Discovery of biased orientations 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 229 biased orientations of DNA motifs (130 FR and 128 RF orientations, 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 orientations of DNA motif sequences tended to be co-localized in the same open chromatin regions. For 51 - 204 of FR and RF orientations of DNA motif sequences in $CD4^+$ T cells, 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 significantly more than other types of EPAs. Moreover, 47 of 69 biased orientations of DNA motifs conserved between monocyte and $CD4^+$ T cells showed significantly reduced correlation in expression level of nearby genes separated by the motif sites, suggesting that the DNA motifs affect gene expression.

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

77 (TF) are found in open chromatin regions near transcriptional start sites (TSS) as well as
78 chromatin interaction anchors.

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

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

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

104 DNA binding proteins involved in chromatin interactions are supposed to affect
105 the transcription of genes in the loops formed by chromatin interactions. In my previous
106 analysis, the expression level of human putative transcriptional target genes was
107 affected, according to the criteria of enhancer-promoter association (EPA) (Fig. 1B;
108 (Osato 2018)). EPI were predicted based on EPA shortened at the genomic locations of
109 FR orientation of CTCF binding sites, and transcriptional target genes of each TF bound
110 in enhancers and promoters were predicted based on the EPI. The EPA affected the
111 expression levels of putative transcriptional target genes the most among three types of
112 EPAs, compared with the expression levels of transcriptional target genes predicted
113 from only promoters (Fig. 2). The expression levels tended to be increased in
114 monocytes and $CD4^+$ T cells, implying that enhancers activated the transcription of
115 genes, and decreased in ES and iPS cells, implying that enhancers repressed the
116 transcription of genes. These analyses suggested that enhancers affected the
117 transcription of genes significantly, when EPI were predicted properly. Other DNA
118 binding proteins involved in chromatin interactions, as well as CTCF, may locate at
119 chromatin interaction anchors with a pair of biased orientation of DNA binding motif
120 sequences, affecting the expression level of putative transcriptional target genes in the

121 loops formed by the chromatin interactions. As experimental issues of the analyses of
122 chromatin interactions, chromatin interaction data are changed, according to
123 experimental techniques, depth of DNA sequencing, and even replication sets of the
124 same cell type. Chromatin interaction data may not be saturated enough to cover all
125 chromatin interactions. Supposing these characters of DNA binding proteins associated
126 with chromatin interactions and avoiding experimental issues of the analyses of
127 chromatin interactions, here I searched for DNA motif sequences of TF and repeat DNA
128 sequences, affecting EPI and the expression level of putative transcriptional target genes
129 in $CD14^+$ monocytes and $CD4^+$ T cells of four people without using chromatin
130 interaction data. Then, putative EPI were compared with chromatin interaction data.
131

132 ## Results

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

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

143 transcriptional target genes of each TF bound in enhancers and promoters were
144 predicted based on the EPI. The expression levels of the putative transcriptional target
145 genes were compared with those predicted from promoters using Mann Whiteney U test,
146 two-sided (p-value < 0.05) (Fig. 2A). The number of TF showing a significant
147 difference of expression level of their putative transcriptional target genes was counted.
148 To examine whether the orientation of a DNA motif sequence, which is supposed to act
149 as insulator sites and shorten EPA, affected the number of TF showing a significant
150 difference of expression level of their putative transcriptional target genes, the number
151 of the TF was compared among forward-reverse (FR), reverse-forward (RF), and any
152 orientation (i.e. without considering orientation) of a DNA motif sequence shortening
153 EPA, using Fisher's exact test (p-value < 0.05). To avoid missing DNA motif sequences
154 showing a relatively weak statistical significance by multiple testing correction, the
155 above analyses were conducted using monocytes of four people independently, and
156 DNA motif sequences found in monocytes of four people in common were selected.
157 Total 229 of biased (130 FR and 128 RF) orientation of DNA binding motif sequences
158 of TF were found in monocytes of four people in common, whereas only 11 any
159 orientation of DNA binding motif sequence was found in monocytes of four people in
160 common (Fig. 2B; Table 1; Supplemental Table S1). FR orientation of DNA motif
161 sequences included CTCF, cohesin (RAD21), and YY1, which are associated with
162 chromatin interactions and EPI. NFI (Nuclear factor-1) has been reported to have
163 enhancer-blocking activity (Gaussin et al. 2012), and NFI showed a biased orientation
164 in monocytes and $CD4^+$ T cells.

165 The same analysis was conducted using DNase-seq data of $CD4^+$ T cells of four
166 people. Total 211 of biased (120 FR and 122 RF) orientation of DNA binding motif
167 sequences of TF were found in T cells of four people in common, whereas only 15 any
168 orientation (i.e. without considering orientation) of DNA binding motif sequences were
169 found in T cells of four people in common (Supplemental Fig. S1 and Supplemental
170 Table S2). Biased orientation of DNA motif sequences in T cells included CTCF,
171 cohesin (RAD21 and SMC3), YY1 and SMARC. SMARCA4 (BRG1) is associated
172 with topologically associated domain (TAD), which is higher-order chromatin
173 organization. The DNA binding motif sequence of SMARCA4 was not registered in the
174 databases used in this study. FR orientation of DNA motif sequences included
175 SMARCC2, a member of the same SWI/SNF family of proteins as SMARCA4. Biased
176 orientation of DNA motif sequences included JUNDM2 (JDP2), which is involved in
177 histone-chaperone activity, promoting nucleosome, and inhibition of histone acetylation
178 (Jin et al. 2006). JDP2 forms a homodimer or heterodimer with various TF
179 (https://en.wikipedia.org/wiki/Jun_dimerization_protein). Between 229 and 211, 69 of
180 biased orientation of DNA binding motif sequences of TF were found in both
181 monocytes and T cells in common (Supplemental Table S5). For each orientation, 42
182 FR and 36 RF orientations of DNA binding motif sequences of TF were found in both
183 monocytes and T cells in common. Without considering the difference of orientation of
184 DNA binding motif sequences, 88 of biased orientation of DNA binding motif
185 sequences of TF were found in both monocytes and T cells. As a reason for the increase
186 of the number (88) from 69, a TF or an isoform of the same TF may bind to a different

187 DNA binding motif sequence according to cell types and/or in the same cell type. About
188 50% or more of alternative splicing isoforms are differently expressed among tissues,
189 indicating that most alternative splicing is subject to tissue-specific regulation (Wang et
190 al. 2008) (Chen and Manley 2009) (Das et al. 2007). The same TF has several DNA
191 binding motif sequences and in some cases one of the motif sequences is almost the
192 same as the reverse complement sequence of another motif sequence of the same TF.
193 For example, RAD21 had both FR and RF orientations of DNA motif sequences, but
194 the number of the FR orientation of DNA motif sequence was relatively small in the
195 genome, and the RF orientation of DNA motif sequence was frequently observed and
196 co-localized with CTCF. I previously found that a complex of TF would bind to a
197 slightly different DNA binding motif sequence from the combination of DNA binding
198 motif sequences of TF composing the complex in *C. elegans* (Tabuchi et al. 2011).
199 From another viewpoint of this study, the expression level of putative transcription
200 target genes of some TF would be different, depending on the genomic locations
201 (enhancers or promoters) of DNA binding motif sequences of the TF in monocytes and
202 T cells of four people.

203 Moreover, using open chromatin regions overlapped with H3K27ac histone
204 modification marks known as enhancer and promoter marks, the same analyses were
205 performed in monocytes and T cells. H3K27ac histone modification marks were used in
206 the analysis of EPI, but were not used in the analysis of TF as insulator like CTCF and
207 cohesin in this study, since new biased orientation of DNA motif sequences were found
208 in this criterion. When H3K27ac histone modification marks were used in the analysis

209 of TF as insulator like CTCF and cohesin, the number of biased orientation of DNA
210 motif sequences was decreased. Total 188 of biased (105 FR and 107 RF) orientations
211 of DNA binding motif sequences of TF were found in monocytes of four people in
212 common, whereas only nine any orientation of DNA binding motif sequence was found
213 (Supplemental Table S3). Though the number of biased orientations of DNA motif
214 sequences was reduced, CTCF, RAD21, SMC3, and YY1 were found. For T cells using
215 H3K27ac histone modification marks, total 178 of biased (108 FR and 96 RF)
216 orientation of DNA binding motif sequences of TF were found in T cells of four people
217 in common, whereas only 15 any orientation of DNA binding motif sequences were
218 found (Supplemental Table S4). Though the number of biased orientations of DNA
219 motif sequences was reduced, CTCF, RAD21, SMC3, and YY1 were found. As
220 summary of the results with and without H3K27ac histone modification marks, total
221 289 of biased (165 FR and 166 RF) orientation of DNA motif sequences were found in
222 monocytes of four people in common. Total 262 of biased (154 FR and 151 RF)
223 orientations of DNA motif sequences were found in T cells of four people in common.
224 Total number of these results in monocytes and T cells was 433 biased (262 FR and 264
225 RF) orientations of DNA motif sequences. Biased orientations of DNA motif sequences
226 found in both monocytes and T cells were listed in Supplemental Table S5.

227 Instead of DNA binding motif sequences of TF, repeat DNA sequences were also
228 examined. The expression levels of transcriptional target genes of each TF predicted
229 based on EPA that were shortened at the genomic locations of a repeat DNA sequence
230 were compared with those predicted from promoters. Three RF orientation of repeat

231 DNA sequences showed a significant difference of expression level of putative
232 transcriptional target genes in monocytes of four people in common (Table 2). The
233 same as $CD14^+$ monocytes, biased orientations of repeat DNA sequences were
234 examined in $CD4^+$ T cells. Two FR and one RF orientations of repeat DNA sequences
235 showed a significant difference of expression level of putative transcriptional target
236 genes in T cells of four people in common (Supplemental Table S6). MIRc was found
237 in the analysis using open chromatin regions overlapped with H3K27ac histone
238 modification marks, which are enhancer and promoter marks. MIR and other transposon
239 sequences are known to act as insulators and enhancers (Bejerano et al. 2006; Rebollo et
240 al. 2012; de Souza et al. 2013; Jjingo et al. 2014; Wang et al. 2015).
241
242 **Co-location of biased orientations of DNA motif sequences**
243 To examine the association of 229 biased (FR and RF) orientations of DNA
244 binding motif sequences of TF, co-location of the DNA binding motif sequences in
245 open chromatin regions was analyzed in monocytes. The number of open chromatin
246 regions with the same pairs of DNA binding motif sequences was counted, and when
247 the pairs of DNA binding motif sequences were enriched with statistical significance
248 (Fisher's exact test, p-value $< 1.0 \times 10^{-9}$), they were listed (Table 3; Supplemental
249 Table S7). As already known, CTCF was found with cohesin RAD21 (Table 3). Top 30
250 pairs of FR and RF orientations of DNA motifs co-occupied in the same open chromatin
251 regions were shown (Table 3). Total number of pairs of DNA motifs was 261,
252 consisting of 148 unique DNA motifs, when the pair of DNA motifs was observed in

253 more than 60% of the number of open chromatin regions with the DNA motifs. Biased
254 orientations of DNA binding motif sequences of TF tended to be co-localized in the
255 same open chromatin regions.

256 To examine the association of 211 biased orientations of DNA binding motif
257 sequences of TF in $CD4^+$ T cells, co-location of the DNA binding motif sequences in
258 open chromatin regions was analyzed. Top 30 pairs of FR and RF orientations of DNA
259 motifs co-occupied in the same open chromatin regions were shown (Supplemental
260 Table S8). Total number of pairs of DNA motifs was 400, consisting of 141 unique
261 DNA motifs, when the pair of DNA motifs was observed in more than 60% of the
262 number of open chromatin regions with the DNA motifs (Fisher's exact test, *p*-value <
263 1.0×10^{-9}). Among them, 21 pairs of DNA motif sequences including a pair of CTCF,
264 and RAD21 in T cells were found in monocytes in common (Supplemental Table S8).

265 In both monocytes and $CD4^+$ T cells, some pairs of DNA motifs including CTCF
266 and RAD21 were found to be perfectly co-localized in the same open chromatin regions.
267 Almost all genomic positions of a DNA motif were located with genomic positions of
268 another DNA motif in the same open chromatin regions. This would be due to the strict
269 criteria of the search for DNA binding motif sequences of TF to reduce false positive
270 prediction of DNA binding sites of TF. Moreover, the number of pairs of DNA motif
271 sequences found in both monocytes and $CD4^+$ T cells in common was increased,
272 compared with general DNA motif searches using position weight matrix and allowing
273 mismatches of DNA sequences during motif searches. The score (*p*-value) of CTCF was
274 relatively high (50.54), but the score of RAD21 was lower (23.98) (Table 1). Almost all

275 DNA binding sequences of CTCF were co-localized with those of RAD21. However,
276 some DNA binding sequences of RAD21 were located in other genomic regions, which
277 may not be associated with chromatin interactions and biased orientations of the DNA
278 motif sequences. This might decrease the score of RAD21, compared with CTCF.
279
280 **Comparison with chromatin interaction data**
281 To examine whether the biased orientations of DNA motif sequences is
282 associated with chromatin interactions, enhancer-promoter interactions (EPI) predicted
283 based on enhancer-promoter associations (EPA) were compared with chromatin
284 interaction data. Due to the resolution of Hi-C experimental data available for this study
285 (50kb), HiChIP chromatin interaction data for $CD4^+$ T cells were used (Mumbach et al.
286 2017). EPI predicted based on EPA were compared with three replications (B2T1,
287 B2T2, and B3T1) of HiChIP chromatin interaction data in $CD4^+$ T cells respectively.
288 The resolutions of HiChIP chromatin interaction data and EPI were adjusted to 5kb. EPI
289 were predicted based on three types of EPA: (i) EPA shortened at the genomic locations
290 of FR or RF orientation of DNA motif sequence of a TF, (ii) EPA shortened at the
291 genomic locations of DNA motif sequence of a TF acting as insulator sites such as
292 CTCF without considering their orientation, and (iii) EPA without being shortened by
293 the genomic locations of a DNA motif sequence. EPI were predicted based on the three
294 types of EPA using top 60% expression level of all transcripts (genes) excluding
295 transcripts not expressed in T cells. The criteria of the analysis were determined to
296 include known DNA motif sequences involved in chromatin interactions such as CTCF

297 and cohesin in the result. EPA (i) showed the highest ratio of EPI overlapped with
298 chromatin interactions (HiChIP) using DNA binding motif sequences of CTCF and
299 cohesin (RAD21 and SMC3), compared with the other two types of EPA (ii) and (iii) (*n*
300 = 4, binomial test, two-sided, 95% confidence interval) (Supplemental Fig. S2). Total
301 51 biased orientation (33 FR and 23 RF) of DNA motif sequences, which included
302 CTCF, cohesin (RAD21 and SMC3), SMARCC1, and NFI in three replications (B2T1,
303 B2T2, and B3T1) showed a significantly higher ratio of EPI overlapped with HiChIP
304 chromatin interactions (more than 1,000 counts for each interaction) than the other types
305 of EPAs in T cells (Table 4). When comparing EPI predicted based on only EPA (i) and
306 (iii) with the chromatin interactions, total 204 biased orientation (116 FR and 116 RF)
307 of DNA motif sequences showed a significantly higher ratio of EPI predicted based on
308 EPA (i) overlapped with the chromatin interactions than EPI predicted based on EPA
309 (iii) (Table 4; Supplemental material 2). The difference between EPI predicted based on
310 EPA (i) and (ii) seemed to be difficult to distinguish using the chromatin interaction
311 data and the statistical test in some cases. However, as for the difference between EPI
312 predicted based on EPA (i) and (iii), a larger number of biased orientations of DNA
313 motif sequences was found to be correlated with chromatin interaction data. Chromatin
314 interaction data were obtained from different samples from DNase-seq, open chromatin
315 regions, so individual differences seemed to exist from the results of this analysis. Most
316 of biased orientations of DNA motif sequences (97%) were found to be correlated with
317 chromatin interactions, when comparing EPI predicted based on EPA (i) and (iii) with
318 HiChIP chromatin interactions.

319 Moreover, to examine the enhancer activity of EPI, the distribution of expression
320 level of putative target genes of EPI was compared between EPI overlapped with
321 HiChIP chromatin interactions and EPI not overlapped with them. Though the target
322 genes of EPI were selected from top 60% expression level of all transcripts (genes)
323 excluding transcripts not expressed in T cells, target genes of EPI overlapped with
324 chromatin interactions showed a significantly higher expression level than EPI not
325 overlapped with them, suggesting that EPI overlapped with chromatin interactions
326 activated the expression of target genes in T cells. All FR and RF orientations of DNA
327 motifs showed a significantly higher expression level of putative target genes of EPI
328 overlapped with chromatin interactions than EPI not overlapped. When a biased
329 orientation of DNA motif showed a significantly higher expression level of putative
330 target genes of EPI overlapped with chromatin interactions than EPI not overlapped, '1'
331 was marked with in the tables of the comparison between EPI and HiChIP chromatin
332 interactions in Supplemental material 2. When a DNA motif showed a significantly
333 lower expression level, '-1' was marked with, and when there was not significant
334 difference of expression level, '0' was marked with. However, they were not observed
335 in this analysis. HiChIP data were produced using H3K27ac antibody, so chromatin
336 interactions acting as repressor would not be identified.

337 The ratio of EPI overlapped with HiChIP chromatin interactions was also
338 increased using H3K27ac marks. Chromatin interaction data were obtained from
339 different samples from DNase-seq, open chromatin regions in $CD4^+$ T cells, so
340 individual differences seemed to exist from the results of this analysis, and (Mumbach

341 et al. 2017) suggested that individual differences of chromatin interactions were larger
342 than those of open chromatin regions. ATAC-seq data, open chromatin regions were
343 available in $CD4^+$ T cells in the paper, however, when using ATAC-seq data, the result
344 of the analysis of biased orientations of DNA motif sequences was different from
345 DNase-seq data, and not included a part of CTCF and cohesin. Thus, DNase-seq data
346 collected from ENCODE and Blueprint projects were employed in this study.
347

348 **Correlation of expression level of genes closest to a DNA motif sequence**

349 To examine the effect of biased orientations of DNA motif sequences on gene
350 expression, I compared the expression level of protein-coding genes the most closely
351 located to upstream and downstream of biased orientations of DNA motif sequences in
352 human genome. Closely located genes are expected to show correlation of gene
353 expression among tissues (Xie et al. 2007). When a biased orientation of DNA motif
354 sequence acts as insulator (e.g. CTCF), the correlation of gene expression level would
355 be reduced. Among 69 biased orientations of DNA motifs conserved between
356 monocytes and T cells, top 15 were selected in ascending order on the median
357 correlation coefficient of expression level of the closest genes of the DNA motif that
358 was observed in more than 500 genomic locations. The number of genomic locations is
359 relatively small, due to the strict criteria of DNA motif search in this study (see
360 Methods). All 15 motifs showed a distribution of significantly lower correlation of gene
361 expression and included known DNA binding motif sequences of TF associated with
362 chromatin interactions such as CTCF, RAD21 and YY1 (Fig 3) (KS test, p-value <

363 0.01). Of 69 biased orientations of DNA motifs conserved between monocytes and T
364 cells, 47 (68%) showed a distribution of significantly lower correlation of gene
365 expression of the closest genes of the DNA motifs among 53 tissues than the correlation
366 of all pairs of genes, where genes are located the most closely to each other (KS test,
367 p-value < 0.05) (Table 5). When biased orientations of DNA motifs conserved between
368 monocytes and T cells were filtered based on the DNA motifs observed in more than
369 500 genomic locations, 40 of 52 (77%) DNA motifs showed a distribution of
370 significantly lower correlation of gene expression. With 1000 genomic locations, 32 of
371 41 (78%) DNA motifs showed a distribution of significantly lower correlation of gene
372 expression. With 2000 genomic locations, 18 of 23 (78%) DNA motifs showed a
373 distribution of significantly lower correlation of gene expression. According to the
374 increase of number of the DNA motifs observed in the genome sequences, the ratio of
375 the DNA motifs showing a significantly lower correlation was increased, suggesting
376 that the DNA motifs affect gene expression.
377

378 **Discussion**

379 To find DNA motif sequences of transcription factors (TF) and repeat DNA
380 sequences affecting the expression level of human putative transcriptional target genes,
381 the DNA motif sequences were searched from open chromatin regions of monocytes of
382 four people. Total 229 biased [130 forward-reverse (FR) and 128 reverse-forward (RF)]
383 orientation of DNA motif sequences of TF were found in monocytes of four people in
384 common, whereas only 11 any orientation (i.e. without considering orientation) of DNA

385 motif sequence of TF was found to affect the expression level of putative transcriptional
386 target genes, suggesting that enhancer-promoter association (EPA) shortened at the
387 genomic locations of FR or RF orientation of the DNA motif sequence of a TF or a
388 repeat DNA sequence is an important character for the prediction of enhancer-promoter
389 interactions (EPI) and the transcriptional regulation of genes.

390 When DNA motif sequences were searched from monocytes of one person, a
391 larger number of biased orientations of DNA motif sequences affecting the expression
392 level of human putative transcriptional target genes were found. When the number of
393 donors, from which experimental data were obtained, was increased, the number of
394 DNA motif sequences found in all people in common decreased and in some cases,
395 known transcription factors involved in chromatin interactions such as CTCF and
396 cohesin (RAD21 and SMC3) were not identified by statistical tests. This would be
397 caused by individual difference of the same cell type, low quality of experimental data,
398 and experimental errors. Moreover, though FR orientation of DNA binding motif
399 sequences of CTCF and cohesin is frequently observed at chromatin interaction anchors,
400 the percentage of FR orientation is not 100, and other orientations of the DNA binding
401 motif sequences are also observed (Guo et al. 2015). Though DNA binding motif
402 sequences of CTCF and cohesin are found in various open chromatin regions, DNA
403 binding motif sequences of some TF would be observed less frequently in open
404 chromatin regions. The analyses of experimental data of a number of people would
405 avoid missing relatively weak statistical significance of DNA motif sequences of TF in
406 experimental data of each person by multiple testing correction of thousands of

407 statistical tests. A DNA motif sequence was found with p-value < 0.05 in experimental
408 data of one person and the DNA motif sequence found in the same cell type of four
409 people in common would have p-value < $0.05^4 = 6.25 \times 10^{-6}$. Actually, DNA motif
410 sequences with p-value slightly less than 0.05 in monocytes of one person were
411 observed in monocytes of four people in common.

412 At first, EPI were compared with chromatin interactions (Hi-C) in monocytes.
413 Using open chromatin regions overlapped with ChIP-seq experimental data of histone
414 modification of an enhancer mark (H3K27ac), the ratio of EPI not overlapped with
415 Hi-C was reduced. (Phanstiel et al. 2017) also reported that there was an especially
416 strong enrichment for loops with H3K27 acetylation peaks at both ends (Fisher's Exact
417 Test, $p = 1.4 \times 10^{-27}$). However, the total number of EPI overlapped with chromatin
418 interactions was also reduced using H3K27ac peaks, so more chromatin interaction data
419 would be needed to obtain reliable results in this analysis. As an issue of experimental
420 data, data for chromatin interactions and open chromatin regions were obtained from
421 different samples and donors, so individual differences would exist in the data. The
422 resolution of chromatin interaction data used in monocytes was about 50kb, thus the
423 number of chromatin interactions was relatively small (72,284 at 50kb resolution with a
424 cutoff score of CHiCAGO tool > 1 and 16,501 with a cutoff score of CHiCAGO tool >
425 5) ('PCHiC_peak_matrix_cutoff0.txt.gz' file was downloaded from 'Promoter Capture
426 Hi-C in 17 human primary blood cell types' website https://osf.io/u8tzp/files/). To
427 examine the difference of the numbers of EPI overlapped with chromatin interactions,
428 according to the three types of EPAs, the number of chromatin interactions should be

429 large enough.

430 As HiChIP chromatin interaction data were available in $CD4^+$ T cells, biased
431 orientations of DNA motif sequences of TF were examined in T cells using DNase-seq
432 data of four people. The resolutions of chromatin interactions and EPI were adjusted to
433 5kb by fragmentation of genome sequences. The number of HiChIP chromatin
434 interactions was 19,926,360 at 5kb resolution, 666,149 at 5kb resolution with chromatin
435 interactions (more than 1,000 counts for each interaction), and 78,209 at 5kb resolution
436 with chromatin interactions (more than 6,000 counts for each interaction). As expected,
437 the number of EPI overlapped with chromatin interactions was increased, and 51 – 204
438 of biased orientations of DNA motif sequences of TF showed a statistical significance
439 in EPI predicted based on EPA shortened at the genomic locations of the DNA motif
440 sequence, compared with the other types of EPAs or EPA not shortened. False positive
441 prediction of EPI would be reduced by using H3K27ac marks and other features. The
442 ratio of EPI overlapped with HiChIP chromatin interactions was increased using
443 H3K27ac marks. However, the number of biased orientations of DNA motif sequences
444 showing a higher ratio of EPI overlapped with HiChIP chromatin interactions than the
445 other types of EPAs was decreased using H3K27ac marks (Supplemental material 2).
446 HiChIP analysis used H3K27ac antibody to capture chromatin interactions and
447 identified twenty millions of chromatin interactions, but the number of peaks of
448 ChIP-seq data of H3K27ac is usually less than a hundred thousand. There seems to be
449 much difference of the number of H3K27ac sites between HiChIP and ChIP-seq, so
450 ChIP-seq of H3K27ac may not cover all H3K27ac sites.

451 When forming a homodimer or heterodimer with another TF, TF may bind to
452 genome DNA with a specific orientation of their DNA binding sequences (Fig. 4). From
453 the analysis of biased orientations of DNA motif sequences of TF, TF forming
454 heterodimer would also be found. If the DNA binding motif sequence of only the mate
455 to a pair of TF was found in EPA, EPA was shortened at one side, which is the genomic
456 location of the DNA binding motif sequence of the mate to the pair, and transcriptional
457 target genes were predicted using the EPA shortened at the side. In this analysis, the
458 mate to both heterodimer and homodimer of TF can be used to examine the effect on
459 the expression level of transcriptional target genes predicted based on the EPA
460 shortened at one side. For heterodimer, biased orientations of DNA motif sequences
461 may also be found in forward-forward or reverse-reverse orientation.

462 In general, DNA motif sequences of TF were searched from genome sequences
463 using position weight matrix or position frequency matrix and allowing mismatch of
464 DNA sequences under a certain threshold. However, the change of the threshold
465 affected the result of analyses and it seems to be difficult to find the best parameter
466 setting to reduce false-positive and false-negative predictions of DNA biding sites of TF.
467 Therefore, to obtain a robust result in this study, position weight matrixes of DNA
468 motifs were converted into consensus DNA sequences including degenerated sites,
469 which consist of A, T, G, C, R, Y, M, K, S, W, B, H, V, D, and N letters. This does not
470 need parameter setting in the search for DNA motif sequences, and is the same as *k*-mer
471 analysis of DNA motif sequences of TF. Thought the number of genomic regions
472 matched with a DNA motif sequence was decreased, about 200 FR and RF orientations

473 of DNA motif sequences of TF were found, and the number of any orientation of DNA
474 motif sequences was quite small. Though the number of biased orientations of DNA
475 motif sequences of TF was decreased, the number of the DNA motif sequences found in
476 both monocytes and $CD4^+$ T cells in common was increased, suggesting that
477 false-positive prediction of DNA biding sites of TF was reduced. However, the length
478 of the consensus DNA sequences of some TF is relatively long. When the core DNA
479 binding sequence of a TF is short (5 or 6 bp), it may be strict to find perfect matches of
480 the long consensus DNA sequence in genome sequences, not allowing a mismatch of a
481 DNA sequence. To obtain more accurate result of analysis, I plan to refine the method
482 to find biased orientations of DNA motif sequences of TF.

483 It has been reported that CTCF and cohesin-binding sites are frequently mutated
484 in cancer (Katainen et al. 2015). Some biased orientations of DNA motif sequences
485 would be associated with chromatin interactions and might be associated with diseases
486 including cancer.

487 The analyses in this study revealed novel characters of DNA binding motif
488 sequences of TF and repeat DNA sequences to analyze TF involved in chromatin
489 interactions, insulator function and forming a homodimer, heterodimer or complex with
490 other TF, affecting the transcriptional regulation of genes.

491

492 ## Methods

493 **Search for biased orientations of DNA motif sequences**

494 To examine transcriptional regulatory target genes of transcription factors (TF),

495 bed files of hg38 of Blueprint DNase-seq data for $CD14^+$ monocytes of four donors
496 (EGAD00001002286; Donor ID: C0010K, C0011I, C001UY, C00408) were obtained
497 from Blueprint project web site (http://dcc.blueprint-epigenome.eu/#/home), and the bed
498 files of hg38 were converted into those of hg19 using Batch Coordinate Conversion
499 (liftOver) web site (https://genome.ucsc.edu/util.html).

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

514 To identify transcription factor binding sites (TFBS) from the DNase-seq data,
515 TRANSFAC (2019.2), JASPAR (2018), UniPROBE (2018), high-throughput SELEX,
516 transcription factor binding sequences of ENCODE ChIP-seq data, and HOCOMOCO

517 version 9 and 11 were used to predict insulator sites (Wingender et al. 1996; Newburger
518 and Bulyk 2009; Portales-Casamar et al. 2010; Xie et al. 2010; Zhao and Stormo 2011;
519 Jolma et al. 2013; Kheradpour and Kellis 2014; Kulakovskiy et al. 2018). TRANSFAC
520 (2019.2) was used to analyze enhancer-promoter interactions, since these data were
521 sufficient to identify biased orientations of DNA motif sequences of TF with less
522 computational time, reducing the number of any orientation of DNA motif sequences of
523 TF. Position weight matrices of transcription factor binding sequences were converted
524 into DNA consensus sequences considering degenerated sites using convert matrix in
525 RSAT website (van Helden 2003). Transcription factor binding sequences of TF
526 derived from vertebrates were used for further analyses. Transcription factor binding
527 sequences were searched from each narrow peak of DNase-seq data in repeat-masked
528 hg19 genome sequences using Match tool in TRANSFAC with similarity score cutoff
529 of 1 after transforming the DNA consensus sequences of TF into position frequency
530 matrixes for Match search (Kel et al. 2003). Repeat-masked hg19 genome sequences
531 were downloaded from UCSC genome browser (http://genome.ucsc.edu/,
532 http://hgdownload.soe.ucsc.edu/goldenPath/hg19/bigZips/hg19.fa.masked.gz). TF
533 corresponding to transcription factor binding sequences were searched computationally
534 and manually by comparing their names and gene symbols of HGNC (HUGO Gene
535 Nomenclature Committee) -approved gene nomenclature and 31,848 UCSC known
536 canonical transcripts
537 (http://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/knownCanonical.txt.gz), as
538 transcription factor binding sequences were not linked to transcript IDs such as UCSC,

539 RefSeq, and Ensembl transcripts.

540 Target genes of a TF were assigned when its TFBS was found in DNase-seq
541 narrow peaks in promoter or extended regions for enhancer-promoter association of
542 genes (EPA). Promoter and extended regions were defined as follows: promoter regions
543 were those that were within distance of ±5 kb from transcriptional start sites (TSS).
544 Promoter and extended regions were defined as per the following association rule,
545 which is the same as that defined in Figure 3A of a previous study (McLean et al.
546 2010): the single nearest gene association rule, which extends the regulatory domain to
547 the midpoint between the TSS of the gene and that of the nearest gene upstream and
548 downstream without the limitation of extension length. Extended regions for EPA were
549 shortened at the genomic locations of DNA binding sites of a TF that was the closest to
550 a transcriptional start site, and transcriptional target genes were predicted from the
551 shortened enhancer regions using TFBS. Furthermore, promoter and extended regions
552 for EPA were shortened at the genomic locations of forward–reverse (FR) orientation of
553 DNA binding sites of a TF. When forward or reverse orientation of DNA binding sites
554 were continuously located in genome sequences several times, the most external
555 forward–reverse orientation of DNA binding sites were selected. The genomic positions
556 of genes were identified using 'knownGene.txt.gz' file in UCSC bioinformatics sites
557 (Karolchik et al. 2014). The file 'knownCanonical.txt.gz' was also utilized for choosing
558 representative transcripts among various alternate forms for assigning promoter and
559 extended regions for EPA. From the list of transcription factor binding sequences and
560 transcriptional target genes, redundant transcription factor binding sequences were

561 removed by comparing the target genes of a transcription factor binding sequence and
562 its corresponding TF; if identical, one of the transcription factor binding sequences was
563 used. When the number of transcriptional target genes predicted from a transcription
564 factor binding sequence was less than five, the transcription factor binding sequence
565 was omitted.

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

570 For gene expression data, RNA-seq reads mapped onto human hg19 genome
571 sequences were obtained from UCSF-UBC human reference epigenome mapping
572 project RNA-seq reads with poly-A of naive $CD4^+$ T cells (GSM669617). FPKMs of
573 the RNA-seq data were calculated using RSeQC (Wang et al. 2012). For monocytes,
574 Blueprint RNA-seq FPKM data
575 ('C0010KB1.transcript_quantification.rsem_grape2_crg.GRCh38.20150622.results',
576 'C0011IB1.transcript_quantification.rsem_grape2_crg.GRCh38.20150622.results', and
577 'C001UYB4.transcript_quantification.rsem_grape2_crg.GRCh38.20150622.results')
578 were downloaded from Blueprint DCC portal
579 (http://dcc.blueprint-epigenome.eu/#/files). However, RNA-seq FPKM data for C00408
580 was not available, so RNA-seq FPKM data for C001UY was used for C00408. Based
581 on log2-transformed FPKM, UCSC transcripts were arranged in descending order of
582 expression level, and top 4,000 transcripts were selected in each cell type.

583 The expression level of transcriptional target genes predicted based on EPA
584 shortened at the genomic locations of DNA motif sequence of a TF or a repeat DNA
585 sequence was compared with the expression level of transcriptional target genes
586 predicted from promoter. For each DNA motif sequence shortening EPA,
587 transcriptional target genes were predicted using about 3,000 – 5,000 DNA binding
588 motif sequences of TF, and the distribution of expression level of putative
589 transcriptional target genes of each TF was compared between EPA and only promoter
590 using Mann-Whitney test, two-sided (p-value < 0.05). The number of TF showing a
591 significant difference of expression level of putative transcriptional target genes
592 between EPA and promoter was compared among forward-reverse (FR),
593 reverse-forward (RF), and any orientation (i.e. without considering orientation) of a
594 DNA motif sequence shortening EPA using Fisher's exact test (p-value < 0.05). When a
595 DNA motif sequence of a TF or a repeat DNA sequence shortening EPA showed a
596 significant difference of expression level of putative transcriptional target genes among
597 FR, RF, or any orientation in monocytes of four people in common, the DNA motif
598 sequence was listed.

599 Though forward-reverse orientation of DNA binding motif sequences of CTCF
600 and cohesin are frequently observed at chromatin interaction anchors, the percentage of
601 forward-reverse orientation is not 100, and other orientations of the DNA binding motif
602 sequences are also observed (Guo et al. 2015). Though DNA binding motif sequences
603 of CTCF and cohesin are found in various open chromatin regions, DNA binding motif
604 sequences of some transcription factors would be observed less frequently in open

605 chromatin regions. The analyses of experimental data of a number of people would
606 avoid missing relatively weak statistical significance of DNA motif sequences in
607 experimental data of each person by multiple testing correction of thousands of
608 statistical tests. A DNA motif sequence was found with p-value < 0.05 in experimental
609 data of one person and the DNA motif sequence found in the same cell type of four
610 people in common would have p-value $< 0.05^4 = 6.25 \times 10^{-6}$.
611
612 **Co-location of biased orientations of DNA motif sequences**
613 Co-location of biased orientations of DNA binding motif sequences of TF was
614 examined. The number of open chromatin regions with the same pair of DNA binding
615 motif sequences was counted, and when the pair of DNA binding motif sequences were
616 enriched with statistical significance (Fisher's exact test, p-value $< 1.0 \times 10^{-9}$), they
617 were listed.
618
619 **Comparison with chromatin interaction data**
620 For comparison of EPA with chromatin interactions (HiChIP) in $CD4^+$ T cells,
621 'GSM2705049_Naive_HiChIP_H3K27ac_B2T1_allValidPairs.txt',
622 'GSM2705050_Naive_HiChIP_H3K27ac_B2T2_allValidPairs.txt', and
623 'GSM2705051_Naive_HiChIP_H3K27ac_B3T1_allValidPairs.txt' files were
624 downloaded from Gene Expression Omnibus (GEO) database (GSM2705049,
625 GSM2705050 and GSM2705051). The resolutions of chromatin interaction data and
626 EPI were adjusted to 5kb before their comparison. Chromatin interactions with more

627 than 6,000 and 1,000 counts for each interaction were used in this study.

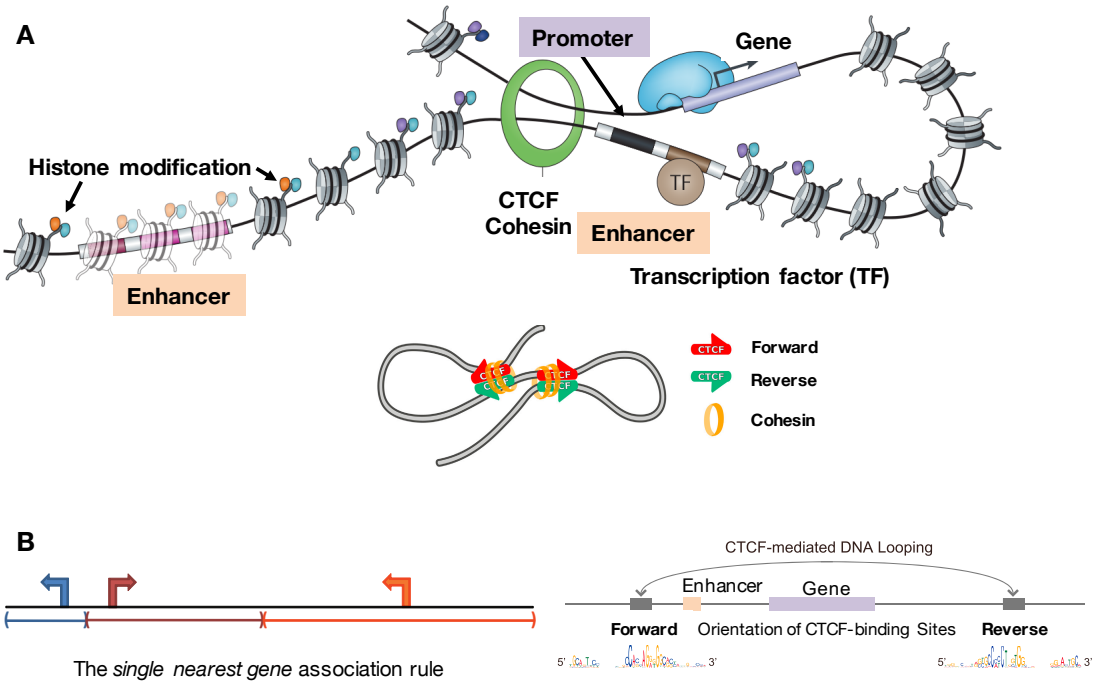
628 Enhancer-promoter interactions (EPI) were predicted using three types of EPAs
629 in monocytes: (i) EPA shortened at the genomic locations of FR or RF orientation of
630 DNA motif sequence of a TF, (ii) EPA shortened at the genomic locations of any
631 orientation (i.e. without considering orientation) of DNA motif sequence of a TF, and
632 (iii) EPA without being shortened by a DNA motif sequence. EPI predicted using the
633 three types of EPAs in common were removed. First, EPI predicted based on EPA (i)
634 were compared with chromatin interactions. The resolution of chromatin interaction
635 data used in this study was 1-5kb, so EPI were adjusted to 5kb before their comparison.
636 The number and ratio of EPI overlapped with chromatin interactions were counted.
637 Second, EPI were predicted based on EPA (ii), and EPI predicted based on EPA (i)
638 were removed from the EPI. The number and ratio of EPI overlapped with chromatin
639 interactions were counted. Third, EPI were predicted based on EPA (iii), and EPI
640 predicted based on EPA (i) and (ii) were removed from the EPI. The number and ratio
641 of EPI overlapped with chromatin interactions were counted. The number and ratio of
642 the EPI were compared two times between EPA (i) and (iii), and EPA (i) and (ii)
643 (binomial distribution, p-value < 0.025 for each test, two-sided, 95% confidence
644 interval).

645 Putative target genes for the analysis of EPI were selected from top 50%
646 expression level of all transcripts excluding transcripts not expressed in monocytes and
647 top 60% expression level of transcripts in $CD4^+$ T cells. The expression level of putative
648 target genes of EPI overlapped with HiChIP chromatin interactions was compared with

649 EPI not overlapped with them. For each FR or RF orientation of DNA motif, EPI were
650 predicted based on EPA and the overlap of EPI with chromatin interactions was
651 examined. When a putative transcriptional target gene of a TF in an enhancer was found
652 in both EPI overlapped with a chromatin interaction and EPI not overlapped with, the
653 target gene was removed. The distribution of expression level of putative target genes
654 was compared using Mann-Whitney test, two-sided (p-value < 0.05).
655
656 **Correlation of gene expression of closest genes to a DNA motif sequence**
657 To examine the correlation of expression level of genes closest to a DNA motif
658 sequence, gene expression data among 53 tissues were obtained from
659 'GTEx_Analysis_2016-01-15_v7_RNASeQCv1.1.8_gene_median_tpm.gct.gz' file in
660 GTEx database (https://gtexportal.org/home/). The correlation of log2-transformed gene
661 expression level was analyzed based on protein-coding transcripts of RefSeq (RefSeq
662 ID with 'NM_'). The same genomic location of RefSeq transcripts was removed. DNA
663 motifs overlapped with RefSeq transcripts were eliminated. Genes closest to DNA
664 motifs and other genes were analyzed using BEDOPS closest-features (Neph et al.
665 2012). KS test was performed using R.
666
667 ## Acknowledgements
668 The supercomputing resource was provided by Human Genome Center of the Institute
669 of Medical Science at the University of Tokyo. Computations were partially performed
670 on the NIG supercomputer at ROIS National Institute of Genetics. Publication charges

671 for this article were funded by JSPS KAKENHI Grant Number 16K00387. This
672 research was partially supported by the Platform Project for Supporting in Drug
673 Discovery and Life Science Research（Platform for Dynamic Approaches to Living
674 System）from Japan Agency for Medical Research and Development (AMED). This
675 research was partially supported by Development of Fundamental Technologies for
676 Diagnosis and Therapy Based upon Epigenome Analysis from Japan Agency for
677 Medical Research and Development (AMED). This work was partially supported by
678 JST CREST Grant Number JPMJCR15G1, Japan.
679

680 **Figures**



681

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

692

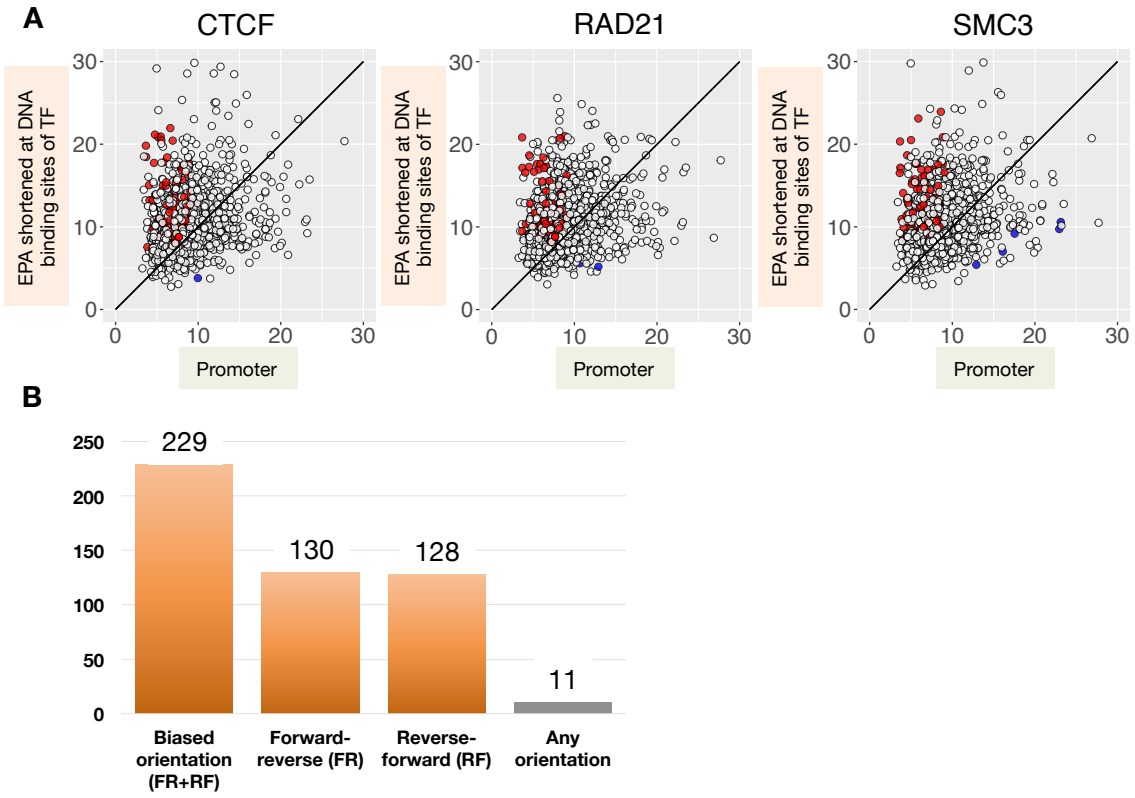
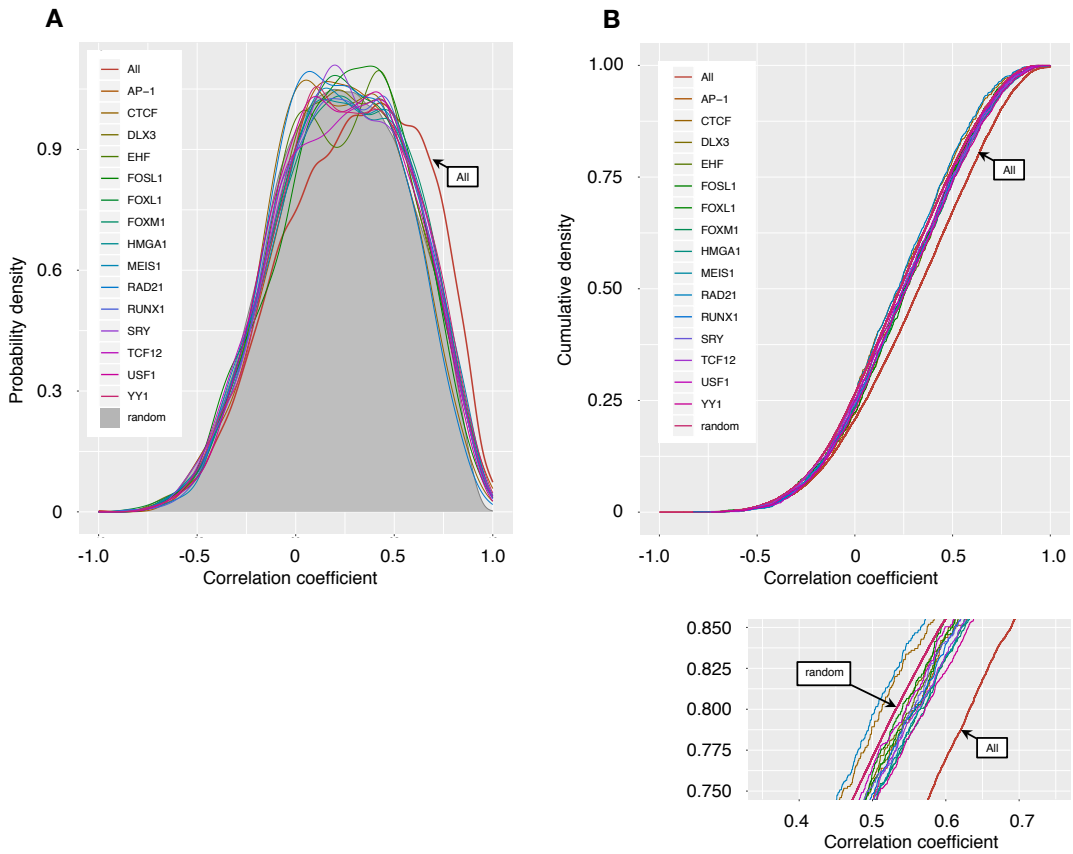


Figure 2. Biased orientations of DNA motif sequences of transcription factors affecting the transcription of genes. (A) Comparison of expression level of putative transcriptional target genes. The median expression levels of target genes of the same transcription factor binding sequences were compared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

705 expression level of target genes predicted based on the EPA tended to be higher than
706 promoters, implying that transcription factors acted as activators of target genes. (B)
707 Biased orientations of DNA motif sequences of transcription factors in monocytes.
708 Total 229 of biased orientations of DNA binding motif sequences of transcription
709 factors were found to affect the expression level of putative transcriptional target genes
710 in monocytes of four people in common, whereas only 11 any orientation (i.e. without
711 considering orientation) of DNA binding motif sequence was found to affect the
712 expression level.
713



714

715 **Figure 3.** Genes separated by predicted CTCF sites are less correlated in gene

716 expression. Correlation coefficient between neighboring gene pairs is shown in terms of

717 probability density (A) and cumulative distribution (B). Brown line (All), correlation

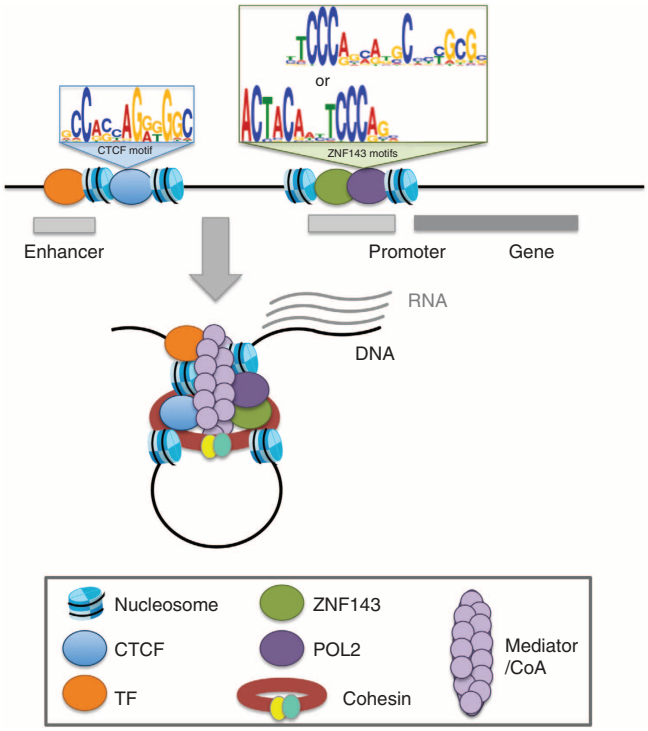
718 between all neighboring genes; gray shading in (A) and red line in (B), correlation

719 between randomly chosen gene pairs.

720

721

722



723

724 **Figure 4.** Schematic representation of chromatin interactions involving gene promoters.

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

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

727 2015).

728

729 **Tables**

730 **Table 1.** Top 90 biased orientations of DNA binding motif sequences of transcription

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

732 $-\log_{10}$ (*p*-value).

733 Forward-reverse orientation

TF	Score	TF	Score	TF	Score
TCF3	123.12	ETS	49.44	DDIT3	29.23
ZNF592	120.55	SOX6	49.42	NRF1	28.86
EOMES	109.78	AP2A	49.36	SREBP2	28.63
GMEB2	75.32	CEBPA	47.31	YY1	28.02
CEBPG	74.27	FOSL1	44.41	FOXM1	27.64
FOXA1	73.14	ATOH1	43.94	VSX1	27.38
FOXA2	72.93	NFATC2	43.63	AP-1	27.13
DBX2	72.66	TBX3	42.82	STAT5A	26.61
NFYA	71.62	CREB	42.45	KLF10	26.24
TP53	66.60	EGR1	41.57	mix-a	26.06
E2F4	65.71	HOXA4	41.47	HLF	25.64
TLX2	64.10	FOXO1	40.92	FOXG1	25.44
IKZF1	63.69	BARHL1	40.43	DLX3	24.92
RFX6	61.62	ZNF189	39.78	ZNF180	24.86
SFN	59.89	SP1	39.54	BARX2	24.45
FOXO3	59.78	ATF3	37.30	REST	24.25
ZBTB18	59.45	CTNNB1	37.01	SMAD	24.05
PPARG	57.25	ZNF579	36.10	RAD21	23.98
ZNF675	54.71	MZF1	35.58	SRY	23.97
HNF1B	53.86	BACH2	35.38	LRF	23.46
ITGB2	53.86	ZIC1	34.59	ZNF300	22.62
RFX4	53.41	CTCFL	34.55	FOXJ2	22.35
TBP	53.06	CXXC1	34.42	TCF7	22.18
CBFB	52.31	GCM2	34.26	MNX1	21.99
TEAD1	51.39	MEIS1:EOMES	34.02	SOX	21.98
ZBTB7A	50.61	HNF4	34.00	MAZ	21.87
CTCF	50.54	HINFP	33.29	HMGIY	21.53
SOX5	50.51	PDX1	32.36	STAT4	21.33
BRCA1	50.47	CEBPD	30.87	TBR1	21.22
SOX3	49.47	RUNX1	29.40	NKX2-5	21.17

734
735

Reverse-forward orientation

TF	Score	TF	Score	TF	Score
ZBTB48	175.35	HOXA11	47.37	CEBP	32.34
ZBTB38	107.55	POU4F3	46.76	RUNX3	32.09
MSX1	105.86	LBX1	46.13	TCF7	31.09
POLYA	98.31	NR5A1	44.93	AR	30.08
RFX5	97.45	ETV5	44.91	ZBTB7B	29.59
SMAD1	83.39	FOXL1	44.50	ESX1	29.45
YBX1	80.34	NFATC2	43.63	CTNNB1	29.43
KLF4	79.72	NR3C1	42.45	HDAC2	29.27
NR2C2	78.27	ETV6	41.44	HIC1	29.16
OTX	74.43	TFAP2A	40.97	NRF1	29.02
CEBPG	74.27	TCF7L2	40.86	HOXD3	28.50
SP1	73.86	PGR	40.58	ZFP42	28.20
BARHL1	72.65	SP2	39.15	NR4A1	27.71
PITX3	71.71	EBF1	38.13	NR4A3	25.50
TEAD3	68.36	ZNF384	37.81	LMX1A	25.31
FOXA2	68.31	ISX	37.47	NFATC4	25.26
ZBTB16	68.09	GFI1	36.99	PAX2	25.04
EN2	64.71	ATF6	36.81	RAX2	24.73
HOXA7	64.52	REST	36.52	FOXD2	24.55
ZBTB7A	63.80	NFYB	36.38	RHOXF1	23.71
NFATC1	61.71	FOS	36.12	CTCFL	23.46
KLF2	61.64	RUNX2	35.21	CREB1	23.28
TEAD1	61.03	YY1	34.76	EN1	22.00
STAT	56.11	CREB3	34.64	SPIB	21.81
SOX6	55.25	MEIS1	34.49	NFAT	20.97
TFAP2C	53.06	CTCF	33.99	LEF1	20.94
HOXA2	52.05	EGR1	33.22	MYB	20.77
FOXO4	51.30	ISL2	33.21	NKX3-2	20.72
ZBTB18	49.30	RUNX1	33.17	CCAAT	20.42
SPDEF	49.26	PDX1	32.57	FLI1	19.65

Table 2. Biased orientation of repeat DNA sequences in monocytes. Score: $-\log_{10}$ (*p*-value).

Reverse-forward (RF) orient

Repeat DNA seq.	Score
AluJo	85.18
AluJb	17.73
L1M5	12.24

Reverse-forward (RF) orientation (using H3K27ac histone modification marks)

Repeat DNA seq.	Score
AluJb	26.27

Table 3. Top 30 of co-locations of biased orientations of DNA binding motif sequences of transcription factors in monocytes. Co-locations of DNA motif sequence of CTCF with another biased orientation of DNA motif sequence were shown in a separate table. Motif 1,2: DNA binding motif sequences of transcription factors. # both: the number of open chromatin regions including both Motif 1 and Motif 2. # motif 1: the number of open chromatin regions including Motif 1. # motif 2: the number of open chromatin regions including Motif 2. # others: the number of open chromatin regions not including Motif 1 and Motif 2. Total number of open chromatin regions is 145,866.

Motif 1	Motif 2	# both	# motif 1	# motif 2	# others	p-value
SOX	TCF7L2	21268	0	129	124469	0
LEF1	TCF7L2	21268	0	126	124472	0
LEF1	SOX	21268	129	126	124343	0
DLX5	PRRX2	20580	0	128	125158	0
NFAT	NFATC4	16105	0	110	129651	0
ETV4	FLI1	15046	3276	0	127544	0
LEF1	SOX10	14773	24783	6621	99689	0
SOX	SOX10	14772	24784	6625	99685	0
SOX10	TCF7L2	14744	6524	24812	99786	0
ELF5	ETV4	13205	5117	6384	121160	0
ELF5	FLI1	12971	6618	2075	124202	0
RUNX2	RUNX3	12875	3088	174	129729	0
RUNX1	RUNX3	12866	0	183	132817	0
RUNX1	RUNX2	12866	3097	0	129903	0
HMGA1	HMX1	12611	0	15149	118106	0
NFATC1	NFATC3	10877	231	0	134758	0
NFATC1	NFATC4	10831	5274	277	129484	0
NFAT	NFATC1	10831	277	5384	129374	0
NFAT	NFATC3	10815	62	5400	129589	0
NFATC3	NFATC4	10815	5290	62	129699	0
BHLHE40	MITF	10319	65	19796	115686	0
HOXA2	MEOX2	9999	0	125	135742	0
EMX1	MEOX2	9999	0	125	135742	0
BCL6	HMGA1	7648	4963	32936	100319	0
MEIS1	MEIS3	6973	4122	13183	121588	0
BHLHE40	TCF12	6343	23772	57	115694	0
HMGA1	STAT5A	4835	34254	2003	104774	0
FOXD2	FOXL1	4606	0	6432	134828	0
DLX3	DLX5	4561	24	16147	125134	0
DLX5	MSX1	4561	24	16147	125134	0

Motif 1	Motif 2	# both	# motif 1	# motif 2	# others	p-value
CTCF	RAD21	1450	0	1306	143110	0
CTCF	CTCFL	1061	389	399	144017	0
CTCF	ZBTB7A	634	14	7766	137452	0
CTCF	ZBTB7	633	517	15	144701	0
CTCF	ZNF300	505	8862	143	136356	0
CTCFL	RAD21	1459	1	1297	143109	0
CTCFL	ZBTB7A	1150	0	7250	137466	0
CTCFL	ZNF300	900	250	8467	136249	0
BHLHE40	CTCF	606	389	29509	115362	8.06E-169

771 **Table 4.** Comparison of enhancer-promoter interactions (EPI) with chromatin
772 interactions in T cells. The upper tables show the number of biased orientations of DNA
773 motifs, where a significantly higher ratio of EPI, which were predicted based on
774 enhancer-promoter association (EPA) (i), overlapped with HiChIP chromatin interaction
775 data than the other types of EPA (ii) and (iii). The middle tables show the 33 FR and 23
776 RF orientations of DNA motifs found in common among B2T1, B2T2 and B3T1
777 replications in the upper table. The lower table shows that among 69 biased orientations
778 of DNA motifs found in both monocytes and T cell, 24 were matched with the analysis
779 of HiChIP for three types of EPA. TF: DNA binding motif sequence of a transcription
780 factor. Score: $-\log_{10}$ (*p*-value). Inf: *p*-value = 0. Score 1: Comparison of EPA shortened
781 at the genomic locations of FR or RF orientation of DNA motif sequence [EPA (i)] with
782 EPA not shortened [EPA (iii)]. Score 2: Comparison of EPA shortened at the genomic
783 locations of FR or RF orientation of DNA motif sequence [EPA (i)] with EPA
784 shortened at the genomic locations of any orientation of DNA motif sequence [EPA
785 (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	71	42	34
B2T2	81	47	40
B3T1	94	55	47
DNA motifs found in common among replications			
B2T1 and B2T2	60	35	30
B2T1 and B3T1	59	38	26
B2T2 and B3T1	68	42	32
B2T1, B2T2 and B3T1	51	33	23

2. Comparison of EPI with HiChIP data between 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	205	116	117
B2T2	205	116	117
B3T1	208	119	119
DNA motifs found in common among replications			
B2T1 and B2T2	204	116	116
B2T1 and B3T1	205	116	117
B2T2 and B3T1	204	116	116
B2T1, B2T2 and B3T1	204	116	116

Forward-reverse (FR)

TF	Score 1	Score 2
AP-1	81.68	15.19
AP2A	142.68	5.30
BPTF	59.08	3.31
CTCF	323.31	7.60
CTCFL	323.31	2.44
DBP	118.88	18.18
DLX3	30.22	4.12
ETS1	82.43	2.00
FOSB	79.38	14.49
FOSL1	42.29	7.04
IKZF1	83.93	8.70
JUN	39.30	5.81
JUNB	110.84	3.66
KLF4	37.29	15.02
KLF5	68.20	9.30
LBX1	154.16	2.69
NEUROD2	48.90	4.73
NFIL3	152.37	2.69
NKX2-2	198.62	2.31
NKX2-5	136.48	8.97
NR2F1	82.58	7.16
NR3C1	77.96	17.53
RAD21	53.81	7.70
RFX7	29.86	1.93
SMARCC1	39.89	6.15
SMC3	300.51	23.65
SPI1	219.80	7.68
SREBF2	68.43	2.84
TEAD4	45.79	10.37
VDR:RXRA	82.58	7.16
ZBTB7A	101.61	5.36
ZEB1	107.06	5.51
ZXDC	23.35	6.26

Reverse-forward (RF)

TF	Score 1	Score 2
AP-1	79.38	14.49
CACD	67.66	8.44
CRX	56.13	4.27
CTCF	5.65	3.43
CTCFL	101.61	5.36
ETV4	103.13	6.18
FOS	76.64	14.21
FOSL1	39.62	6.42
FOSL2	42.00	6.41
FOXA2	323.31	1.69
GATA	293.73	3.88
HDAC1	323.31	11.46
HEY2	257.47	3.23
HIC2	62.15	14.91
JUND:FOSL1	40.68	6.95
MEIS1	72.99	10.02
MEQ	118.91	5.86
NFI	101.48	10.50
NFIL3	114.14	2.15
SOX4	151.86	4.45
TFAP2C	136.94	5.62
YBX1	111.24	2.03
ZNF354C	54.72	6.02

791 Forward-reverse (FR)

TF	Score 1	Score 2
AP-1	81.68	15.19
AP2A	142.68	5.30
CTCF	323.31	7.60
CTCFL	323.31	2.44
DLX3	30.22	4.12
ETS1	82.43	2.00
FOSL1	42.29	7.04
IKZF1	83.93	8.70
JUN	39.30	5.81
KLF4	37.29	15.02
NKX2-5	136.48	8.97
RAD21	53.81	7.70
SPI1	219.80	7.68
ZBTB7A	101.61	5.36
ZEB1	107.06	5.51

Reverse-forward (RF)

TF	Score 1	Score 2
AP-1	79.38	14.49
CTCF	5.65	3.43
CTCFL	101.61	5.36
FOS	76.64	14.21
FOSL1	39.62	6.42
FOXA2	323.31	1.69
MEIS1	72.99	10.02
TFAP2C	136.94	5.62
YBX1	111.24	2.03

792

793

794

795 **Table 5.** Correlation of expression level of genes closest to a DNA motif sequence

796 among 53 tissues. Score: $-\log_{10}$ (*p*-value) of KS test. Median correlation: median

797 correlation coefficient of expression level of nearby genes of a DNA motif. The median

798 correlation coefficient of expression level of all pairs of nearby genes is 0.33. # DNA

799 site: number of genomic locations of DNA motif sequence of TF where upstream and

800 downstream genes closest to the motif sequence have expression data.

TF	Score	Median correlation	# DNA site
AP-1	3.03	0.26	538
AP2A	1.86	0.29	1209
BARHL1	5.16	0.27	2360
BRCA1	2.94	0.27	1260
CEBPB	3.27	0.27	764
CREB1	2.15	0.30	1264
CTCF	9.33	0.23	878
CTCFL	3.09	0.27	548
CXXC1	1.60	0.31	3605
DLX3	5.72	0.25	1288
EHF	5.13	0.27	1230
ETS1	1.46	0.28	906
FOS	2.73	0.29	2917
FOSL1	4.65	0.26	1003
FOXA2	2.95	0.26	361
FOXL1	5.52	0.25	1277
FOXM1	4.58	0.27	1459
FOXO4	2.05	0.24	223
HDAC2	2.11	0.19	67
HMGA1	5.46	0.26	2655
HNF4	4.96	0.27	3508
IKZF1	1.91	0.29	3988
JUN	4.72	0.27	1328
LEF1	4.99	0.27	4036

TF	Score	Median correlation	# DNA site
MEIS1	5.56	0.26	2680
NFATC1	4.15	0.28	2613
NFE2L1	5.58	0.27	3190
NKX2-5	3.74	0.27	1205
NKX3-2	2.21	0.29	2303
NKX6-1	5.21	0.27	2998
NR2C2	2.15	0.21	96
NR4A1	3.60	0.28	1869
PAX8	2.16	0.23	48
PDX1	5.21	0.27	2546
PRDM1	2.39	0.24	77
RAD21	11.74	0.22	1220
REST	3.58	0.27	735
RUNX1	6.54	0.26	2976
SOX10	4.38	0.28	5732
SPI1	3.26	0.27	579
SPIB	5.00	0.27	4705
SRY	4.78	0.24	883
TCF12	6.42	0.26	2101
TCF3	2.77	0.26	239
TFAP2C	1.83	0.29	1229
USF1	7.23	0.25	1638
YY1	4.90	0.26	2463

801

802

803 **References**

804

805 Bailey SD, Zhang X, Desai K, Aid M, Corradin O, Cowper-Sal Lari R, Akhtar-Zaidi B,
806 Scacheri PC, Haibe-Kains B, Lupien M. 2015. ZNF143 provides sequence
807 specificity to secure chromatin interactions at gene promoters. *Nat Commun* **2**:
808 6186.

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

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

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

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

822 de Souza FS, Franchini LF, Rubinstein M. 2013. Exaptation of transposable elements
823 into novel cis-regulatory elements: is the evidence always strong? *Mol Biol Evol*
824 **30**: 1239-1251.

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

828 Gaussin A, Modlich U, Bauche C, Niederlander NJ, Schambach A, Duros C, Artus A,
829 Baum C, Cohen-Haguenauer O, Mermod N. 2012. CTF/NF1 transcription
830 factors act as potent genetic insulators for integrating gene transfer vectors.
831 *Gene Ther* **19**: 15-24.

832 Guo Y, Xu Q, Canzio D, Shou J, Li J, Gorkin DU, Jung I, Wu H, Zhai Y, Tang Y et al.
833 2015. CRISPR Inversion of CTCF Sites Alters Genome Topology and
834 Enhancer/Promoter Function. *Cell* **162**: 900-910.

835 Javierre BM, Burren OS, Wilder SP, Kreuzhuber R, Hill SM, Sewitz S, Cairns J,
836 Wingett SW, Varnai C, Thiecke MJ et al. 2016. Lineage-Specific Genome
837 Architecture Links Enhancers and Non-coding Disease Variants to Target Gene
838 Promoters. *Cell* **167**: 1369-1384 e1319.
839 Jeong M, Huang X, Zhang X, Su J, Shamim M, Bochkov I, Reyes J, Jung H, Heikamp
840 E, Presser Aiden A et al. 2017. A Cell Type-Specific Class of Chromatin Loops
841 Anchored at Large DNA Methylation Nadirs. *bioRxiv*.
842 Ji X, Dadon DB, Abraham BJ, Lee TI, Jaenisch R, Bradner JE, Young RA. 2015.
843 Chromatin proteomic profiling reveals novel proteins associated with
844 histone-marked genomic regions. *Proc Natl Acad Sci U S A* **112**: 3841-3846.
845 Jin C, Kato K, Chimura T, Yamasaki T, Nakade K, Murata T, Li H, Pan J, Zhao M, Sun
846 K et al. 2006. Regulation of histone acetylation and nucleosome assembly by
847 transcription factor JDP2. *Nat Struct Mol Biol* **13**: 331-338.
848 Jjingo D, Conley AB, Wang J, Marino-Ramirez L, Lunyak VV, Jordan IK. 2014.
849 Mammalian-wide interspersed repeat (MIR)-derived enhancers and the
850 regulation of human gene expression. *Mob DNA* **5**: 14.
851 John S, Sabo PJ, Thurman RE, Sung MH, Biddie SC, Johnson TA, Hager GL,
852 Stamatoyannopoulos JA. 2011. Chromatin accessibility pre-determines
853 glucocorticoid receptor binding patterns. *Nature genetics* **43**: 264-268.
854 Jolma A, Yan J, Whitington T, Toivonen J, Nitta KR, Rastas P, Morgunova E, Enge M,
855 Taipale M, Wei G et al. 2013. DNA-binding specificities of human transcription
856 factors. *Cell* **152**: 327-339.
857 Karolchik D, Barber GP, Casper J, Clawson H, Cline MS, Diekhans M, Dreszer TR,
858 Fujita PA, Guruvadoo L, Haeussler M et al. 2014. The UCSC Genome Browser
859 database: 2014 update. *Nucleic acids research* **42**: D764-770.
860 Katainen R, Dave K, Pitkanen E, Palin K, Kivioja T, Valimaki N, Gylfe AE,
861 Ristolainen H, Hanninen UA, Cajuso T et al. 2015. CTCF/cohesin-binding sites
862 are frequently mutated in cancer. *Nature genetics* **47**: 818-821.
863 Kel AE, Gossling E, Reuter I, Cheremushkin E, Kel-Margoulis OV, Wingender E. 2003.
864 MATCH: A tool for searching transcription factor binding sites in DNA
865 sequences. *Nucleic acids research* **31**: 3576-3579.
866 Kheradpour P, Kellis M. 2014. Systematic discovery and characterization of regulatory
867 motifs in ENCODE TF binding experiments. *Nucleic acids research* **42**:

868 2976-2987.
869 Kulakovskiy IV, Vorontsov IE, Yevshin IS, Sharipov RN, Fedorova AD, Rumynskiy EI,
870 Medvedeva YA, Magana-Mora A, Bajic VB, Papatsenko DA et al. 2018.
871 HOCOMOCO: towards a complete collection of transcription factor binding
872 models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic acids*
873 *research* **46**: D252-D259.
874 Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler
875 transform. *Bioinformatics (Oxford, England)* **25**: 1754-1760.
876 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G,
877 Durbin R, Genome Project Data Processing S. 2009. The Sequence
878 Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)* **25**:
879 2078-2079.
880 McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM,
881 Bejerano G. 2010. GREAT improves functional interpretation of cis-regulatory
882 regions. *Nature biotechnology* **28**: 495-501.
883 Mumbach MR, Satpathy AT, Boyle EA, Dai C, Gowen BG, Cho SW, Nguyen ML,
884 Rubin AJ, Granja JM, Kazane KR et al. 2017. Enhancer connectome in primary
885 human cells identifies target genes of disease-associated DNA elements. *Nature*
886 *genetics* **49**: 1602-1612.
887 Neph S, Kuehn MS, Reynolds AP, Haugen E, Thurman RE, Johnson AK, Rynes E,
888 Maurano MT, Vierstra J, Thomas S et al. 2012. BEDOPS: high-performance
889 genomic feature operations. *Bioinformatics (Oxford, England)* **28**: 1919-1920.
890 Newburger DE, Bulyk ML. 2009. UniPROBE: an online database of protein binding
891 microarray data on protein-DNA interactions. *Nucleic acids research* **37**:
892 D77-82.
893 Osato N. 2018. Characteristics of functional enrichment and gene expression level of
894 human putative transcriptional target genes. *BMC Genomics* **19**: 957.
895 Phanstiel DH, Van Bortle K, Spacek D, Hess GT, Shamim MS, Machol I, Love MI,
896 Aiden EL, Bassik MC, Snyder MP. 2017. Static and Dynamic DNA Loops form
897 AP-1-Bound Activation Hubs during Macrophage Development. *Molecular cell*
898 **67**: 1037-1048.e1036.
899 Portales-Casamar E, Thongjuea S, Kwon AT, Arenillas D, Zhao X, Valen E, Yusuf D,
900 Lenhard B, Wasserman WW, Sandelin A. 2010. JASPAR 2010: the greatly

901 expanded open-access database of transcription factor binding profiles. *Nucleic*
902 *acids research* **38**: D105-110.
903 Rebollo R, Romanish MT, Mager DL. 2012. Transposable elements: an abundant and
904 natural source of regulatory sequences for host genes. *Annu Rev Genet* **46**:
905 21-42.
906 Schreiber J, Libbrecht M, Bilmes J, Noble W. 2017. Nucleotide sequence and DNaseI
907 sensitivity are predictive of 3D chromatin architecture. *bioRxiv*.
908 Tabuchi TM, Deplancke B, Osato N, Zhu LJ, Barrasa MI, Harrison MM, Horvitz HR,
909 Walhout AJ, Hagstrom KA. 2011. Chromosome-biased binding and gene
910 regulation by the Caenorhabditis elegans DRM complex. *PLoS Genet* **7**:
911 e1002074.
912 van Helden J. 2003. Regulatory sequence analysis tools. *Nucleic acids research* **31**:
913 3593-3596.
914 Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF,
915 Schroth GP, Burge CB. 2008. Alternative isoform regulation in human tissue
916 transcriptomes. *Nature* **456**: 470-476.
917 Wang J, Vicente-Garcia C, Seruggia D, Molto E, Fernandez-Minan A, Neto A, Lee E,
918 Gomez-Skarmeta JL, Montoliu L, Lunyak VV et al. 2015. MIR retrotransposon
919 sequences provide insulators to the human genome. *Proc Natl Acad Sci U S A*
920 **112**: E4428-4437.
921 Wang L, Wang S, Li W. 2012. RSeQC: quality control of RNA-seq experiments.
922 *Bioinformatics (Oxford, England)* **28**: 2184-2185.
923 Weintraub AS, Li CH, Zamudio AV, Sigova AA, Hannett NM, Day DS, Abraham BJ,
924 Cohen MA, Nabet B, Buckley DL et al. 2017. YY1 Is a Structural Regulator of
925 Enhancer-Promoter Loops. *Cell* **171**: 1573-1588 e1528.
926 Wingender E, Dietze P, Karas H, Knuppel R. 1996. TRANSFAC: a database on
927 transcription factors and their DNA binding sites. *Nucleic acids research* **24**:
928 238-241.
929 Xie X, Mikkelsen TS, Gnirke A, Lindblad-Toh K, Kellis M, Lander ES. 2007.
930 Systematic discovery of regulatory motifs in conserved regions of the human
931 genome, including thousands of CTCF insulator sites. *Proc Natl Acad Sci U S A*
932 **104**: 7145-7150.
933 Xie Z, Hu S, Blackshaw S, Zhu H, Qian J. 2010. hPDI: a database of experimental

934 human protein-DNA interactions. *Bioinformatics (Oxford, England)* **26**:
935 287-289.
936 Zhang H, Li F, Jia Y, Xu B, Zhang Y, Li X, Zhang Z. 2017. Characteristic arrangement
937 of nucleosomes is predictive of chromatin interactions at kilobase resolution.
938 *Nucleic acids research* **45**: 12739-12751.
939 Zhao Y, Stormo GD. 2011. Quantitative analysis demonstrates most transcription
940 factors require only simple models of specificity. *Nature biotechnology* **29**:
941 480-483.
942