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. Other DNA binding proteins such as YY1, ZNF143, and SMARCA4 are also reported to be associated with chromatin interactions. It is still unclear what proteins are associated with chromatin interactions and insulator functions.

To find DNA binding motif sequences of transcription factors (TF), such as CTCF, 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, T cells, HMEC and NPC, 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 closed chromatin regions.

Total 96 biased orientations of DNA motifs (64 FR and 52 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 monocytes, T cells, HMEC and NPC in common. For 44 (69%) FR and 28 (54%) RF of 64 FR and 52 RF orientations of DNA motif sequences in $CD4^+$ T cells, EPI predicted from EPA that was 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, 43 (72%) FR and 40 (80%) RF of 64 FR and 52 RF orientations of DNA motifs found in the four cell types in common showed significantly reduced correlation in expression level of nearby genes separated by the motif sites, implying that the DNA motifs were associated with insulator functions.

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

55 anchors – forward-reverse (FR) orientation is frequently observed (Rao et al. 2014; de
56 Wit et al. 2015; Guo et al. 2015). Other DNA binding proteins such as ZNF143, YY1,
57 and SMARCA4 (BRG1) are found to be associated with chromatin interactions and EPI
58 (Bailey et al. 2015; Barutcu et al. 2016; Weintraub et al. 2017). CTCF, cohesin, ZNF143,
59 YY1 and SMARCA4 have other biological functions as well as chromatin interactions
60 and EPI. The DNA binding motif sequences of the transcription factors (TF) are found
61 in open chromatin regions near transcriptional start sites (TSS) as well as chromatin
62 interaction anchors.

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

70 DNA binding motif sequence of YY1 does not seem to be enriched at both
71 chromatin interaction anchors (Z-score < 2), whereas DNA binding motif sequence of
72 ZNF143 is significantly enriched (Z-score > 7; Bailey et al. 2015 Figure 2a). In the
73 analysis of YY1, to identify a protein factor that might contribute to EPI, (Ji et al. 2015)
74 performed chromatin immune precipitation with mass spectrometry (ChIP-MS), using
75 antibodies directed toward histones with modifications characteristic of enhancer and
76 promoter chromatin (H3K27ac and H3K4me3, respectively). Of 26 transcription factors

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

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

88 DNA binding proteins involved in chromatin interactions are supposed to affect
89 the transcription of genes in the loops formed by chromatin interactions. In my previous
90 analysis, the expression level of human putative transcriptional target genes was
91 affected, according to the criteria of enhancer-promoter association (EPA) (Fig. 1B;
92 (Osato 2018)). EPA was shortened at the genomic locations of FR orientation of CTCF
93 binding sites, and transcriptional target genes of each TF bound in enhancers and
94 promoters were predicted based on the EPA. The EPA affected the expression levels of
95 putative transcriptional target genes, compared with the expression levels of
96 transcriptional target genes predicted from promoters (or closed chromatin regions) (Fig.
97 2A). The expression levels tend to be increased in monocytes and $CD4^+$ T cells,
98 implying that enhancers activated the transcription of genes. The expression levels tend

99 to be decreased in ES and iPS cells, implying that enhancers repressed the transcription
100 of genes. These analyses suggested that enhancers affected the transcription of genes
101 significantly, when EPI were predicted properly. Other DNA binding proteins involved
102 in chromatin interactions, as well as CTCF, may locate at chromatin interaction anchors
103 with a pair of biased orientation of DNA binding motif sequences, affecting the
104 expression level of putative transcriptional target genes in the loops formed by the
105 chromatin interactions. As experimental issues of the analyses of chromatin interactions,
106 chromatin interaction data are changed, according to experimental techniques, depth of
107 DNA sequencing, and even replication sets of the same cell type. Chromatin interaction
108 data may not be saturated enough to cover all chromatin interactions. Supposing these
109 characters of DNA binding proteins associated with chromatin interactions and avoiding
110 experimental issues of the analyses of chromatin interactions, here I searched for DNA
111 motif sequences of TF, affecting EPI and the expression level of putative transcriptional
112 target genes in monocytes, T cells, HMEC and NPC without using chromatin interaction
113 data. Then, putative EPI were compared with chromatin interaction and gene expression
114 data.
115

116 ## Results

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

118 Transcription factor binding sites (TFBS) were predicted using open chromatin
119 regions and DNA motif sequences of transcription factors (TF) collected from various
120 databases and journal papers (see Methods). Transcriptional target genes were predicted

121 using TFBS in promoters and enhancers. Enhancer-promoter association (EPA) was
122 shortened at the genomic locations of DNA binding motif sequence of a TF acting as
123 insulator such as CTCF and cohesin (RAD21 and SMC3) (Fig. 1B). To find DNA motif
124 sequences of TF acting as insulator, other than CTCF and cohesin, affecting the
125 expression level of genes, EPA was shortened at the genomic locations of DNA motif
126 sequence of a TF, and transcriptional target genes of each TF bound in (i) distal open
127 chromatin regions (enhancers) and (ii) closed chromatin regions in promoters were
128 predicted based on the EPA, in order to estimate enhancer activity of each TF. The ratio
129 of expression level of putative transcriptional target genes of each TF between (i) and
130 (ii) was calculated. The distribution of the ratios of all TF was compared among
131 forward-reverse (FR), reverse-forward (RF), and any orientation (i.e. without
132 considering orientation) of a DNA motif sequence shortening EPA using Mann-Whitney
133 test, two-sided (p-value $< 10^{-7}$), and the number of TF with significant difference of the
134 distributions among FR, RF and Any was counted (Fig. 2A). Several hundreds of FR
135 and RF orientations of DNA motifs of TF were found in monocytes, T cells, HMEC and
136 NPC. The number of the TF seemed to be changed, according to the number of
137 DNase-seq reads, since the number of the reads of monocyte is larger than other three
138 cell types (i.e. more DNA motif sites would be predicted to some extent, if DNase-seq
139 reads increase). Total 96 of biased (64 FR and 52 RF) orientations of DNA binding
140 motif sequences of TF were found in the four cell types in common, whereas any
141 orientation of DNA binding motif sequence was not found (Fig. 2B; Table 1;
142 Supplemental material). FR orientation of DNA motif sequences included CTCF,

143 cohesin (RAD21 and SMC3), YY1 and ZNF143 which are associated with chromatin
144 interactions and EPI.

145 Without considering the difference of orientation of DNA binding motif
146 sequences, 175 of biased orientations of DNA binding motif sequences of TF were
147 found in the four cell types in common. As a reason for the increase of the number (175)
148 from 96, a TF or an isoform of the same TF may bind to a different DNA binding motif
149 sequence according to cell types and/or in the same cell type. About 50% or more of
150 alternative splicing isoforms are differently expressed among tissues, indicating that
151 most alternative splicing is subject to tissue-specific regulation (Das et al. 2007; Wang
152 et al. 2008; Chen and Manley 2009). The same TF has several DNA binding motif
153 sequences and in some cases one of the motif sequences is almost the same as the
154 reverse complement sequence of another motif sequence of the same TF. I previously
155 found that a complex of TF would bind to a slightly different DNA binding motif
156 sequence from the combination of DNA binding motif sequences of TF composing the
157 complex in *C. elegans* (Tabuchi et al. 2011). The difference of binding specificity of
158 pairs of transcription factors was reported (Jolma et al. 2015). Cofactors also contribute
159 to alter binding specificity of transcription factors (Ansari and Peterson-Kaufman 2011;
160 Slattery et al. 2011; Merabet and Mann 2016). These increase the variety and number of
161 DNA binding motif sequences of a TF.

162

163 **Comparison with chromatin interaction data**

164 To examine whether the biased orientations of DNA motif sequences is associated

165 with chromatin interactions, enhancer-promoter interactions (EPI) predicted based on
166 enhancer-promoter associations (EPA) were compared with chromatin interaction data.
167 HiChIP chromatin interaction data for $CD4^+$ T cells were used (Mumbach et al. 2017).
168 EPI predicted based on EPA were compared with three replications (B2T1, B2T2, and
169 B3T1) of HiChIP chromatin interaction data in $CD4^+$ T cells respectively. The
170 resolutions of HiChIP chromatin interaction data and EPI were adjusted to 5kb. EPI
171 were predicted based on three types of EPA: (i) EPA shortened at the genomic locations
172 of FR or RF orientation of DNA motif sequence of a TF acting as insulator sites such as
173 CTCF, (ii) EPA shortened at the genomic locations of DNA motif sequence of a TF
174 without considering their orientation, and (iii) EPA without being shortened by the
175 genomic locations of a DNA motif sequence. Total 201 biased orientations [133 (49%)
176 FR and 90 (43%) RF of 273 FR and 211 RF] of DNA motif sequences in T cells, which
177 included CTCF, cohesin (RAD21 and SMC3), ZNF143, and YY1 in three replications
178 (B2T1, B2T2, and B3T1) showed a significantly higher ratio of EPI overlapped with
179 HiChIP chromatin interactions (more than 1,000 counts for each interaction) than the
180 other types of EPAs in T cells (Table 2). When comparing EPI predicted based on only
181 EPA (i) and (iii) with the chromatin interactions, total 390 biased orientations [261
182 (96%) FR and 199 (94%) RF of 273 FR and 211 RF] of DNA motif sequences in T cells
183 showed a significantly higher ratio of EPI predicted based on EPA (i) overlapped with
184 the chromatin interactions than EPI predicted based on EPA (iii) (Table 2; Supplemental
185 material). The difference between EPI predicted based on EPA (i) and (ii) seemed to be
186 difficult to distinguish using the chromatin interaction data and the statistical test in

187 some cases. However, as for the difference between EPI predicted based on EPA (i) and
188 (iii), a larger number of biased orientations of DNA motif sequences was found to be
189 correlated with chromatin interaction data. Chromatin interaction data were obtained
190 from different samples from DNase-seq, open chromatin regions, so individual
191 differences may exist. (Mumbach et al. 2017) suggested that individual differences of
192 chromatin interactions were larger than those of open chromatin regions. Most of biased
193 orientations of DNA motif sequences (95%) were found to be correlated with chromatin
194 interactions, when comparing EPI predicted based on EPA (i) and (iii) with HiChIP
195 chromatin interactions.

196 Moreover, to examine the enhancer activity of EPI, the distribution of expression
197 level of putative target genes of EPI was compared between EPI overlapped with
198 HiChIP chromatin interactions and EPI not overlapped with them. Though the target
199 genes of EPI were selected from top 4,000 transcripts (genes) in terms of the expression
200 level of all transcripts (genes) excluding transcripts not expressed in T cells, target
201 genes of EPI overlapped with chromatin interactions showed a significantly higher
202 expression level than EPI not overlapped with them, suggesting that EPI overlapped
203 with chromatin interactions activated the expression of target genes in T cells. All FR
204 and RF orientations of DNA motifs showed a significantly higher expression level of
205 putative target genes of EPI overlapped with chromatin interactions than EPI not
206 overlapped. When a biased orientation of DNA motif showed a significantly higher
207 expression level of putative target genes of EPI overlapped with chromatin interactions
208 than EPI not overlapped, ‘1’ was marked with in the tables of comparison between EPI

209 and HiChIP chromatin interactions in Supplemental material. DNA motifs showing a
210 significantly lower expression level or not showing significant difference of expression
211 level were not observed in this analysis. HiChIP data were produced using H3K27ac
212 antibody, so chromatin interactions acting as repressor would not be identified.
213
214 **Correlation of expression level of gene pairs separated by a DNA motif sequence**
215 To examine the effect of biased orientations of DNA motif sequences on gene
216 expression, I compared the expression level of protein-coding genes the most closely
217 located to upstream and downstream of biased orientations of DNA motif sequences in
218 human genome. Closely located genes are expected to show correlation of gene
219 expression among tissues (Xie et al. 2007). When a biased orientation of DNA motif
220 sequence acts as insulator (e.g. CTCF), the correlation of gene expression level would
221 be reduced. Among 96 biased (64 FR and 52 RF) orientations of DNA motifs found in
222 common in monocytes, T cells, HMEC, and NPC, top 20 were selected in ascending
223 order of the median correlation coefficient of expression level of the closest gene pairs
224 of the DNA motif that was observed in more than 50 genomic locations. The number of
225 genomic locations was relatively small, when ubiquitously expressed genes among 53
226 tissues and DNA motif sites near the genes were eliminated with coefficient of variance
227 < 90 (see Methods). Pairs of genes near all 20 motifs showed a distribution of
228 significantly lower correlation of gene expression level than all pairs of genes, and
229 included known DNA binding motif sequences of TF associated with chromatin
230 interactions such as CTCF, RAD21, and SMC3 (Fig 3) (Mann-Whitney test, p-value <

231 0.01). Of 96 biased orientations of DNA motifs found in common in the four cell types,
232 43 (72%) FR and 40 (80%) RF of 60 FR and 50 RF with variable expression data of
233 nearby genes among 53 tissues showed a distribution of significantly lower correlation
234 of gene expression of the closest genes of the DNA motifs than the correlation of all
235 pairs of neighbor genes (Mann-Whitney test, p-value < 0.05) (Table 3). The DNA
236 motifs included known DNA binding motif sequences of TF associated with chromatin
237 interactions such as CTCF, RAD21, SMC3, YY1 and ZNF143. Though the number of
238 DNA motif sites was reduced, compared with all DNA motif sites of a TF, DNA motif
239 sites examined for biased orientation in EPA showed a distribution of significantly
240 lower correlation of expression level of gene pairs separated by the sites (33 (60%) FR
241 and 36 (77%) RF of 55 FR and 47 RF orientations of DNA motifs of TF with variable
242 expression data of nearby genes) (Supplemental material). Some DNA motif sites of
243 biased orientations of DNA motifs were found from genomic regions not including
244 known DNA motifs of TF associated with chromatin interactions such as CTCF,
245 RAD21, SMC3, YY1 and ZNF143, and located in significantly lower correlation of
246 expression level of gene pairs. Significantly lower correlation of gene expression tend to
247 be observed in DNA motifs with relatively a large number of genomic sites (Median
248 number of sites of statistically significant DNA motifs is 656 and 479 for FR and RF
249 respectively, and not significant DNA motifs is 122 and 115 for FR and RF respectively.
250 p-value = 0 and $< 10^{-10}$ for FR and RF respectively, Mann-Whitney test). Many DNA
251 motifs showed significantly lower correlation of gene expression, so I wondered there
252 may be a possibility that when we take a small number of gene pairs randomly (about

253 hundreds as the same as the number of DNA motif sites), the correlation of gene pairs
254 might be significantly low occasionally. However, the possibility that happen was low
255 and less than expectation from *p*-value threshold by computer simulation.
256
257 **Co-location of biased orientations of DNA motif sequences**
258 To examine the association of biased orientations of DNA binding motif
259 sequences of TF, co-location of the DNA binding motif sequences within 200bp to each
260 other was analyzed in monocytes, T cells, HMEC, and NPC. The number of genes with
261 the same pairs of DNA binding motif sequences in upstream and downstream of the
262 genes in EPA was counted, and the pairs of DNA binding motif sequences found in
263 more than 100 genes were listed (Table 4; Supplemental material). Majority of pairs of
264 DNA motif sites were overlapped with more than 1bp, but the same tendency was found
265 in the analysis of DNA motif sites of all TF (Supplemental material). As already known,
266 CTCF was found with cohesin RAD21 and SMC3 (Table 4). Top 30 pairs of FR and RF
267 orientations of DNA motifs overlapped or co-localized were shown (Table 4). Total
268 number of pairs of DNA motifs overlapped or co-localized was 433 (FR) and 708 (RF)
269 in monocytes, consisting of 178 (39% of all FR) and 213 (42% of all RF) unique DNA
270 motifs. Only eight overlapping or co-location of DNA motifs of TF were found in the
271 four cell types in common, including pairs of CTCF, RAD21 and SMC3. This implied
272 that most of overlapping and co-location of DNA motifs of TF were cell-type-specific
273 (Supplemental material). Pairs of biased orientations of DNA binding motifs of TF tend
274 to be overlapped, and relatively a small number of motifs were co-localized in upstream

275 and downstream of genes. Overlapping DNA motifs act as DNA binding sites of several
276 TF simultaneously or competitively (Ackerman et al. 1991; Yoon and Chikaraishi 1992;
277 Ansari and Peterson-Kaufman 2011; Slattery et al. 2011; Chatterjee et al. 2012; He et al.
278 2015; Kin et al. 2016; Merabet and Mann 2016). CTCF and cohesin have also been
279 reported to act cooperatively with other TF and protein such as Klf4 and Brd2 (Wei et al.
280 2013; Hsu et al. 2017). Klf4 recruits cohesin to the Oct4 enhancer region. BRD2
281 supports boundary (insulator) activity of CTCF. The analysis of 457 ChIP-seq data sets
282 on 119 human TFs showed that secondary motifs were detected in addition to the
283 canonical motifs of the TFs, indicating tethered binding and co-binding between
284 multiple TFs (Wang et al. 2012a). The authors observed significant position and
285 orientation preferences between many co-binding TFs. Overlapping DNA motif sites of
286 TF found in this study may not be artifact, but biologically functional. Though repeat
287 sequences in human genome were masked, more than a hundred of pairs of DNA motif
288 sites overlapped at more than a hundred or a thousand upstream and downstream
289 regions of genes.
290

291 **Discussion**

292 To find DNA motif sequences of transcription factors (TF) affecting the
293 expression level of human putative transcriptional target genes, DNA motif sequences
294 were searched from open chromatin regions of monocytes, T cells, HMEC and NPC.
295 Total 96 biased [64 forward-reverse (FR) and 52 reverse-forward (RF)] orientations of
296 DNA motif sequences of TF were found in the four cell types in common, whereas any

297 orientation (i.e. without considering orientation) of DNA motif sequence of TF was not
298 found to affect the expression level of putative transcriptional target genes, suggesting
299 that enhancer-promoter association (EPA) shortened at the genomic locations of FR or
300 RF orientation of DNA motif sequence of a TF is effective for the prediction of
301 enhancer-promoter interactions (EPI).

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

319 large enough.

320 As HiChIP chromatin interaction data were available in $CD4^+$ T cells, biased
321 orientations of DNA motif sequences of TF were examined in T cell. The resolutions of
322 chromatin interactions and EPI were adjusted to 5kb by fragmentation of genome
323 sequences. The number of HiChIP chromatin interactions was 19,926,360 at 5kb
324 resolution, 666,149 at 5kb resolution with chromatin interactions (more than 1,000
325 counts for each interaction), and 78,209 at 5kb resolution with chromatin interactions
326 (more than 6,000 counts for each interaction). As expected, the number of EPI
327 overlapped with chromatin interactions was increased, and 133 FR and 90 RF
328 orientations of DNA motif sequences of TF showed a statistical significance in EPI
329 predicted based on EPA shortened at the genomic locations of the DNA motif sequence,
330 compared with the other types of EPAs or EPA not shortened. HiChIP analysis used
331 H3K27ac antibody to capture chromatin interactions and identified twenty millions of
332 chromatin interactions, but the number of peaks of ChIP-seq data of H3K27ac is usually
333 less than a hundred thousand. There seems to be much difference of the number of
334 H3K27ac sites between HiChIP and ChIP-seq, so ChIP-seq of H3K27ac may not cover
335 all H3K27ac sites.

336 When forming a homodimer or heterodimer with another TF, TF may bind to
337 genome DNA with a specific orientation of their DNA binding sequences (Fig. 4). From
338 the analysis of biased orientations of DNA motif sequences of TF, TF forming
339 heterodimer would also be found. If the DNA binding motif sequence of only the mate
340 to a pair of TF was found in EPA, EPA was shortened at one side, which is the genomic

341 location of the DNA binding motif sequence of the mate to the pair, and transcriptional
342 target genes were predicted using the EPA shortened at the side. In this analysis, the
343 mate to both heterodimer and homodimer of TF can be used to examine the effect on the
344 expression level of transcriptional target genes predicted based on the EPA shortened at
345 one side. Biased orientations of DNA motif sequences may also be found in
346 forward-forward or reverse-reverse orientation.

347 In general, DNA motif sequences of TF were searched from genome sequences
348 using position weight matrix or position frequency matrix and allowing mismatch of
349 DNA sequences under a certain threshold. However, the change of the threshold
350 affected the result of analyses and it seems to be difficult to find the best parameter
351 setting to reduce false-positive and false-negative predictions of DNA biding sites of TF.
352 Therefore, firstly, to obtain a robust result in this study, position weight matrixes of
353 DNA motifs were converted into consensus DNA sequences including degenerated sites,
354 which consist of A, T, G, C, R, Y, M, K, S, W, B, H, V, D, and N letters. This does not
355 need parameter setting in the search for DNA motif sequences, and is the same as *k*-mer
356 analysis of DNA motif sequences of TF. Thought the number of genomic regions
357 matched with a DNA motif sequence was decreased, about 200 FR and RF orientations
358 of DNA motif sequences of TF were found, and the number of any orientation of DNA
359 motif sequences was quite small, suggesting that false-positive prediction of DNA
360 biding sites of TF was reduced. However, the length of the consensus DNA sequences
361 of some TF is relatively long. When the core DNA binding sequence of a TF is short (5
362 or 6 bp), it may be strict to find perfect matches of the long consensus DNA sequence in

363 genome sequences, not allowing a mismatch of a DNA sequence.

364 Secondly, protein interaction quantitation (PIQ) tool was used to find DNA motif
365 sites of TF (Sherwood et al. 2014). PIQ predicts DNA motifs sites of TF by reducing
366 noise of DNase-seq data, and the comparison of the result of prediction with ChIP-seq
367 data revealed that PIQ showed a better performance than other tools and digital genome
368 footprinting (DNase-DGF). The results of prediction of DNA motif sites of TF using
369 PIQ in this study showed an increase of biased orientations of DNA motif sequences of
370 TF, reducing the number of any orientation of DNA motif sequences of TF. The number
371 of DNA motif sites of TF was expected to increase, compared with the above method to
372 use consensus DNA sequences of TF. However, the number of DNA motif sites was
373 almost the same or slightly decreased, compared with the result of consensus DNA
374 sequences, suggesting that PIQ increases true positive prediction of DNA motif sites,
375 reducing false positive prediction.

376 To estimate enhancer activities of DNA motifs of TF, the expression level of
377 putative transcriptional target genes of a TF was compared between DNA motifs of TF
378 in distal open chromatin regions (enhancers) and closed chromatin regions in promoters.
379 Previously, the expression level was compared between distal open chromatin regions
380 and open chromatin regions in promoters. However, it would be easy to understand the
381 comparison of DNA motif sites of a TF between distal open chromatin regions and
382 closed chromatin regions to estimate the enhancer activity of TF. The number of DNA
383 motif sites in closed chromatin regions was large, so the number in closed chromatin
384 regions in promoter was used.

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

389 The analyses in this study revealed novel characters of DNA binding motif
390 sequences of TF involved in chromatin interactions, insulator function and potentially
391 forming a homodimer, heterodimer or complex with other TF, affecting the
392 transcriptional regulation of genes.

393

394 ## Methods

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

396 To identify transcription factor binding sites (TFBS) from open and closed
397 chromatin regions, TRANSFAC (2019.2), JASPAR (2018), UniPROBE (2018),
398 high-throughput SELEX, transcription factor binding sequences of ENCODE ChIP-seq
399 data, and HOCOMOCO version 9 and 11 were used to predict insulator sites
400 (Wingender et al. 1996; Newburger and Bulyk 2009; Portales-Casamar et al. 2010; Xie
401 et al. 2010; Zhao and Stormo 2011; Jolma et al. 2013; Kheradpour and Kellis 2014;
402 Kulakovskiy et al. 2018). TRANSFAC (2019.2) was used to analyze enhancer-promoter
403 interactions, since these data were sufficient to identify biased orientations of DNA
404 motif sequences of TF with less computational time, reducing the number of any
405 orientation of DNA motif sequences of TF. To reduce false positive prediction of DNA
406 motif sites of TF in open chromatin regions, protein interaction quantitation (PIQ) tool

407 was used with 12,249 position frequency matrices converted from DNA motifs of
408 vertebrate TF in the above databases and DNase-seq data from GEO database
409 (GSM1024791 $CD14^+$ monocytes; GSM665812 $CD4^+$ T cell; GSM736634 HMEC;
410 GSM878615 NPC) (Sherwood et al. 2014). To predict DNA motif sites of TF from
411 closed chromatin regions, another method was used. Position weight matrices of
412 vertebrate transcription factor binding sequences were converted into DNA consensus
413 sequences considering degenerated sites using convert matrix in RSAT website (van
414 Helden 2003). Transcription factor binding sequences of vertebrate TF were used for
415 further analyses. Transcription factor binding sequences were searched from each
416 narrow peak of DNase-seq data in repeat-masked hg19 genome sequences using Match
417 tool in TRANSFAC with similarity score cutoff of 1 after transforming the DNA
418 consensus sequences of TF into position frequency matrixes for Match search (Kel et al.
419 2003). Repeat-masked hg19 genome sequences were downloaded from UCSC genome
420 browser (http://genome.ucsc.edu/,
421 http://hgdownload.soe.ucsc.edu/goldenPath/hg19/bigZips/hg19.fa.masked.gz). TF
422 corresponding to transcription factor binding sequences were searched computationally
423 and manually by comparing their names and gene symbols of HGNC (HUGO Gene
424 Nomenclature Committee) -approved gene nomenclature and 31,848 UCSC known
425 canonical transcripts
426 (http://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/knownCanonical.txt.gz), as
427 transcription factor binding sequences were not linked to transcript IDs such as UCSC,
428 RefSeq, and Ensembl transcripts.

429 Target genes of a TF were assigned when its TFBS was found in DNase-seq
430 narrow peaks in promoter or extended regions for enhancer-promoter association of
431 genes (EPA). Promoter and extended regions were defined as follows: promoter regions
432 were those that were within distance of ±5 kb from transcriptional start sites (TSS).
433 Promoter and extended regions were defined as per the following association rule,
434 which is the same as that defined in Figure 3A of a previous study (McLean et al. 2010):
435 the single nearest gene association rule, which extends the regulatory domain to the
436 midpoint between the TSS of the gene and that of the nearest gene upstream and
437 downstream without the limitation of extension length. Extended regions for EPA were
438 shortened at the genomic locations of DNA binding sites of a TF that was the closest to
439 a transcriptional start site, and transcriptional target genes were predicted from the
440 shortened enhancer regions using TFBS. Furthermore, promoter and extended regions
441 for EPA were shortened at the genomic locations of forward–reverse (FR) orientation of
442 DNA binding sites of a TF. When forward or reverse orientation of DNA binding sites
443 were continuously located in genome sequences several times, the most internal (i.e.
444 closest to a transcript) forward–reverse orientation of DNA binding sites were selected.
445 The genomic positions of genes were identified using 'knownGene.txt.gz' file in UCSC
446 bioinformatics sites (Karolchik et al. 2014). The file 'knownCanonical.txt.gz' was also
447 utilized for choosing representative transcripts among various alternate forms for
448 assigning promoter and extended regions for EPA. From the list of transcription factor
449 binding sequences and transcriptional target genes, redundant transcription factor
450 binding sequences were removed by comparing the target genes of a transcription factor

451 binding sequence and its corresponding TF; if identical, one of the transcription factor
452 binding sequences was used. When the number of transcriptional target genes predicted
453 from a transcription factor binding sequence was less than five, the transcription factor
454 binding sequence was omitted.

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

459 For gene expression data, RNA-seq reads mapped onto human hg19 genome
460 sequences were obtained from UCSF-UBC human reference epigenome mapping
461 project RNA-seq reads with poly-A of naive $CD4^+$ T cells (GEO: GSM669617).
462 FPKMs of the RNA-seq data were calculated using RSeQC (Wang et al. 2012b). For
463 monocytes, Blueprint RNA-seq FPKM data
464 ('C0010KB1.transcript_quantification.rsem_grape2_crg.GRCh38.20150622.results')
465 were downloaded from Blueprint DCC portal
466 (http://dcc.blueprint-epigenome.eu/#/files). RNA-seq reads with poly-A of human
467 skeletal muscle myoblasts (HSMM) (GEO: GSM958744, Encode: ENCSR000CWN)
468 and H1 derived neural progenitor cell (NPC) (GEO: GSM915326, Encode:
469 ENCFF529SIO) were used. Based on log2-transformed FPKM, UCSC transcripts were
470 arranged in descending order of expression level, and top 4,000 transcripts were
471 selected in each cell type.

472 The expression level of transcriptional target genes predicted based on EPA

473 shortened at the genomic locations of DNA motif sequence of a TF was compared with
474 the expression level of transcriptional target genes predicted from closed chromatin
475 regions in promoters. For each DNA motif sequence shortening EPA, transcriptional
476 target genes were predicted using DNA binding motif sequences of vertebrate TF in
477 TRANSFAC database, and the ratio of expression level of putative transcriptional target
478 genes of each TF was calculated between EPA and closed chromatin regions in
479 promoters. The distribution of the ratios of all TF was compared among forward-reverse
480 (FR), reverse-forward (RF), and any orientation (i.e. without considering orientation) of
481 a DNA motif sequence shortening EPA using Mann-Whitney test, two-sided (p-value <
482 10^{-7}). Other parameters for the analysis were that the number of transcriptional target
483 genes of a TF was more than 50, and the number of TF (not insulator TF) to predict
484 transcriptional target genes was more than 50. The number of DNA motif sites of TF
485 (insulator TF) shortening EPA was more than 100. When a DNA motif sequence of a TF
486 shortening EPA showed a significant difference of expression level of putative
487 transcriptional target genes among FR, RF or any orientation, the DNA motif sequence
488 was listed.
489

490 **Comparison with chromatin interaction data**

491 For comparison of EPA with chromatin interactions (HiChIP) in $CD4^+$ T cells,
492 'GSM2705049_Naive_HiChIP_H3K27ac_B2T1_allValidPairs.txt',
493 'GSM2705050_Naive_HiChIP_H3K27ac_B2T2_allValidPairs.txt', and
494 'GSM2705051_Naive_HiChIP_H3K27ac_B3T1_allValidPairs.txt' files were

495 downloaded from Gene Expression Omnibus (GEO) database (GSM2705049,
496 GSM2705050 and GSM2705051). The resolutions of chromatin interaction data and
497 EPI were adjusted to 5kb before their comparison. Chromatin interactions with more
498 than 6,000 and 1,000 counts for each interaction were used in this study.

499 Enhancer-promoter interactions (EPI) were predicted using three types of EPAs in
500 monocytes: (i) EPA shortened at the genomic locations of FR or RF orientation of DNA
501 motif sequence of a TF, (ii) EPA shortened at the genomic locations of any orientation
502 (i.e. without considering orientation) of DNA motif sequence of a TF, and (iii) EPA
503 without being shortened by a DNA motif sequence. EPI predicted in the three types of
504 EPAs in common were removed. First, EPI predicted based on EPA (i) were compared
505 with chromatin interactions. The resolution of chromatin interaction data used in this
506 study was 1-5kb, so EPI were adjusted to 5kb before their comparison. The number and
507 ratio of EPI overlapped with chromatin interactions were counted. Second, EPI were
508 predicted based on EPA (ii), and EPI predicted based on EPA (i) were removed from the
509 EPI. The number and ratio of EPI overlapped with chromatin interactions were counted.
510 Third, EPI were predicted based on EPA (iii), and EPI predicted based on EPA (i) and
511 (ii) were removed from the EPI. The number and ratio of EPI overlapped with
512 chromatin interactions were counted. The number and ratio of the EPI were compared
513 two times between EPA (i) and (iii), and EPA (i) and (ii) (binomial distribution, p-value
514 < 0.025 for each test, two-sided, 95% confidence interval).

515 Putative target genes for the analysis of EPI were selected from top 4000
516 transcripts in term of the expression level. The expression level of putative target genes

517 of EPI overlapped with HiChIP chromatin interactions was compared with EPI not
518 overlapped with them. For each FR or RF orientation of DNA motif, EPI were predicted
519 based on EPA and the overlap of EPI with chromatin interactions was examined. When
520 a putative transcriptional target gene of a TF in an enhancer was found in both EPI
521 overlapped with a chromatin interaction and EPI not overlapped with, the target gene
522 was removed. The distribution of expression level of putative target genes was
523 compared using Mann-Whitney test, two-sided (p-value < 0.05).
524
525 **Correlation of expression level of gene pairs separated by a DNA motif sequence**
526 To examine the correlation of expression level of genes pairs separated by a DNA
527 motif sequence, gene expression data among 53 tissues were obtained from
528 'GTEx_Analysis_2016-01-15_v7_RNASeQCv1.1.8_gene_median_tpm.gct.gz' file in
529 GTEx database (https://gtexportal.org/home/). The correlation of log2-transformed gene
530 expression level was analyzed based on protein-coding transcripts of RefSeq (RefSeq
531 ID with 'NM_'). The same genomic location of RefSeq transcripts was removed. DNA
532 motifs overlapped with RefSeq transcripts were eliminated. Genes closest to DNA
533 motifs and other genes were analyzed using BEDOPS closest-features (Neph et al.
534 2012). Mann-Whitney test was performed using R.
535
536 **Co-location of biased orientations of DNA motif sequences**
537 Co-location of biased orientations of DNA binding motif sequences of TF was
538 examined. The number of genes with the same pair of DNA motifs of biased

539 orientations of DNA motifs of TF in upstream and downstream of genes in EPA was
540 counted, and the pairs of DNA motifs found in more than 100 genes were listed.
541

542 **Acknowledgements**

543 The supercomputing resource was provided by Human Genome Center of the Institute
544 of Medical Science at the University of Tokyo. Computations were partially performed
545 on the NIG supercomputer at ROIS National Institute of Genetics. Publication charges
546 for this article were funded by JSPS KAKENHI Grant Number 16K00387. This
547 research was partially supported by the Platform Project for Supporting in Drug
548 Discovery and Life Science Research（Platform for Dynamic Approaches to Living
549 System）from Japan Agency for Medical Research and Development (AMED). This
550 research was partially supported by Development of Fundamental Technologies for
551 Diagnosis and Therapy Based upon Epigenome Analysis from Japan Agency for
552 Medical Research and Development (AMED). This work was partially supported by
553 JST CREST Grant Number JPMJCR15G1, Japan.
554

555 **Consent for publication**

556 Images in Figure 2 adapted from , which is distributed under the terms of the Creative
557 Commons Attribution 4.0 International License
558 (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use,
559 distribution, and reproduction in any medium, provided you give appropriate credit to
560 the original author(s) and the source, provide a link to the Creative Commons license,

and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

565 **Figures**

566 Web links to original figures.

567 Figure 1. (A) Figure 1. Enhancers and their features. (Shlyueva et al. 2014)

568 https://static-content.springer.com/esm/art%3A10.1038%2Fnrg3682/MediaObjects/415

569 76_2014_BFnrg3682_MOESM18_ESM.ppt

570 Graphical Abstract (de Wit et al. 2015)

571 https://marlin-prod.literatumonline.com/cms/attachment/7d219941-8398-4373-9ac9-8bf

572 92c5eba1d/fx1.jpg

573

574 (B) Supplementary Figure 2. Computationally-defined regulatory domain (McLean et al.

575 2010)

576 https://static-content.springer.com/esm/art%3A10.1038%2Fnbt.1630/MediaObjects/415

577 87_2010_BFnbt1630_MOESM11_ESM.pdf

578 Figure 4. The Role of CBS Location and Orientation in CTCF-Mediated Genome-wide

579 DNA Looping (Guo et al. 2015)

580 https://marlin-prod.literatumonline.com/cms/attachment/297a06ad-7568-483e-9904-9ae

581 2af081251/gr4.jpg

582

583 **Figure 1.** Chromatin interactions and enhancer-promoter association. (A) Forward–

584 reverse orientation of CTCF-binding sites are frequently found in chromatin interaction

585 anchors. CTCF can block the interaction between enhancers and promoters limiting the

586 activity of enhancers to certain functional domains (Rao et al. 2014; Shlyueva et al.

587 2014; de Wit et al. 2015; Guo et al. 2015). (B) Computationally-defined regulatory
588 domains for enhancer-promoter association (McLean et al. 2010). The *single nearest*
589 *gene* association rule extends the regulatory domain to the midpoint between this gene's
590 TSS and the nearest gene's TSS both upstream and downstream. Enhancer-promoter
591 association was shortened at the genomic locations of forward-reverse orientation of
592 DNA binding motif sequence of a transcription factor (e.g. CTCF in the figure).
593

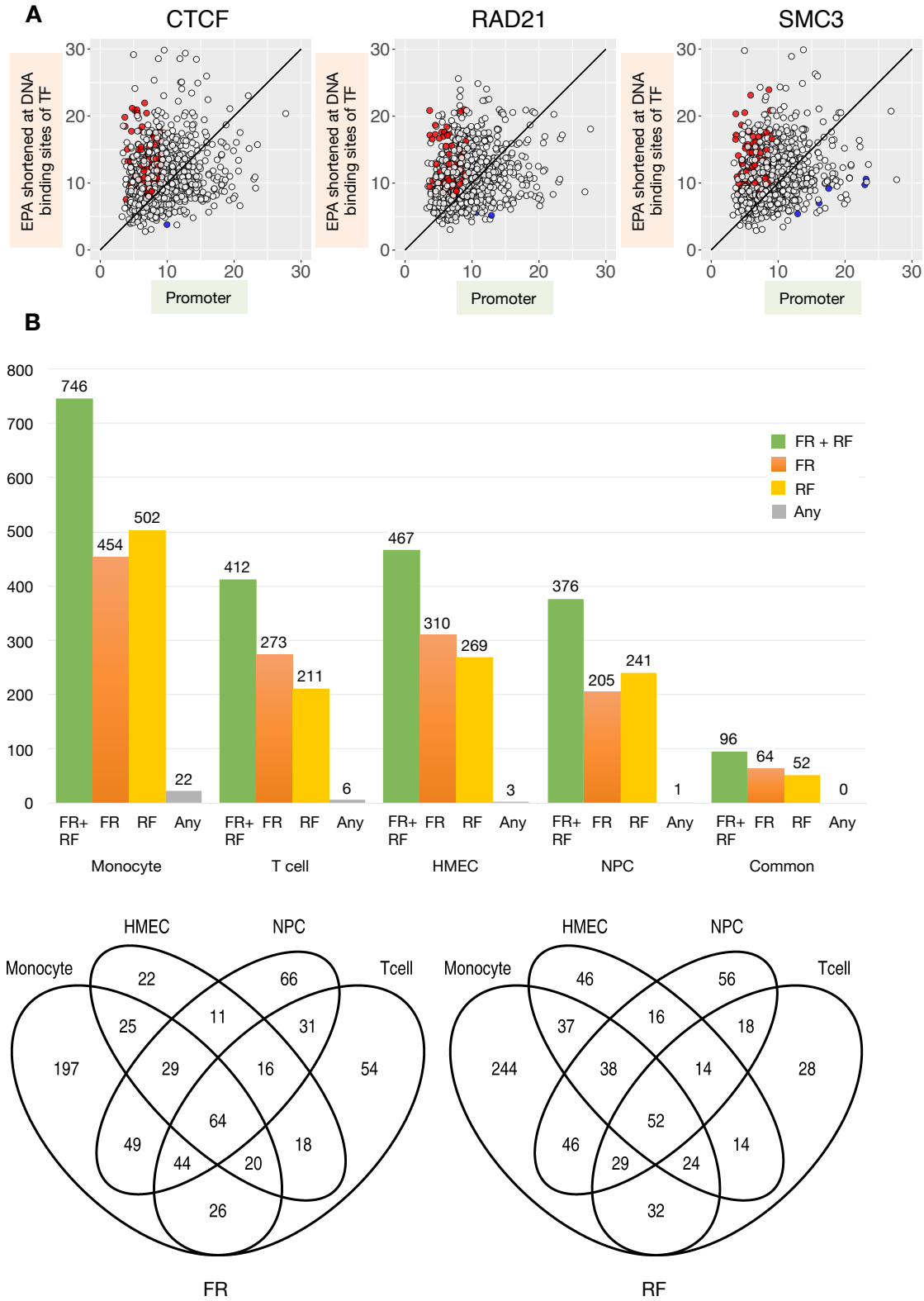
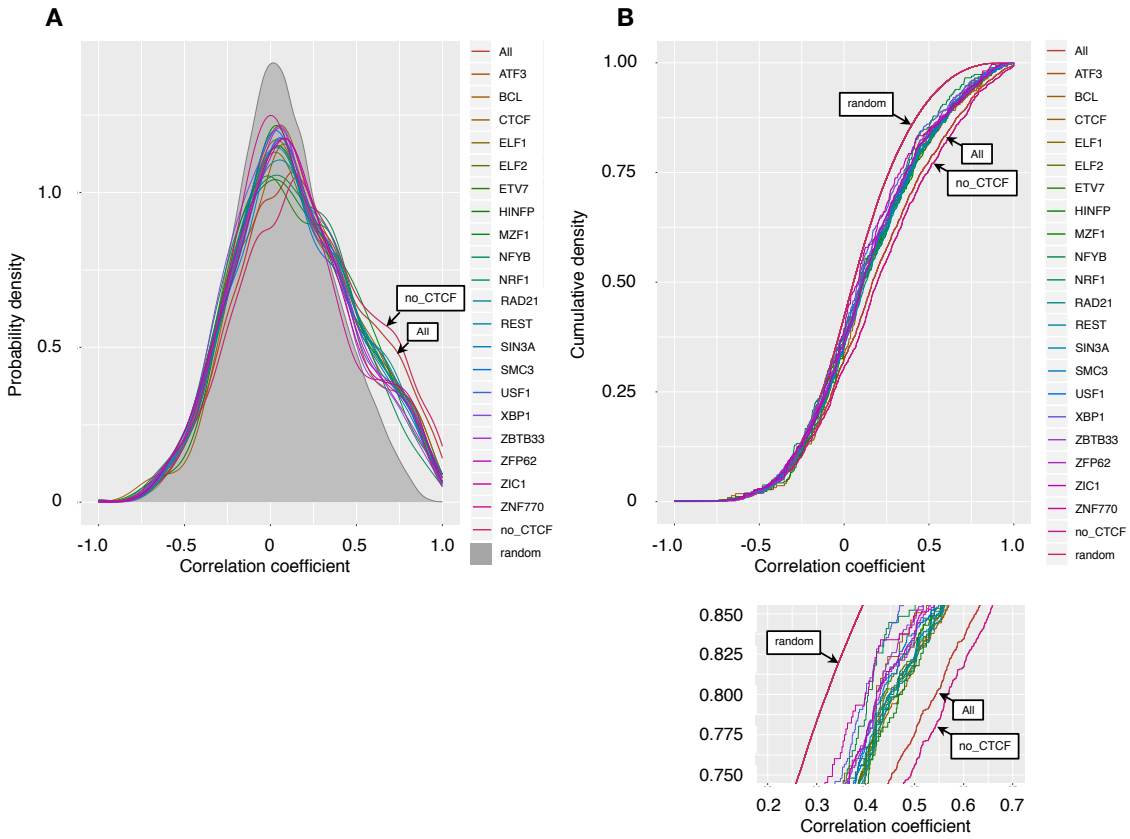


Figure 2. Biased orientations of DNA motif sequences of transcription factors affecting the transcription of genes. (A) Comparison of expression level of putative

597 transcriptional target genes. The median expression levels of target genes of the same
598 transcription factor binding sequences were compared between promoters (or closed
599 chromatin regions) and enhancer-promoter association (EPA) shortened at the genomic
600 locations of forward-reverse orientation of DNA motif sequence of CTCF and cohesin
601 (RAD21 and SMC3) respectively in monocytes. Red and blue dots show statistically
602 significant difference (Mann-Whitney test) of the distribution of expression levels of
603 target genes between promoters and the EPA. Red dots show the median expression
604 level of target genes was higher based on the EPA than promoters, and blue dots show
605 the median expression level of target genes was lower based on the EPA than promoters.
606 The expression level of target genes predicted based on the EPA tend to be higher than
607 promoters, implying that transcription factors acted as activators of target genes (Osato
608 2018). (B) Biased orientations of DNA motif sequences of transcription factors. Total 96
609 of biased orientations of DNA binding motif sequences of transcription factors were
610 found to affect the expression level of putative transcriptional target genes in monocytes,
611 T cells, HMEC and NPC in common, whereas any orientation (i.e. without considering
612 orientation) of DNA binding motif sequence was not found to affect the expression
613 level.
614



615

616 **Figure 3.** Genes separated by predicted DNA binding sites of biased orientations of
617 DNA motifs of TFs are less correlated in gene expression. Top 20 TF are shown
618 according to ascending order of median correlation coefficient. Correlation coefficient
619 between neighboring gene pairs is shown in terms of probability density (A) and
620 cumulative distribution (B). Brown line (All), correlation between all neighboring
621 genes; gray shading in (A) and red line (random) in (B), correlation between randomly
622 chosen gene pairs; Purple line (no_CTCF), correlation between gene pairs not separated
623 by predicted CTCF sites.

624

625

626 Web links to original figures.

627 Figure 5. Schematic representation of chromatin interactions involving gene promoters

628 (Bailey et al. 2015).

629 https://www.nature.com/articles/ncomms7186/figures/5

630

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

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

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

634 2015).

635

636 **Tables**

637 **Table 1.** Biased orientations of DNA binding motif sequences of transcription factors in

638 monocytes, T cells, HMEC and NPC. TF: DNA binding motif sequences of

639 transcription factors.

640

641 Forward-reverse orientation

TF	P-value
AHR	0
ARNT	0
ARNTL	7.96E-08
ATF2	3.09E-14
ATF3	2.22E-15
BCL	0
BHLHE41	5.55E-14
CREB3L1	0
CTCF	0
CXXC1	0
E2F	0
E2F1	0
EGR1	0
EGR3	0
EGR4	0
ELF1	0
ELF2	5.74E-13
ELK3	3.33E-15
ELK4	6.19E-11
ERG	0
ETS	0
ETS1	0
ETV7	2.31E-08
HEY2	0
HIC1	0
HINFP	6.46E-14
INSM1	2.22E-16
KLF4	0
KLF7	9.07E-09
MAX	0

TF	P-value
MYC	0
MYCN	0
MZF1	0
NFY	1.72E-11
NFYB	0
NONO	0
NRF1	1.63E-11
OVOL2	0
PLAG1	0
PLAGL1	0
RAD21	0
REST	0
RFX1	0
SIN3A	0
SIX5	0
SMC3	0
SP2	1.04E-11
TFAP2A	0
TFAP2B	2.22E-16
TFAP2C	0
TP73	0
USF1	0
XBP1	0
YY1	0
YY2	0
ZBTB33	0
ZBTB7A	0
ZFP62	0
ZFX	0
ZIC1	0

TF	P-value
ZNF143	0
ZNF740	0
ZNF770	6.51E-08
ZNF84	0

Reverse-forward orientation

TF	P-value
AP1	0
ARNT	1.26E-11
ATF3	0
BHLHE40	8.88E-16
CREB1	0
CTCF	0
E2F	2.32E-13
E2F3	0
EFNA2	6.61E-11
EGR1	0
EGR2	0
ETS1	0
FLI1	1.05E-13
FLI1:FIGLA	0
FOS	0
GMEB2	0
HIF1A	0
KLF4	0
MITF	0
MYB	0
MYC	0
MYCN	0
MZF1	0
NFKB1	0
NFYA	2.31E-13
NR3C1	0
NRF1	0
OSR2	0
PBX3	0
PLAG1	2.78E-14

TF	P-value
RFX2	0
RFX5	0
RUNX3	0
SP1	0
SP2	0
SP4	0
SPDEF	0
SREBF2	0
STAT	0
TATA	0
TFAP2B	2.00E-15
TFAP2C	0
TFCP2L1	0
USF1	7.55E-10
YBX1	0
YY1	0
YY2	0
ZBTB33	0
ZBTB7B	0
ZNF100	6.22E-14
ZNF143	0
ZNF432	0

642

643

644 **Table 2.** Comparison of enhancer-promoter interactions (EPI) with chromatin
645 interactions in T cells. The upper tables show the number of biased orientations of DNA
646 motifs, where a significantly higher ratio of EPI, which were predicted based on
647 enhancer-promoter association (EPA) (i) overlapped with HiChIP chromatin interaction
648 data than the other types of EPA (ii) and (iii). The middle tables show 133 FR and 90
649 RF orientations of DNA motifs in T cells found in common among B2T1, B2T2 and
650 B3T1 replications in the upper table. The lower table shows that among 64 FR and 52
651 RF biased orientations of DNA motifs found in common in monocytes, T cell, HMEC
652 and NPC, 44 (69%) FR and 28 (54%) RF were matched with the analysis of HiChIP for
653 three types of EPA. TF: DNA binding motif sequence of a transcription factor. Score: –
654 $\log_{10}$ (*p*-value). Score 1: Comparison of EPA shortened at the genomic locations of FR
655 or RF orientation of DNA motif sequence [EPA (i)] with EPA not shortened [EPA (iii)].
656 Score 2: Comparison of EPA shortened at the genomic locations of FR or RF orientation
657 of DNA motif sequence [EPA (i)] with EPA shortened at the genomic locations of any
658 orientation of DNA 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	226	153	101
B2T2	221	149	98
B3T1	237	160	109
DNA motifs found in common among replications			
B2T1 and B2T2	212	141	95
B2T1 and B3T1	212	143	93
B2T2 and B3T1	208	138	93
B2T1, B2T2 and B3T1	201	133	90

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	395	262	205
B2T2	398	264	204
B3T1	394	263	203
DNA motifs found in common among replications			
B2T1 and B2T2	394	262	202
B2T1 and B3T1	391	261	202
B2T2 and B3T1	392	262	200
B2T1, B2T2 and B3T1	390	261	199

Forward-reverse (FR)

TF	Score 1	Score 2	TF	Score 1	Score 2	TF	Score 1	Score 2
AHR	38.81	13.41	ETS2	208.78	13.99	MNT	215.63	8.00
ARNTL	155.78	1.96	ETV1	153.32	7.02	MTF1	323.31	14.21
ATF1	323.31	15.00	ETV3	140.11	14.58	MXI1	323.31	70.98
ATF3	323.31	35.22	ETV5:EVX1	145.48	17.33	MYC	84.79	24.93
BCL	153.72	3.52	ETV5:HES7	8.76	8.74	MYCN	165.16	25.87
CREB	90.91	2.32	ETV5:TCF3	203.52	5.90	MZF1	81.83	2.29
CREB3	323.31	38.74	FOXN4	73.43	5.59	NFKB1	279.24	14.27
CREB3L1	323.31	27.96	FOXO1:ELK3	323.31	43.51	NFKB1:NFKB1	323.31	13.79
CTCF	323.31	3.50	GABPA	197.57	20.39	NFY	323.31	3.42
CTCFL	174.90	2.09	GABPA:GABPB1	142.28	6.10	NONO	323.31	9.24
CXXC1	69.92	3.30	GCM2	13.68	18.80	NR2C2	323.31	3.64
E2	323.31	20.15	GDNF	286.73	3.15	NRF1	45.13	30.75
E2F	122.12	21.87	GMEB1	136.68	10.56	OVOL2	323.31	31.61
E2F1	43.22	2.85	HIC1	223.55	4.56	PATZ1	101.85	9.30
EGR1	79.09	3.79	HIF1A	241.47	1.71	PLAG1	177.90	11.02
EGR2	176.69	4.53	HOXB2:ETV1	102.22	2.42	PURA	196.04	17.98
EGR3	85.25	34.18	ID2	323.31	14.52	R	22.49	3.99
EHF	323.31	8.66	INSM1	193.85	15.87	**RAD21**	258.21	4.81
ELF1	203.38	3.86	JUN:JUNB	275.78	6.48	RBAK	267.43	11.25
ELF2	227.62	7.83	JUNB:JUN	202.73	5.07	RELA	323.31	4.68
ELF5	323.31	8.76	KLF	252.10	8.35	RFX1	159.78	17.05
ELK1:TBX21	126.46	8.41	KLF15	323.31	22.56	RFX4	323.31	35.28
ELK1:TEF	323.31	7.99	KLF17	87.51	14.77	RFX5	323.31	26.52
ELK4	117.83	4.81	KLF2	90.74	15.34	SAP1	124.23	14.63
ENO1	80.42	3.13	KLF4	209.71	9.38	**SMC3**	222.09	1.76
EPAS1	65.43	7.76	KLF7	229.95	11.96	SP1	235.48	29.09
ERF	207.67	6.25	LHX8	244.38	2.95	SP2	94.11	9.54
ERG	199.81	25.70	MAX	323.31	8.35	SP4	14.30	6.31
ETS	196.80	1.63	MAZ	109.42	3.37	SPDEF	179.39	16.81
ETS1	308.00	3.92	MBD2	8.41	1.95	SPIC	323.31	28.49

TF	Score 1	Score 2
STAT	84.60	1.69
STAT1	323.31	5.38
TCF12	111.44	16.79
TCF3	323.31	14.05
TEAD4:ETV4	226.28	5.90
TFAP2A	25.86	20.26
TFAP2B	90.03	6.10
TFAP2C	169.60	8.55
TFAP4	93.98	2.36
TFE3	139.21	16.71
TP73	198.45	7.75
TR4	39.01	4.97
USF1	194.06	3.56
VEZF1	134.36	3.49
XBP1	323.31	44.58
XRCC5:XRCC6	132.56	3.92
YY1	62.59	5.30
ZBED6	183.14	20.23
ZBTB17	48.37	7.27
ZBTB26	151.91	16.84
ZBTB5	120.19	9.58
ZBTB7A	52.89	30.58
ZFX	13.33	14.55
ZFY	36.57	4.04
ZIC3	266.75	10.54
ZIC4	17.27	9.30
ZNF140	323.31	4.99
ZNF143	323.31	35.38
ZNF165	323.31	7.80
ZNF25	281.69	7.72

TF	Score 1	Score 2
ZNF322	120.49	11.55
ZNF325	69.42	2.74
ZNF408	214.73	2.22
ZNF467	127.21	4.64
ZNF479	226.76	3.26
ZNF48	106.17	9.88
ZNF511	84.74	9.87
ZNF553	93.45	9.17
ZNF555	232.83	7.63
ZNF624	235.50	3.84
ZNF672	266.64	10.80
ZNF675	323.31	5.81
ZNF676	323.31	16.54

Reverse-forward (RF)

TF	Score 1	Score 2
ARNT	323.31	14.12
ARNT2	41.41	42.78
ASCL2	138.15	6.41
ATF3	323.31	14.64
ATF4	323.31	25.14
ATF7	79.30	4.96
BCLAF1	148.24	8.52
BHLHE40	323.31	12.42
CDCA7L	84.72	8.42
CREB3	206.55	11.07
CREB3L1	323.31	7.63
CREB3L2	67.51	7.98
CTCF	240.61	4.56
E2F	78.16	7.07
E2F1	124.10	15.73
E2F3	65.67	8.85
E2F6	165.75	30.69
EFNA2	255.01	10.95
ELK3	143.48	7.57
EP300	121.01	10.90
ETV5:CLOCK	42.10	7.29
GLIS2	37.40	17.34
GMEB2	323.31	7.81
HES1	88.18	9.46
HES7	166.52	14.60
HIVEP1	323.31	4.90
HOXB2:ELK1	217.94	11.92
IRC900814	323.31	4.89
KLF4	213.35	9.49
KLF8	323.31	17.42

TF	Score 1	Score 2
MAX	304.34	8.64
MLXIPL	64.44	10.74
MUSCLE	23.05	3.47
MYB	323.31	17.68
MYBL1	316.23	23.27
MYC	323.31	12.23
MZF1	200.49	22.73
NFY	209.97	3.22
NFYA	323.31	11.75
NHLH1	105.19	4.42
NR1H3	323.31	9.38
NR3C1	43.16	8.31
OSR2	111.77	14.65
RFX1	323.31	22.99
RXRA	146.30	15.11
SALL1	277.84	1.91
SCRT2	301.48	9.60
SMAD6	259.24	7.43
SP100	20.75	7.99
SP2	294.86	2.59
SP4	110.94	19.63
SPDEF	185.85	14.35
SPZ1	323.31	9.87
SREBF1	67.27	15.74
SREBF2	297.27	1.62
TAF1	252.27	17.47
TATA	2.16	4.02
TFAP2A	20.17	29.15
TFAP2B	29.46	12.62
TFAP2C	112.10	51.21

TF	Score 1	Score 2
TFE3	95.23	10.59
THAP1	127.72	5.07
YY1	323.31	9.92
YY2	47.49	2.61
ZBTB33	90.29	7.78
ZBTB45	323.31	57.08
ZBTB7A	89.19	16.19
ZBTB7B	43.90	23.59
ZFP64	147.00	13.59
ZIC3	42.65	15.39
ZNF100	27.67	4.75
ZNF143	63.03	9.69
ZNF148	213.77	18.93
ZNF16	323.31	47.43
ZNF37A	310.67	14.79
ZNF410	140.45	13.57
ZNF419	247.63	11.25
ZNF429	87.05	32.39
ZNF436	323.31	11.48
ZNF468	21.12	7.67
ZNF557	299.44	9.58
ZNF575	39.45	14.38
ZNF586	260.73	6.18
ZNF611	279.80	15.65
ZNF616	189.53	30.98
ZNF692	202.47	2.33
ZNF727	214.01	17.22
ZNF770	300.52	4.71
ZNF773	150.45	8.84
ZSCAN21	323.31	5.55

Forward-reverse (FR)

TF	Score 1	Score 2
AHR	38.81	13.41
ARNTL	155.78	1.96
ATF3	323.31	35.22
BCL	153.72	3.52
CREB3L1	323.31	27.96
CTCF	323.31	3.50
CXXC1	69.92	3.30
E2F	122.12	21.87
E2F1	43.22	2.85
EGR1	79.09	3.79
EGR3	85.25	34.18
ELF1	203.38	3.86
ELF2	227.62	7.83
ELK4	117.83	4.81
ERG	199.81	25.70
ETS	196.80	1.63
ETS1	308.00	3.92
HIC1	223.55	4.56
INSM1	193.85	15.87
KLF4	209.71	9.38
KLF7	229.95	11.96
MAX	323.31	8.35
MYC	84.79	24.93
MYCN	165.16	25.87
MZF1	81.83	2.29
NFY	323.31	3.42
NONO	323.31	9.24
NRF1	45.13	30.75
OVOL2	323.31	31.61
PLAG1	177.90	11.02

TF	Score 1	Score 2
RAD21	258.21	4.81
RFX1	159.78	17.05
SMC3	222.09	1.76
SP2	94.11	9.54
TFAP2A	25.86	20.26
TFAP2B	90.03	6.10
TFAP2C	169.60	8.55
TP73	198.45	7.75
USF1	194.06	3.56
XBP1	323.31	44.58
YY1	62.59	5.30
ZBTB7A	52.89	30.58
ZFX	13.33	14.55
ZNF143	323.31	35.38

Reverse-forward (RF)

TF	Score 1	Score 2
ARNT	323.31	14.12
ATF3	323.31	14.64
BHLHE40	323.31	12.42
CTCF	240.61	4.56
E2F	78.16	7.07
E2F3	65.67	8.85
EFNA2	255.01	10.95
GMEB2	323.31	7.81
KLF4	213.35	9.49
MYB	323.31	17.68
MYC	323.31	12.23
MZF1	200.49	22.73
NFYA	323.31	11.75
NR3C1	43.16	8.31
OSR2	111.77	14.65
SP2	294.86	2.59
SP4	110.94	19.63
SPDEF	185.85	14.35
SREBF2	297.27	1.62
TATA	2.16	4.02
TFAP2B	29.46	12.62
TFAP2C	112.10	51.21
YY1	323.31	9.92
YY2	47.49	2.61
ZBTB33	90.29	7.78
ZBTB7B	43.90	23.59
ZNF100	27.67	4.75
ZNF143	63.03	9.69

Table 3. Correlation of expression level of gene pairs separated by a DNA motif site among 53 tissues. Pairs of genes separated by 64 FR and 52 RF orientations of DNA motifs of TF found in common in monocytes, T cells, HMEC and NPC were analyzed to find significantly lower correlation of their expression level than the correlation of all pairs of genes. Score: $-\log_{10}$ (p-value) of Mann-Whitney test. Median correlation: median correlation coefficient of expression level of nearby gene pairs separated by a DNA motif. The median correlation coefficient of expression level of all pairs of nearby genes is 0.17. # DNA site: number of DNA motif sites of TF separating gene pairs with expression data. Ubiquitously expressed genes with coefficient of variance < 90 and a

681 DNA motif site nearby were not used for statistical test, so the number of DNA site is
682 relatively small. The results using all genes were shown in Supplemental material. All
683 DNA motif sites of a TF were used including DNA motif sites examined for biased
684 orientation in EPA. The results using only biased orientation of DNA motif sites
685 examined in this study were shown in Supplemental material 2.

686

687 Forward-reverse (FR)

TF	Score	Median correlation	# DNA site
AHR	1.80	0.12	520
ARNT	1.35	0.09	141
ATF2	1.40	0.11	166
ATF3	3.61	0.11	795
BHLHE41	2.02	0.07	118
CREB3L1	2.88	0.05	127
CTCF	6.02	0.09	837
E2F	3.73	0.10	444
E2F1	3.79	0.11	868
EGR1	2.97	0.10	391
ELF1	4.86	0.10	714
ELF2	3.88	0.11	781
ELK3	1.86	0.12	547
ELK4	2.44	0.11	663
ETS	3.34	0.11	523
ETS1	4.27	0.11	687
ETV7	3.68	0.11	662
HIC1	1.65	0.11	380
HINFP	1.33	0.09	258
KLF4	2.00	0.13	718
MYC	3.86	0.11	762
MYCN	2.36	0.11	583
MZF1	1.74	0.10	261
NFYB	2.97	0.10	264
NRF1	2.21	0.11	640
PLAG1	2.93	0.10	503
PLAGL1	2.55	0.10	414
RAD21	6.08	0.09	852
REST	4.35	0.11	920
RFX1	2.38	0.08	81
SIN3A	3.31	0.11	813
SMC3	6.18	0.09	625
SP2	2.30	0.12	738
TFAP2A	4.17	0.11	767
TFAP2C	2.13	0.11	470
USF1	4.24	0.10	628
XBP1	6.65	0.07	379
ZBTB33	4.90	0.10	701
ZBTB7A	2.82	0.10	485
ZFP62	2.95	0.11	565
ZFX	2.69	0.11	581
ZIC1	4.47	0.11	829
ZNF770	3.38	0.08	271

Reverse-forward (RF)

TF	Score	Median correlation	# DNA site
AP1	3.87	0.11	744
ARNT	2.95	0.10	453
ATF3	4.34	0.10	665
BHLHE40	2.49	0.07	155
CREB1	1.79	0.11	303
CTCF	5.96	0.09	721
E2F	3.95	0.10	663
E2F3	4.05	0.09	616
EFNA2	3.51	0.10	543
EGR1	2.58	0.10	428
EGR2	1.39	0.13	747
ETS1	3.75	0.11	667
FLI1:FIGLA	5.41	0.08	428
FOS	3.51	0.11	777
GMEB2	1.36	0.09	310
HIF1A	2.34	0.11	448
KLF4	2.21	0.12	744
MITF	3.14	0.10	419
MYC	4.85	0.10	704
MYCN	2.79	0.10	507
NFKB1	1.59	0.11	342
NFYA	2.15	0.10	229
NRF1	2.01	0.12	615
OSR2	1.99	0.12	301
PBX3	2.20	0.11	253
RFX2	2.12	0.09	168
RFX5	3.03	0.05	75
RUNX3	2.01	0.11	241
SP1	2.73	0.11	725
SP2	2.15	0.12	1090
SP4	2.33	0.11	479
SPDEF	5.29	0.07	449
TFAP2B	3.42	0.11	855
TFAP2C	2.05	0.11	576
TFCP2L1	1.45	0.13	362
YBX1	1.34	0.13	160
YY1	2.23	0.01	50
ZBTB33	3.44	0.11	886
ZBTB7B	2.27	0.13	892
ZNF143	2.26	0.07	81

688

689

690 **Table 4.** Top 30 of overlapping biased orientations of DNA binding motif sequences of
691 transcription factors in monocytes. Overlapping DNA motif sequence of CTCF with
692 another biased orientation of DNA motif sequence were shown in a separate table.
693 Motif 1,2: DNA binding motif sequences of transcription factors. No.: the number of
694 genes with both Motif 1 and Motif 2 in upstream and downstream of genes (< 1 Mbp
695 distance from genes). Total number of genes is 31,227.

696

697 Forward-reverse (FR)

Motif 1	Motif 2	No.
FOS	JUN	3647
FOS	FOSL2	3284
CTCF	SMC3	3131
FOSL2	JUN	3072
FOSB	FOSL2	2897
FOSL2	JUNB	2870
EGR1	EGR2	2609
RAD21	SMC3	2580
CTCF	RAD21	2574
ATF3	FOSL2	2553
ETS2	ETSLIKE	2420
FOS	JUNB	2353
JUN	JUNB	2324
EGR1	WT1	2182
IRF	IRF4	2157
FOSL2	SMARCC1	2137
FOSB	JUN	1975
ELF1	ETS1	1920
ATF3	JUN	1882
JUN	TFAP2A	1790
FOSL2	TFAP2A	1769
BACH2	SMARCC1	1598
CHD2	E2F	1558
E2F1	RBL1	1485
JUN	SMARCC1	1434
JUN	NR3C1	1406
USF1	USF2	1399
BACH2	JUNB	1351
ETSLIKE	GABPA	1303
BACH2	FOSB	1298

Reverse-forward (RF)

Motif 1	Motif 2	No.
SP1	SP3	3763
EGR1	EGR3	3530
NR1I2	NR1I3	3199
ZIC1	ZIC3	3176
EGR	WT1	2917
FOS	FOS:JUN	2801
EGR	EGR3	2528
AP1	FOS	2524
MAFG	NFE2L1:MAFG	2496
KLF10	SP1	2485
EGR	EGR2	2387
KLF3	SP3	2267
GABPA	ZNF200	2113
KLF1	SP3	1961
KLF10	SP3	1898
AP1	JUNB	1843
FOS:JUN	JUNB	1800
KLF7	SP4	1791
KLF1	SP1	1753
KLF1	SP2	1746
SP2	SP3	1706
EGR1	WT1	1681
SP1	SP2	1644
CTCF	RXRA	1524
EHF	ELF1	1514
EGR1	SP1:SP3	1507
KLF1	KLF3	1463
ATF1	ATF3	1335
ELK1	ZNF200	1296
AP1	FOS:JUN	1288

FR

Motif 1	Motif 2	No.
CTCF	SMC3	3131
CTCF	RAD21	2574
CTCF	CTCFL	1263
CTCF	ZBTB7A	515
CTCF	E2F	140
CTCF	WT1	134
CTCF	MYC	120
CTCF	PLAGL1	108

RF

Motif 1	Motif 2	No.
CTCF	RXRA	1524
CTCF	CTCFL	788
CTCF	RAD21	439
CTCF	ZIC1	141
CTCF	ZIC3	120

698

699

700

701 **References**

702

703 Ackerman SL, Minden AG, Williams GT, Bobonis C, Yeung CY. 1991. Functional

704 significance of an overlapping consensus binding motif for Sp1 and Zif268 in
705 the murine adenosine deaminase gene promoter. *Proc Natl Acad Sci U S A* **88**:
706 7523-7527.
707 Ansari AZ, Peterson-Kaufman KJ. 2011. A partner evokes latent differences between
708 Hox proteins. *Cell* **147**: 1220-1221.
709 Bailey SD, Zhang X, Desai K, Aid M, Corradin O, Cowper-Sal Lari R, Akhtar-Zaidi B,
710 Scacheri PC, Haibe-Kains B, Lupien M. 2015. ZNF143 provides sequence
711 specificity to secure chromatin interactions at gene promoters. *Nat Commun* **2**:
712 6186.
713 Barutcu AR, Lajoie BR, Fritz AJ, McCord RP, Nickerson JA, van Wijnen AJ, Lian JB,
714 Stein JL, Dekker J, Stein GS et al. 2016. SMARCA4 regulates gene expression
715 and higher-order chromatin structure in proliferating mammary epithelial cells.
716 *Genome research* **26**: 1188-1201.
717 Chatterjee R, Zhao J, He X, Shlyakhtenko A, Mann I, Waterfall JJ, Meltzer P,
718 Sathyanarayana BK, FitzGerald PC, Vinson C. 2012. Overlapping ETS and CRE
719 Motifs ((G/C)CGGAAGTGACGTCA) preferentially bound by GABPalpha and
720 CREB proteins. *G3 (Bethesda)* **2**: 1243-1256.
721 Chen M, Manley JL. 2009. Mechanisms of alternative splicing regulation: insights from
722 molecular and genomics approaches. *Nat Rev Mol Cell Biol* **10**: 741-754.
723 Das D, Clark TA, Schweitzer A, Yamamoto M, Marr H, Arribere J, Minovitsky S,
724 Poliakov A, Dubchak I, Blume JE et al. 2007. A correlation with exon
725 expression approach to identify cis-regulatory elements for tissue-specific
726 alternative splicing. *Nucleic acids research* **35**: 4845-4857.
727 de Wit E, Vos ES, Holwerda SJ, Valdes-Quezada C, Verstegen MJ, Teunissen H,
728 Splinter E, Wijchers PJ, Krijger PH, de Laat W. 2015. CTCF Binding Polarity
729 Determines Chromatin Looping. *Molecular cell* **60**: 676-684.
730 Guo Y, Xu Q, Canzio D, Shou J, Li J, Gorkin DU, Jung I, Wu H, Zhai Y, Tang Y et al.
731 2015. CRISPR Inversion of CTCF Sites Alters Genome Topology and
732 Enhancer/Promoter Function. *Cell* **162**: 900-910.
733 He X, Syed KS, Tillo D, Mann I, Weirauch MT, Vinson C. 2015. GABPalpha Binding
734 to Overlapping ETS and CRE DNA Motifs Is Enhanced by CREB1: Custom
735 DNA Microarrays. *G3 (Bethesda)* **5**: 1909-1918.
736 Hsu SC, Gilgenast TG, Bartman CR, Edwards CR, Stonestrom AJ, Huang P, Emerson

737 DJ, Evans P, Werner MT, Keller CA et al. 2017. The BET Protein BRD2
738 Cooperates with CTCF to Enforce Transcriptional and Architectural Boundaries.
739 *Molecular cell* **66**: 102-116 e107.
740 Ji X, Dadon DB, Abraham BJ, Lee TI, Jaenisch R, Bradner JE, Young RA. 2015.
741 Chromatin proteomic profiling reveals novel proteins associated with
742 histone-marked genomic regions. *Proc Natl Acad Sci U S A* **112**: 3841-3846.
743 Jolma A, Yan J, Whitington T, Toivonen J, Nitta KR, Rastas P, Morgunova E, Enge M,
744 Taipale M, Wei G et al. 2013. DNA-binding specificities of human transcription
745 factors. *Cell* **152**: 327-339.
746 Jolma A, Yin Y, Nitta KR, Dave K, Popov A, Taipale M, Enge M, Kivioja T, Morgunova
747 E, Taipale J. 2015. DNA-dependent formation of transcription factor pairs alters
748 their binding specificity. *Nature* **527**: 384-388.
749 Karolchik D, Barber GP, Casper J, Clawson H, Cline MS, Diekhans M, Dreszer TR,
750 Fujita PA, Guruvadoo L, Haeussler M et al. 2014. The UCSC Genome Browser
751 database: 2014 update. *Nucleic acids research* **42**: D764-770.
752 Katainen R, Dave K, Pitkanen E, Palin K, Kivioja T, Valimaki N, Gylfe AE, Ristolainen
753 H, Hanninen UA, Cajuso T et al. 2015. CTCF/cohesin-binding sites are
754 frequently mutated in cancer. *Nature genetics* **47**: 818-821.
755 Kel AE, Gossling E, Reuter I, Cheremushkin E, Kel-Margoulis OV, Wingender E. 2003.
756 MATCH: A tool for searching transcription factor binding sites in DNA
757 sequences. *Nucleic acids research* **31**: 3576-3579.
758 Kheradpour P, Kellis M. 2014. Systematic discovery and characterization of regulatory
759 motifs in ENCODE TF binding experiments. *Nucleic acids research* **42**:
760 2976-2987.
761 Kin K, Chen X, Gonzalez-Garay M, Fakhouri WD. 2016. The effect of non-coding
762 DNA variations on P53 and cMYC competitive inhibition at cis-overlapping
763 motifs. *Hum Mol Genet* **25**: 1517-1527.
764 Kulakovskiy IV, Vorontsov IE, Yevshin IS, Sharipov RN, Fedorova AD, Rumynskiy EI,
765 Medvedeva YA, Magana-Mora A, Bajic VB, Papatsenko DA et al. 2018.
766 HOCOMOCO: towards a complete collection of transcription factor binding
767 models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic acids*
768 *research* **46**: D252-D259.
769 McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM,

770 Bejerano G. 2010. GREAT improves functional interpretation of cis-regulatory
771 regions. *Nature biotechnology* **28**: 495-501.
772 Merabet S, Mann RS. 2016. To Be Specific or Not: The Critical Relationship Between
773 Hox And TALE Proteins. *Trends Genet* **32**: 334-347.
774 Mumbach MR, Satpathy AT, Boyle EA, Dai C, Gowen BG, Cho SW, Nguyen ML,
775 Rubin AJ, Granja JM, Kazane KR et al. 2017. Enhancer connectome in primary
776 human cells identifies target genes of disease-associated DNA elements. *Nature*
777 *genetics* **49**: 1602-1612.
778 Neph S, Kuehn MS, Reynolds AP, Haugen E, Thurman RE, Johnson AK, Rynes E,
779 Maurano MT, Vierstra J, Thomas S et al. 2012. BEDOPS: high-performance
780 genomic feature operations. *Bioinformatics (Oxford, England)* **28**: 1919-1920.
781 Newburger DE, Bulyk ML. 2009. UniPROBE: an online database of protein binding
782 microarray data on protein-DNA interactions. *Nucleic acids research* **37**:
783 D77-82.
784 Osato N. 2018. Characteristics of functional enrichment and gene expression level of
785 human putative transcriptional target genes. *BMC Genomics* **19**: 957.
786 Phanstiel DH, Van Bortle K, Spacek D, Hess GT, Shamim MS, Machol I, Love MI,
787 Aiden EL, Bassik MC, Snyder MP. 2017. Static and Dynamic DNA Loops form
788 AP-1-Bound Activation Hubs during Macrophage Development. *Molecular cell*
789 **67**: 1037-1048.e1036.
790 Portales-Casamar E, Thongjuea S, Kwon AT, Arenillas D, Zhao X, Valen E, Yusuf D,
791 Lenhard B, Wasserman WW, Sandelin A. 2010. JASPAR 2010: the greatly
792 expanded open-access database of transcription factor binding profiles. *Nucleic*
793 *acids research* **38**: D105-110.
794 Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn
795 AL, Machol I, Omer AD, Lander ES et al. 2014. A 3D map of the human
796 genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**:
797 1665-1680.
798 Schreiber J, Libbrecht M, Bilmes J, Noble W. 2017. Nucleotide sequence and DNaseI
799 sensitivity are predictive of 3D chromatin architecture. *bioRxiv*.
800 Sherwood RI, Hashimoto T, O'Donnell CW, Lewis S, Barkal AA, van Hoff JP, Karun V,
801 Jaakkola T, Gifford DK. 2014. Discovery of directional and nondirectional
802 pioneer transcription factors by modeling DNase profile magnitude and shape.

803 *Nature biotechnology* **32**: 171-178.
804 Shlyueva D, Stampfel G, Stark A. 2014. Transcriptional enhancers: from properties to
805 genome-wide predictions. *Nature reviews Genetics* **15**: 272-286.
806 Slattery M, Riley T, Liu P, Abe N, Gomez-Alcala P, Dror I, Zhou T, Rohs R, Honig B,
807 Bussemaker HJ et al. 2011. Cofactor binding evokes latent differences in DNA
808 binding specificity between Hox proteins. *Cell* **147**: 1270-1282.
809 Tabuchi TM, Deplancke B, Osato N, Zhu LJ, Barrasa MI, Harrison MM, Horvitz HR,
810 Walhout AJ, Hagstrom KA. 2011. Chromosome-biased binding and gene
811 regulation by the Caenorhabditis elegans DRM complex. *PLoS Genet* **7**:
812 e1002074.
813 van Helden J. 2003. Regulatory sequence analysis tools. *Nucleic acids research* **31**:
814 3593-3596.
815 Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF,
816 Schroth GP, Burge CB. 2008. Alternative isoform regulation in human tissue
817 transcriptomes. *Nature* **456**: 470-476.
818 Wang J, Zhuang J, Iyer S, Lin X, Whitfield TW, Greven MC, Pierce BG, Dong X,
819 Kundaje A, Cheng Y et al. 2012a. Sequence features and chromatin structure
820 around the genomic regions bound by 119 human transcription factors. *Genome*
821 *research* **22**: 1798-1812.
822 Wang L, Wang S, Li W. 2012b. RSeQC: quality control of RNA-seq experiments.
823 *Bioinformatics (Oxford, England)* **28**: 2184-2185.
824 Wei Z, Gao F, Kim S, Yang H, Lyu J, An W, Wang K, Lu W. 2013. Klf4 organizes
825 long-range chromosomal interactions with the oct4 locus in reprogramming and
826 pluripotency. *Cell Stem Cell* **13**: 36-47.
827 Weintraub AS, Li CH, Zamudio AV, Sigova AA, Hannett NM, Day DS, Abraham BJ,
828 Cohen MA, Nabet B, Buckley DL et al. 2017. YY1 Is a Structural Regulator of
829 Enhancer-Promoter Loops. *Cell* **171**: 1573-1588 e1528.
830 Wingender E, Dietze P, Karas H, Knuppel R. 1996. TRANSFAC: a database on
831 transcription factors and their DNA binding sites. *Nucleic acids research* **24**:
832 238-241.
833 Xie X, Mikkelsen TS, Gnirke A, Lindblad-Toh K, Kellis M, Lander ES. 2007.
834 Systematic discovery of regulatory motifs in conserved regions of the human
835 genome, including thousands of CTCF insulator sites. *Proc Natl Acad Sci U S A*

836 **104**: 7145-7150.
837 Xie Z, Hu S, Blackshaw S, Zhu H, Qian J. 2010. hPDI: a database of experimental
838 human protein-DNA interactions. *Bioinformatics (Oxford, England)* **26**:
839 287-289.
840 Yoon SO, Chikaraishi DM. 1992. Tissue-specific transcription of the rat tyrosine
841 hydroxylase gene requires synergy between an AP-1 motif and an overlapping E
842 box-containing dyad. *Neuron* **9**: 55-67.
843 Zhang H, Li F, Jia Y, Xu B, Zhang Y, Li X, Zhang Z. 2017. Characteristic arrangement
844 of nucleosomes is predictive of chromatin interactions at kilobase resolution.
845 *Nucleic acids research* **45**: 12739-12751.
846 Zhao Y, Stormo GD. 2011. Quantitative analysis demonstrates most transcription factors
847 require only simple models of specificity. *Nature biotechnology* **29**: 480-483.
848