8.44
KLF13	173.56	6.79
NFE2L2	82.85	1.82
NFKB1	110.00	5.78
NFYA	196.79	2.04
NKX2-1	323.31	7.60
NKX2-4	323.31	6.17

TF	Score 1	Score 2
NKX2-5	Inf	5.84
RUNX2	182.38	5.57
SIX6	82.33	5.12
SMARC	323.31	3.22
SRF	323.31	5.58
TAL1	134.12	4.06
ZBTB18	230.07	8.48
ZBTB4	Inf	4.88
ZFP2	323.31	5.36
ZNF257	146.66	7.87
ZNF28	323.31	4.50
ZNF282	133.41	4.00
ZNF300	323.31	2.29
ZNF341	102.58	3.74
ZNF44	194.89	5.88
ZNF484	323.31	11.42
ZNF526	76.12	2.62
ZNF625	Inf	5.86
ZNF658	99.09	7.89
ZNF721	93.57	3.96
ZNF773	175.66	6.26

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

896 Enhancer/Promoter Function. *Cell* **162**: 900-910.
897 Javierre BM, Burren OS, Wilder SP, Kreuzhuber R, Hill SM, Sewitz S, Cairns J,
898 Wingett SW, Varnai C, Thiecke MJ et al. 2016. Lineage-Specific Genome
899 Architecture Links Enhancers and Non-coding Disease Variants to Target Gene
900 Promoters. *Cell* **167**: 1369-1384 e1319.
901 Jeong M, Huang X, Zhang X, Su J, Shamim M, Bochkov I, Reyes J, Jung H, Heikamp
902 E, Presser Aiden A et al. 2017. A Cell Type-Specific Class of Chromatin Loops
903 Anchored at Large DNA Methylation Nadirs. *bioRxiv*.
904 Ji X, Dadon DB, Abraham BJ, Lee TI, Jaenisch R, Bradner JE, Young RA. 2015.
905 Chromatin proteomic profiling reveals novel proteins associated with
906 histone-marked genomic regions. *Proc Natl Acad Sci U S A* **112**: 3841-3846.
907 Jin C, Kato K, Chimura T, Yamasaki T, Nakade K, Murata T, Li H, Pan J, Zhao M, Sun
908 K et al. 2006. Regulation of histone acetylation and nucleosome assembly by
909 transcription factor JDP2. *Nat Struct Mol Biol* **13**: 331-338.
910 Jjingo D, Conley AB, Wang J, Marino-Ramirez L, Lunyak VV, Jordan IK. 2014.
911 Mammalian-wide interspersed repeat (MIR)-derived enhancers and the
912 regulation of human gene expression. *Mob DNA* **5**: 14.
913 John S, Sabo PJ, Thurman RE, Sung MH, Biddie SC, Johnson TA, Hager GL,
914 Stamatoyannopoulos JA. 2011. Chromatin accessibility pre-determines
915 glucocorticoid receptor binding patterns. *Nature genetics* **43**: 264-268.
916 Jolma A, Yan J, Whitington T, Toivonen J, Nitta KR, Rastas P, Morgunova E, Enge M,
917 Taipale M, Wei G et al. 2013. DNA-binding specificities of human transcription
918 factors. *Cell* **152**: 327-339.
919 Karolchik D, Barber GP, Casper J, Clawson H, Cline MS, Diekhans M, Dreszer TR,
920 Fujita PA, Guruvadoo L, Haeussler M et al. 2014. The UCSC Genome Browser
921 database: 2014 update. *Nucleic acids research* **42**: D764-770.
922 Katainen R, Dave K, Pitkanen E, Palin K, Kivioja T, Valimaki N, Gylfe AE,
923 Ristolainen H, Hanninen UA, Cajuso T et al. 2015. CTCF/cohesin-binding sites
924 are frequently mutated in cancer. *Nature genetics* **47**: 818-821.
925 Kheradpour P, Kellis M. 2014. Systematic discovery and characterization of regulatory
926 motifs in ENCODE TF binding experiments. *Nucleic acids research* **42**:
927 2976-2987.
928 Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler

929 transform. *Bioinformatics (Oxford, England)* **25**: 1754-1760.
930 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G,
931 Durbin R, Genome Project Data Processing S. 2009. The Sequence
932 Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)* **25**:
933 2078-2079.
934 McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM,
935 Bejerano G. 2010. GREAT improves functional interpretation of cis-regulatory
936 regions. *Nature biotechnology* **28**: 495-501.
937 Mumbach MR, Satpathy AT, Boyle EA, Dai C, Gowen BG, Cho SW, Nguyen ML,
938 Rubin AJ, Granja JM, Kazane KR et al. 2017. Enhancer connectome in primary
939 human cells identifies target genes of disease-associated DNA elements. *Nature*
940 *genetics* **49**: 1602-1612.
941 Newburger DE, Bulyk ML. 2009. UniPROBE: an online database of protein binding
942 microarray data on protein-DNA interactions. *Nucleic acids research* **37**:
943 D77-82.
944 Osato N. 2018. Characteristics of functional enrichment and gene expression level of
945 human putative transcriptional target genes. *BMC Genomics* **19**: 957.
946 Phanstiel DH, Van Bortle K, Spacek D, Hess GT, Shamim MS, Machol I, Love MI,
947 Aiden EL, Bassik MC, Snyder MP. 2017. Static and Dynamic DNA Loops form
948 AP-1-Bound Activation Hubs during Macrophage Development. *Molecular cell*
949 **67**: 1037-1048.e1036.
950 Portales-Casamar E, Thongjuea S, Kwon AT, Arenillas D, Zhao X, Valen E, Yusuf D,
951 Lenhard B, Wasserman WW, Sandelin A. 2010. JASPAR 2010: the greatly
952 expanded open-access database of transcription factor binding profiles. *Nucleic*
953 *acids research* **38**: D105-110.
954 Ramani V, Cusanovich DA, Hause RJ, Ma W, Qiu R, Deng X, Blau CA, Disteche CM,
955 Noble WS, Shendure J et al. 2016. Mapping 3D genome architecture through in
956 situ DNase Hi-C. *Nat Protoc* **11**: 2104-2121.
957 Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn
958 AL, Machol I, Omer AD, Lander ES et al. 2014. A 3D map of the human
959 genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**:
960 1665-1680.
961 Rebollo R, Romanish MT, Mager DL. 2012. Transposable elements: an abundant and

962 natural source of regulatory sequences for host genes. *Annu Rev Genet* **46**:
963 21-42.
964 Saeed S, Quintin J, Kerstens HH, Rao NA, Aghajanirefah A, Matarese F, Cheng SC,
965 Ratter J, Berentsen K, van der Ent MA et al. 2014. Epigenetic programming of
966 monocyte-to-macrophage differentiation and trained innate immunity. *Science*
967 *(New York, NY)* **345**: 1251086.
968 Schreiber J, Libbrecht M, Bilmes J, Noble W. 2017. Nucleotide sequence and DNaseI
969 sensitivity are predictive of 3D chromatin architecture. *bioRxiv*.
970 Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N,
971 Schwikowski B, Ideker T. 2003. Cytoscape: a software environment for
972 integrated models of biomolecular interaction networks. *Genome research* **13**:
973 2498-2504.
974 Tabuchi TM, Deplancke B, Osato N, Zhu LJ, Barrasa MI, Harrison MM, Horvitz HR,
975 Walhout AJ, Hagstrom KA. 2011. Chromosome-biased binding and gene
976 regulation by the Caenorhabditis elegans DRM complex. *PLoS Genet* **7**:
977 e1002074.
978 Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF,
979 Schroth GP, Burge CB. 2008. Alternative isoform regulation in human tissue
980 transcriptomes. *Nature* **456**: 470-476.
981 Wang J, Vicente-Garcia C, Seruggia D, Molto E, Fernandez-Minan A, Neto A, Lee E,
982 Gomez-Skarmeta JL, Montoliu L, Lunyak VV et al. 2015. MIR retrotransposon
983 sequences provide insulators to the human genome. *Proc Natl Acad Sci U S A*
984 **112**: E4428-4437.
985 Wang L, Wang S, Li W. 2012. RSeQC: quality control of RNA-seq experiments.
986 *Bioinformatics (Oxford, England)* **28**: 2184-2185.
987 Weintraub AS, Li CH, Zamudio AV, Sigova AA, Hannett NM, Day DS, Abraham BJ,
988 Cohen MA, Nabet B, Buckley DL et al. 2017. YY1 Is a Structural Regulator of
989 Enhancer-Promoter Loops. *Cell* **171**: 1573-1588 e1528.
990 Wingender E, Dietze P, Karas H, Knuppel R. 1996. TRANSFAC: a database on
991 transcription factors and their DNA binding sites. *Nucleic acids research* **24**:
992 238-241.
993 Xie Z, Hu S, Blackshaw S, Zhu H, Qian J. 2010. hPDI: a database of experimental
994 human protein-DNA interactions. *Bioinformatics (Oxford, England)* **26**:

995 287-289.
996 Zhang H, Li F, Jia Y, Xu B, Zhang Y, Li X, Zhang Z. 2017. Characteristic arrangement
997 of nucleosomes is predictive of chromatin interactions at kilobase resolution.
998 *Nucleic acids research* **45**: 12739-12751.
999 Zhang Y, An L, Xu J, Zhang B, Zheng WJ, Hu M, Tang J, Yue F. 2018. Enhancing
1000 Hi-C data resolution with deep convolutional neural network HiCPlus. *Nat*
1001 *Commun* **9**: 750.
1002 Zhao Y, Stormo GD. 2011. Quantitative analysis demonstrates most transcription
1003 factors require only simple models of specificity. *Nature biotechnology* **29**:
1004 480-483.

1005