Discovery of biased orientations of regulatory motifs affecting human gene transcription and including chromatin-associated motifs

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 play essential roles in enhancer-promoter interactions (EPIs) and regulating the transcription of genes. CTCF and cohesin proteins are located at the anchors of chromatin interactions, forming their loop structures. Other DNA binding proteins such as YY1, ZNF143 and SMARCA4 are also reported to be involved in chromatin interactions. However, the overall picture of proteins associated with chromatin interactions and insulator function is still unclear. Here I describe a systematic and comprehensive approach to discover DNA binding motif sequences of transcription factors (TFs), affecting the interaction between enhancers and promoters of genes and their expression. This analysis identified 96 biased orientations of DNA motifs of TFs that affected the expression level of putative transcriptional target genes significantly in monocytes, T cells, HMEC and NPC in common, and included known TFs involved in chromatin interactions and insulator function such as CTCF, cohesin (RAD21 and SMC3), YY1 and ZNF143. The 96 biased orientations of DNA motifs contain 64 forward-reverse (FR) and 52 reverse-forward (RF) orientations of DNA motifs (Some TFs have more than one DNA motifs in databases, so the total number is smaller than the sum of FR and RF). The merger of DNA motif sites of biased orientation of a TF and CTCF showed a higher expression level of putative transcriptional target genes than DNA motif sites of only CTCF, and 42 TFs, except for CTCF, RAD21 and SMC3, indicated a higher expression level. Among 42 TFs, KLF4 and ERG were reported to be associated with chromatin interactions. Most of the other TFs were also known to have chromatin-associated functions. Moreover, to confirm the

predicted DNA motifs, first, they were compared with chromatin interaction data. Despite the limitation of resolution of chromatin interaction data, for 44 (69%) FR and 28 (54%) RF orientations of DNA motif sequences in $CD4^+$ T cells, EPIs predicted based on an enhancer-promoter association (EPA) that was shortened at the biased orientations of DNA motif sites, overlapped with chromatin interactions significantly more than other types of EPAs. Second, using gene expression data among 53 tissues, 43 (72%) FR and 40 (80%) RF orientations of DNA motifs showed significantly reduced correlation in expression level of nearby genes separated by the motif sites. Some DNA motif sites of biased orientations of DNA motifs of TFs were found between nearby genes not including DNA motif sites of the known TFs involved in chromatin interactions and between low correlation of expression level of nearby genes. These analyses suggest 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 play essential roles in enhancer-promoter interactions (EPIs) 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). The DNA

55 binding sequence of CTCF indicate its orientation bias at chromatin interaction anchors
56 – forward-reverse (FR) orientation is frequently observed (Rao et al. 2014; de Wit et al.
57 2015; Guo et al. 2015). Other DNA binding proteins such as ZNF143, YY1 and
58 SMARCA4 (BRG1) are involved in chromatin interaction and EPI (Bailey et al. 2015;
59 Barutcu et al. 2016; Weintraub et al. 2017). ZNF143 is known to bind more to promoter
60 of genes than distal regions, establishing looping with distal element bound by CTCF
61 (Bailey et al. 2015). DNA binding motif sequence of ZNF143 is significantly enriched
62 at both chromatin interaction anchors (Z-score > 7), whereas DNA binding motif
63 sequence of YY1 does not seem to be enriched (Z-score < 2; Bailey et al. 2015 Figure
64 2a). CTCF, cohesin, ZNF143, YY1 and SMARCA4 have various biological functions
65 other than chromatin interaction. These transcription factors (TFs) are bound in open
66 chromatin regions near transcriptional start sites (TSS) as well as chromatin interaction
67 anchors, which may reduce the enrichment of the DNA motif sites of the TFs at
68 chromatin interaction anchors.

69 Computational approaches to predict chromatin interactions were developed
70 (Schreiber et al. 2017; Zhang et al. 2017). However, they were not intended to find
71 DNA motif sequences of TFs involved in chromatin interactions. (Xie et al. 2005)
72 mentioned in Conclusion section that the next challenge will be to develop systematic
73 methods to discern the specific functions of these motifs in a genome-wide fashion.

74 DNA binding proteins involved in chromatin interactions are considered to affect
75 the transcription of genes in the loops formed by chromatin interactions. To confirm this
76 notion, I previously found that the expression level of human putative transcriptional

77 target genes of a TF was significantly changed, according to the criteria of
78 enhancer-promoter association (EPA) (Osato 2018). An EPA was shortened at the FR
79 orientation of CTCF binding sites, and transcriptional target genes of each TF bound in
80 enhancers and promoters were predicted based on the EPA (Fig. 1B). The expression
81 levels of the transcriptional target genes were significantly different, compared with the
82 expression levels of transcriptional target genes of TFs predicted based on only
83 promoters (or closed chromatin regions). The expression levels tend to be increased in
84 monocytes and $CD4^+$ T cells, implying that enhancers activate the transcription of genes.
85 The expression levels tend to be decreased in ES and iPS cells, implying that enhancers
86 repress the transcription of genes. The expression level of putative transcriptional target
87 genes of a TF bound in enhancers is changed significantly, when EPIs are predicted
88 properly using the insulator sites of CTCF in cell types. This suggests that the difference
89 of the expression level would reflect the accuracy of prediction of transcriptional target
90 genes of a TF based on a given EPA. Using this approach, I planned to search for other
91 DNA binding proteins involved in chromatin interactions, affecting the expression level
92 of putative transcriptional target genes through their insulator function.

93 Experimental analyses of chromatin interactions in molecular biology have
94 technical limitations: chromatin interaction data are altered, according to experimental
95 techniques, the number of reads of DNA sequencing and even replication sets of the
96 same cell type. Chromatin interaction data may not be saturated enough to contain all
97 chromatin interactions and may include experimental noises such as ligation error.
98 Avoiding the technical limitations of experimental analyses, here I searched

99 computationally for DNA motif sequences of TFs, affecting EPI and the expression
100 level of putative transcriptional target genes in monocytes, T cells, HMEC and NPC
101 without using chromatin interaction data. Then, putative EPIs and DNA motif sites of
102 TFs were compared with chromatin interaction and gene expression data.
103

104 ## Results

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

106 Transcription factor binding sites (TFBSs) were predicted using open chromatin
107 regions and DNA motif sequences of transcription factors (TF) collected from various
108 databases and journal papers (see Methods). Transcriptional target genes were predicted
109 based on TFBS in promoters and enhancers. An enhancer-promoter association (EPA)
110 was shortened at the DNA binding motif sites of a TF acting as insulator such as CTCF
111 and cohesin (RAD21 and SMC3) (Fig. 1B). To find DNA motif sequences of TFs with
112 insulator function, other than CTCF and cohesin, affecting the expression level of genes,
113 An EPA was shortened at the DNA motif sites of a putative insulator TF, and
114 transcriptional target genes of each general TF bound in (i) distal open chromatin
115 regions (enhancers) and (ii) closed chromatin regions in promoters were predicted based
116 on the EPA, in order to estimate the enhancer activity of each general TF. The ratio of
117 expression level of putative transcriptional target genes of each TF between (i) and (ii)
118 was calculated. The distribution of the ratios of all general TFs was compared among
119 forward-reverse (FR), reverse-forward (RF) orientations and non-orientation (i.e.
120 without considering orientation) of DNA motif sites of an insulator TF shortening an

121 EPA using Mann-Whitney test, two-sided (p-value < 10^{-7}), and putative insulator TFs
122 with significant differences of the distributions of the ratios of expression level among
123 FR, RF orientations and non-orientation were enumerated (Fig. 2A).

124 To confirm the prediction of TFBS, the distributions of TFBS were examined in
125 the human genome. TFBS involved in chromatin interactions were identified near TSS
126 (Fig. 2B). ZNF143 is known to bind to promoter of genes establishing looping with
127 distal element bound by CTCF (Fig. 5) (Bailey et al. 2015). YY1 is multi-functional
128 transcriptional regulator and is involved in transcriptional activation, repression and
129 initiation (Shrivastava and Calame 1994; Xi et al. 2007). Most of DNA motif sites of
130 the TFs were identified in intergenic and intron regions of the human genome (Fig. 2C).

131 Several hundreds of FR and RF orientations of DNA motifs of TFs were found in
132 monocytes, T cells, human mammary epithelial cells (HMEC) and neural progenitor
133 cells (NPC) (Fig. 3). The number of the TFs seemed to be changed, according to the
134 number of DNase-seq reads, since the number of the reads in monocyte is larger than
135 the other three cell types. When DNase-seq reads increase, more TFBS may be
136 predicted to some extent. Total 96 of biased (64 FR and 52 RF) orientations of DNA
137 binding motif sequences of TFs were discovered in the four cell types in common,
138 whereas non-orientation of DNA binding motif sequences of TFs were not detected (Fig.
139 3; Table 1; Supplemental material). The FR orientation of DNA motif sequences
140 included CTCF, cohesin (RAD21 and SMC3), YY1 and ZNF143, which are involved in
141 chromatin interaction and EPI.

142 Without considering the difference of orientations of DNA binding motif

143 sequences among cell types, 175 biased orientations of DNA binding motif sequences
144 of TFs were found in the four cell types in common. These numbers (96 and 175) are
145 unique numbers of TFs with biased orientations of DNA motifs, and not the unique
146 numbers of DNA motifs. Possible causes of the number 175 increased from 96 are as
147 follows: (1) a TF or an isoform of the same TF may bind to a different DNA binding
148 motif sequence, according to cell types and/or in the same cell type. About 50% or more
149 of alternative splicing isoforms are differently expressed among tissues, indicating that
150 most alternative splicing is subject to tissue-specific regulation (Das et al. 2007; Wang
151 et al. 2008; Chen and Manley 2009). (2) The same TF has several DNA binding motif
152 sequences, and in some cases one of the motif sequences is almost the same as the
153 reverse complement sequence of another motif sequence of the same TF in the same or
154 different databases of TFs. (3) I previously found that a complex of TFs would bind to a
155 slightly different DNA binding motif sequence from the combination of DNA binding
156 motif sequences of TFs composing the complex in *C. elegans* (Tabuchi et al. 2011). In
157 other reports, the difference of binding specificity of pairs of TFs was investigated
158 (Jolma et al. 2015). Cofactors also contribute to alter binding specificity of TFs (Ansari
159 and Peterson-Kaufman 2011; Slattery et al. 2011; Merabet and Mann 2016). These
160 increase the variety and number of (biased orientations of) DNA binding motif
161 sequences of a TF, which are potentially involved in different biological functions
162 among cell types.

163 There may be a possibility that DNA binding sites of biased orientations of TFs
164 were located near those of CTCF, so the biased orientations of TFs seem to have

165 insulator function. To address the issue, an EPA was shortened at the DNA binding sites
166 of biased orientations of CTCF and another biased orientation of a TF in 96 TFs found
167 in common among the four cell types. Transcriptional target genes of TFs were
168 predicted based on the EPA, and the ratio of expression level of transcriptional target
169 genes of a TF in open chromatin regions in the EPA and closed chromatin regions in
170 promoter was examined. The ratio was compared between an EPA shortened at the
171 DNA binding sites of CTCF and the merger of those of CTCF and another biased
172 orientation of a TF. If the DNA binding sites of biased orientation of a TF include
173 insulator sites not associated with those of CTCF, the ratio would be higher than an EPA
174 shortened at the DNA binding sites of only CTCF. However, there are many DNA
175 binding sites of CTCF in the human genome, the difference between the ratios may not
176 be large. The biased orientations of CTCF consist of 19 DNA binding motif sequences
177 in monocyte. When comparing the DNA sites of each DNA binding motif sequence of
178 CTCF with those of merger of the same CTCF and another biased orientation of a TF,
179 among 96 biased orientations of TFs, 79, 70, 95 and 62 TFs showed a higher ratio in
180 monocyte, $CD4^+$ T cell, HMEC and NPC respectively (p-value < 0.05) (Supplemental
181 material). YY1 was in a higher rank (the top in HMEC and NPC, second in T cell, and
182 fifth in monocyte). The rank of each TF was normalized by the lowest rank in each cell
183 type and was summed in the four cell types, where 45 TFs were found in common
184 (Table 2). The result is consistent that the DNA binding sites of YY1 are not frequently
185 located close to those of CTCF and are associated with chromatin interactions inside
186 chromatin interactions formed by CTCF (Weintraub et al. 2017) (Figure 2C, D, E and

187 Figure 7E in the reference). About 45 – 65% of DNA binding sites of ZNF143 are
188 closely located to those of CTCF in three cell types (Wen et al. 2018) (Figure 1a in the
189 reference). DNA binding sites of ZNF143 tend to be close to transcriptional start sites in
190 promoters. Other biased orientations of TFs potentially have DNA sites with insulator
191 function other than DNA sites of CTCF. Actually, KLF4 (Di Giammartino et al. 2019)
192 and ERG (Zhang et al. 2019; Li et al. 2020) were found to be associated with chromatin
193 interactions. Myc (Knoepfler et al. 2006; Sabo and Amati 2014), ETS family including
194 ELK3 (Yang et al. 2015), RUNX3 (Lee et al. 2019), USF1 (Huang et al. 2007; Li et al.
195 2016), AHR/ARNT (Swanson 2002), SP2 (Volkel et al. 2015), EGR2 (Mager et al.
196 2008; Frob and Wegner 2020; Wust et al. 2020), ZBTB33 (Blattler et al. 2013), NF-Y
197 (Oldfield et al. 2014), MITF (Levy and Fisher 2011; Laurette et al. 2015), NONO (Ma
198 et al. 2016), FLI1 (Riggi et al. 2014), E2F (Taubert et al. 2004; Blais and Dynlacht
199 2007; Subtil-Rodriguez et al. 2014) were associated with chromatin remodeling,
200 chromatin accessibility and/or histone modification. DNA sites of these TFs would
201 affect gene expression by chromatin-associated functions. (Huang et al. 2007)
202 mentioned that the insulator element at the 5' end of the chicken beta-globin locus acts
203 as a barrier, protecting transgenes against silencing effects of adjacent heterochromatin.
204 The transcription factor USF1 binds within the insulator and that this site is important
205 for generating in adjacent nucleosomes histone modifications associated with active
206 chromatin and, by inference, with barrier function. As another criterion of comparison,
207 when the DNA sites of DNA motif sequences of all biased orientations of CTCF were
208 merged, and an EPA was shortened at the merged DNA sites, among 96 biased

209 orientations of TFs, 1, 0, 34 and 14 TFs showed a higher ratio in monocyte, $CD4^+$ T cell,
210 HMEC and NPC respectively (p-value < 0.05), and included YY1 in HMEC and NPC,
211 but did not include YY1 and ZNF143 in monocyte and T cell (Supplemental material).
212 The criterion of the comparison seemed to be strict. DNA binding sites of CTCF
213 involved in non-insulator function and false-positive predictions of DNA binding sites
214 of CTCF might affect the result of the analysis.
215
216 **Comparison with chromatin interaction data**
217 To ask whether the biased orientations of DNA motif sequences are associated
218 with chromatin interaction, enhancer-promoter interactions (EPIs) predicted based on an
219 enhancer-promoter association (EPA) were compared with chromatin interaction data.
220 EPIs predicted based on an EPA were compared with three replications (B2T1, B2T2
221 and B3T1) of HiChIP chromatin interaction data in $CD4^+$ T cells separately (Mumbach
222 et al. 2017). The resolutions of HiChIP chromatin interaction data and EPIs were
223 adjusted to 5 kilobases (kb). The number of unique HiChIP cis (i.e. in the same
224 chromatin) chromatin interactions was 19,926,360 at 5 kb resolution in B2T1
225 replication of T cells, 666,149 at 5 kb resolution with ≥1,000 counts for each
226 interaction and 78,209 at 5 kb resolution with ≥6,000 counts for each interaction
227 (Supplemental material). EPI were predicted based on three types of EPAs: (i) EPA
228 shortened at the FR or RF orientation of DNA motif sites of a TF with putative insulator
229 function such as CTCF, (ii) EPA shortened at the DNA motif sites of a TF without
230 considering their orientation, and (iii) EPA without being shortened by DNA motif sites.

231 For total 201 biased orientations [133 (49%) FR and 90 (43%) RF of 273 FR and 211
232 RF] of DNA motif sequences in T cells, which included CTCF, cohesin (RAD21 and
233 SMC3), ZNF143 and YY1 in three replications (B2T1, B2T2, and B3T1), a
234 significantly higher ratio of EPIs overlapped with HiChIP chromatin interactions with
235 ≥1,000 counts for each interaction in EPA (i) than the other types of EPA (ii) and (iii) in
236 T cells (Table 3). When comparing EPIs predicted based on only EPA (i) and (iii) with
237 the chromatin interactions, for total 390 biased orientations [261 (96%) FR and 199
238 (94%) RF of 273 FR and 211 RF] of DNA motif sequences in T cells, a significantly
239 higher ratio of EPIs overlapped with the chromatin interactions in EPA (i) than EPA (iii)
240 (Table 3; Supplemental material). The difference between EPIs predicted based on EPA
241 (i) and (ii) seemed to be difficult to distinguish using the chromatin interaction data and
242 a statistical test in some cases. However, as for the difference between EPIs predicted
243 based on EPA (i) and (iii), a larger number of biased orientations of DNA motif
244 sequences were correlated with chromatin interaction data. Most of biased orientations
245 of DNA motif sequences (95%) were correlated with chromatin interactions, when
246 comparing EPIs predicted based on EPA (i) and (iii) with HiChIP chromatin interactions.
247 Chromatin interaction data were obtained from different samples from DNase-seq open
248 chromatin regions, so individual differences may exist. Individual differences of
249 chromatin interactions were larger than those of open chromatin regions (Mumbach et
250 al. 2017).

251 Additionally, to confirm these tendencies of comparison of EPIs with HiChIP
252 chromatin interactions, the same analysis was conducted using HiChIP chromatin

253 interaction data in Th17, Treg and GM12878 cells. For total 128 biased orientations [76
254 (38%) FR and 64 (35%) RF of 200 FR and 183 RF] of DNA motif sequences in Th17
255 cells, which included CTCF, cohesin (RAD21 and SMC3), ZNF143 and YY1 in three
256 replications (B1T2, B2T1, and B3T1), a significantly higher ratio of EPIs overlapped
257 with HiChIP chromatin interactions in EPA (i) than the other types of EPA (ii) and (iii)
258 (Supplemental material). When comparing EPIs predicted based on only EPA (i) and
259 (iii) with the chromatin interactions, for total 282 biased orientations [182 (91%) FR
260 and 162 (89%) RF of 200 FR and 183 RF] of DNA motif sequences, a significantly
261 higher ratio of EPIs overlapped with the chromatin interactions in EPA (i) than EPA (iii)
262 (Supplemental material).

263 In Treg cells, chromatin interactions in one biological replication were overlapped
264 with EPIs, but those in the other two biological replications were less overlapped with
265 EPIs, so the one replication (B3T1) was used for this analysis. For total 281 biased
266 orientations [154 (53%) FR and 176 (54%) RF of 290 FR and 323 RF] of DNA motif
267 sequences, which included CTCF, cohesin (RAD21 and SMC3), ZNF143 and YY1, a
268 significantly higher ratio of EPIs overlapped with HiChIP chromatin interactions in EPA
269 (i) than the other types of EPA (ii) and (iii) (Supplemental material). When comparing
270 EPIs predicted based on only EPA (i) and (iii) with the chromatin interactions, for total
271 482 biased orientations [283 (98%) FR and 309 (96%) RF of 290 FR and 323 RF] of
272 DNA motif sequences, a significantly higher ratio of EPIs overlapped with the
273 chromatin interactions in EPA (i) than EPA (iii) (Supplemental material).

274 In GM12878 cells, for total 168 biased orientations [96 (38%) FR and 98 (35%)

275 RF of 229 FR and 237 RF] of DNA motif sequences, which included CTCF, cohesin
276 (RAD21 and SMC3), ZNF143 and YY1 in two replications (B1 and B2), a significantly
277 higher ratio of EPIs overlapped with HiChIP chromatin interactions in EPA (i) than the
278 other types of EPA (ii) and (iii) (Supplemental material). When comparing EPIs
279 predicted based on only EPA (i) and (iii) with the chromatin interactions, for total 366
280 biased orientations [218 (95%) FR and 227 (96%) RF of 229 FR and 237 RF] of DNA
281 motif sequences, a significantly higher ratio of EPIs overlapped with the chromatin
282 interactions in EPA (i) than EPA (iii) (Supplemental material).

283 In situ Hi-C data in HMEC, which are available in 4D nucleome data portal, were
284 compared with putative EPIs. The number of unique in situ Hi-C cis chromatin
285 interactions was 121,873,642 at 3 kb resolution. For total 222 biased orientations [155
286 (50%) FR and 108 (40%) RF of 310 FR and 269 RF] of DNA motif sequences, which
287 included CTCF, cohesin (RAD21 and SMC3), ZNF143 and YY1, a significantly higher
288 ratio of EPIs overlapped with in situ Hi-C chromatin interactions (≥1 count for each
289 interaction, i.e. all interactions) in EPA (i) than the other types of EPA (ii) and (iii)
290 (Supplemental material). When comparing EPIs predicted based on only EPA (i) and
291 (iii) with the chromatin interactions, for total 406 biased orientations [268 (86%) FR
292 and 230 (86%) RF of 310 FR and 269 RF] of DNA motif sequences, a significantly
293 higher ratio of EPIs overlapped with the chromatin interactions in EPA (i) than EPA (iii)
294 (Supplemental material). There were six replications of in situ Hi-C experimental data
295 of HMEC, but other replications contained much smaller number of chromatin
296 interactions. The HiChIP chromatin interaction data in T cells were expected to be

297 enriched with enhancer-promoter interactions, since they were produced with H3K27ac
298 antibody to capture chromatin interactions. In situ Hi-C data consist of all types of
299 chromatin interactions and thus, contain a smaller number of enhancer-promoter
300 interactions than HiChIP data. Chromatin interactions with a small number of counts
301 may be artifacts occurred by ligation errors during experiments (Lieberman-Aiden et al.
302 2009; Yardimci et al. 2019). Therefore, other replications of in situ Hi-C data with a
303 smaller number of chromatin interactions will not be available for this analysis.

304 Promoter capture Hi-C data of NPC was downloaded and analyzed in the same
305 procudure as T cell and HMEC (Jung et al. 2019). The number of unique promoter
306 capture Hi-C cis chromatin interactions was 9,436,689 at 5 kb resolution. For total 109
307 biased orientations [61 (28%) FR and 60 (25%) RF of 205 FR and 241 RF] of DNA
308 motif sequences, which included CTCF, cohesin (RAD21 and SMC3), ZNF143 and
309 YY1, a significantly higher ratio of EPIs overlapped with promoter capture Hi-C
310 chromatin interactions (≥1 count for each interaction, i.e. all interactions) in EPA (i)
311 than the other types of EPA (ii) and (iii) (Supplemental material). When comparing EPIs
312 predicted based on only EPA (i) and (iii) with the chromatin interactions, for total 186
313 biased orientations [99 (48%) FR and 103 (43%) RF of 205 FR and 241 RF] of DNA
314 motif sequences, a significantly higher ratio of EPIs overlapped with the chromatin
315 interactions in EPA (i) than EPA (iii) (Supplemental material).

316 All three types of chromatin interaction data (HiChIP, in situ Hi-C and promoter
317 capture Hi-C) indicated that EPIs predicted based on EPA (i) that was shortened at DNA
318 motif sites of known TFs involved in chromatin interactions (CTCF, RAD21, SMC3,

319 ZNF143 and YY1), overlapped with more chromatin interactions than the other EPA (ii)
320 and (iii) in six cell types (T cells, Th17, Treg, GM12878, HMEC and NPC). In
321 monocyte, Hi-C data was available, but it was not analyzed, due to the low resolution of
322 chromatin interaction data (50 kb).

323 The biased orientations of DNA motif sites of known TFs involved in chromatin
324 interactions such as CTCF, RAD21, SMC3, YY1 and ZNF143 may be overlapped with
325 the biased orientations of DNA motif sites of other TFs, and thus, they showed a
326 significantly higher ratio of EPIs overlapped with chromatin interactions in EPA (i) than
327 the other types of EPA (ii) and (iii). However, among 273 FR and 211 RF orientations of
328 DNA motifs of TFs not including the known TFs in T cells, the DNA motif sites of only
329 15 (6%) FR and 9 (4%) RF orientations of DNA motifs of TFs overlapped with $\geq 60\%$
330 of DNA motif sites of biased orientations of DNA motifs of either of the known TFs at 5
331 kb resolution.

332 Moreover, to investigate the enhancer activity of EPIs, the distribution of
333 expression level of putative target genes of EPIs was compared between EPIs
334 overlapped with HiChIP chromatin interactions and EPIs not overlapped with them.
335 Though the target genes of EPIs were selected from top 4,000 transcripts (genes) in
336 terms of the expression level of all transcripts (genes) excluding transcripts not
337 expressed in T, Th17, Treg and GM12878 cells separately, target genes of EPIs
338 overlapped with chromatin interactions indicated a significantly higher expression level
339 than EPIs not overlapped with them, suggesting that EPIs overlapped with chromatin
340 interactions activated the expression of target genes in the four cell types. For almost all

341 (99% - 100%) FR and RF orientations of DNA motifs, a significantly higher expression
342 level of putative target genes of EPIs overlapped with chromatin interactions than EPIs
343 not overlapped in the four cell types. When a biased orientation of DNA motif showed a
344 significantly higher expression level of putative target genes of EPIs overlapped with
345 chromatin interactions than EPIs not overlapped, '1' was marked with in the tables of
346 comparison between EPIs and HiChIP chromatin interactions in Supplemental material.
347 DNA motifs showing a significantly lower expression level or not showing a significant
348 difference of expression level were not observed in this analysis. HiChIP data were
349 produced using H3K27ac antibody, so chromatin interactions acting as repressor would
350 not be identified. However, for in situ Hi-C and promoter capture Hi-C data in HMEC
351 and NPC, the significant difference of expression level of putative target genes of EPIs
352 was observed in a small percentage (less than 1%) of biased orientations of DNA
353 motifs.

354 Recently high resolution of chromatin interaction data were publicly available
355 (Krietenstein et al. 2020). The biased orientations of DNA sites of TFs were examined
356 using Micro-C chromatin interaction data in human fibroblast. The resolution of the
357 data was adjusted to 200 b. DNA sites of CTCF, RAD21 and SMC3 showed a
358 significant directional bias at the pairs of chromatin interaction anchors. DNA sites of
359 SMRC2 showed a directional bias at the anchors. SMARCA4 is known to be involved
360 in higher-order chromatin structure such as topologically associating domain (TAD)
361 (Barutcu et al. 2016). DNA motifs of SMARCA4 did not exist in the databases of DNA
362 motifs of TFs used in this analysis. SMRC2 was not identified in this analysis, because I

363 searched for DNA sites of TFs affecting EPI and gene expression, which are not
364 high-order chromatin structure. ATF3 and ZBTB7A in Table 2 were found among top 32
365 biased orientations of DNA sites of TFs at the pairs of chromatin interaction anchors
366 with p-value < 10^{-50}. Deletion of an ATF3 binding site affected the expression of a target
367 gene, which was far apart from the site (Kim et al. 2017). ZBTB7A is involved in the
368 changes in chromatin accessibility and the regulation of gene expression (Ramos Pittol
369 et al. 2018). DNA sites of YY1 and ZNF143 also indicated a weak directional bias at the
370 chromatin interaction anchors. In this analysis, biased orientations of DNA sites of TFs
371 were investigated based on the difference of distribution of expression level of putative
372 transcriptional target genes, and YY1 and ZNF143 showed significant directional biases,
373 implying that the analysis in this study is able to find TFs that do not have a strong
374 directional bias based on the number of DNA sites at chromatin interaction anchors.
375 Most of TFs (37/39), except for CTCF, RAD21 and SMC3, in Table 2 showed weak
376 directional biases at the chromatin interaction anchors.
377
378 **Correlation of expression level of gene pairs separated by DNA motif sites of a TF**
379 To ask whether the biased orientations of DNA motif sites affect gene expression,
380 I compared the expression level in monocytes of protein-coding genes the most closely
381 located to upstream and downstream of biased orientations of DNA motif sites in the
382 human genome. The expression levels of closely located genes including divergent gene
383 pairs tend to be correlated among tissues (Purmann et al. 2007; Xie et al. 2007). When a
384 biased orientation of DNA motif sites serve as insulator (e.g. CTCF), the correlation of

385 gene expression level would be reduced. Among 96 biased (64 FR and 52 RF)
386 orientations of DNA motifs of TFs found in common in monocytes, T cells, HMEC and
387 NPC, TFs with ≥50 genomic locations of the DNA motif sites were selected, after the
388 elimination of ubiquitously expressed genes among 53 tissues and DNA motif sites near
389 the genes with coefficient of variance (CV) <90. After this filtering, 43 (72%) FR and
390 40 (80%) RF of 60 FR and 50 RF showed significantly lower correlations of expression
391 levels of gene pairs separated by the DNA motif sites of a TF than the correlations of all
392 pairs of neighbor genes with intergene distance <1 megabase (Mb) (Mann-Whitney test,
393 p-value < 0.05) (Table 4). The FR and RF orientations of TFs included known TFs
394 involved in chromatin interactions such as CTCF, RAD21, SMC3, YY1 and ZNF143.
395 Top 20 of the FR and RF orientations of DNA motifs were selected in ascending order
396 of the median correlation coefficient of expression levels of gene pairs separated by the
397 DNA motif sites of a TF. The 20 TFs included known TFs involved in chromatin
398 interactions such as CTCF, RAD21 and SMC3 (Fig. 4) (Mann-Whitney test, p-value <
399 0.01). Instead of using all DNA motif sites of a TF, when DNA motif sites of a TF were
400 limited to sites examined for biased orientations of DNA motifs in an EPA, 33 (60%)
401 FR and 36 (77%) RF of 55 FR and 47 RF showed significantly lower correlations of
402 expression levels of gene pairs separated by the DNA motif sites (Supplemental
403 material). With intergene distance <20 kb, 21 (34%) FR and 8 (15%) RF of 62 FR and
404 52 RF showed significantly lower correlations of expression levels of gene pairs
405 separated by the DNA motif sites of a TF than the correlations of all pairs of neighbor
406 genes (Supplemental material).

407 A large number of DNA motif sites of a TF tend to show significantly lower
408 correlations of gene expression (Median numbers of DNA motif sites showing
409 significantly lower correlations were 656 and 479 for FR and RF respectively, and
410 median numbers of DNA motif sites not showing lower correlations significantly were
411 122 and 115 for FR and RF respectively. Mann-Whitney test p-value = 0 and < 10^{-10} for
412 FR and RF respectively). This suggests that if the criteria to examine the correlations of
413 expression levels of gene pairs separated by DNA motif sites are more stringent, and
414 thus the number of DNA motif sites is reduced, it would be more difficult to identify
415 using a statistical test the difference of correlations of expression levels of gene pairs
416 separated by the reduced DNA motif sites of TFs, particularly for TFs with a small
417 number of DNA motif sites.

418 Among 91 biased orientations of DNA motifs of TFs found in common in the
419 four cell types, after removing five known TFs involved in chromatin interactions such
420 as CTCF, RAD21, SMC3, YY1 and ZNF143, DNA motif sites of 88 TFs were located
421 at ≥1 genomic regions between gene pairs not including DNA motif sites of the known
422 TFs, and at ≥1 genomic regions between low correlation (< 0.13) of expression level of
423 gene pairs (when ubiquitously expressed genes among 53 tissues were eliminated with
424 CV <90. The median correlation coefficient of expression level of all pairs of nearby
425 genes is 0.17, Table 4). DNA motif sites of 52 TFs were placed at ≥10 genomic
426 regions.

427

428 **Co-location of biased orientations of DNA motif sites of TFs**

429 To examine the association of biased orientations of DNA binding motif
430 sequences of TFs, co-location of the DNA binding motif sequences within 200bp to
431 each other, which is close to the length of core DNA around octamer of core histone
432 (146bp), that is, the length of the minimum open chromatin region, was analyzed in
433 monocytes, T cells, HMEC and NPC. The same pairs of DNA binding motif sequences
434 in upstream and downstream of genes in EPAs were enumerated, and the pairs of DNA
435 binding motif sequences found in ≥ 100 genomic regions were listed (Table 5;
436 Supplemental material). Majority of pairs of DNA motif sites overlapped with more
437 than 1 base, but the same tendency was found in the analysis of DNA motif sites of all
438 TFs (Supplemental material). CTCF was partially overlapped with cohesin RAD21 and
439 SMC3 (Table 5). Top 30 pairs of FR and RF orientations of DNA motifs overlapped or
440 co-localized were shown (Table 5). Total numbers of unique pairs of DNA motif sites of
441 TFs overlapped or co-localized were 404 (FR) and 662 (RF) in monocytes, consisting of
442 177 (39% of all FR) and 218 (43% of all RF) unique DNA motifs. Only ten overlapping
443 or co-location of DNA motifs of TFs were identified in the four cell types in common,
444 including pairs of CTCF, RAD21, SMC3 and YY1. This implies that most of
445 overlapping and co-location of DNA motifs of TFs serve in cell-type-specific manner
446 (Supplemental material). Pairs of biased orientations of DNA binding motifs of TFs
447 tend to overlap, and relatively a small number of motifs are co-localized in upstream
448 and downstream of genes. Overlapping DNA motifs act as DNA binding sites of several
449 TFs simultaneously or competitively (Ackerman et al. 1991; Yoon and Chikaraishi
450 1992; Ansari and Peterson-Kaufman 2011; Slattery et al. 2011; Chatterjee et al. 2012;

451 He et al. 2015; Kin et al. 2016; Merabet and Mann 2016). CTCF and cohesin function
452 coordinately with another TF and a protein such as Klf4 and Brd2 (Wei et al. 2013; Hsu
453 et al. 2017). Klf4 recruits cohesin to the Oct4 enhancer region. BRD2 supports
454 boundary (insulator) activity of CTCF. The analysis of 457 ChIP-seq data sets on 119
455 human TFs showed that secondary motifs were detected in addition to the canonical
456 motifs of the TFs, indicating tethered binding and co-binding between multiple TFs
457 (Wang et al. 2012a). The authors observed significant position and orientation
458 preferences between many co-binding TFs. Overlapping DNA motif sites of TFs found
459 in this study may not be artifact, but biologically functional. Though repeat sequences in
460 human genome were masked, more than a hundred of pairs of DNA motif sites
461 overlapped at hundreds or thousands upstream and downstream regions of genes.
462

463 **Discussion**

464 The approach in this study is an initial comprehensive method to find DNA motif
465 sequences of transcription factors (TFs) affecting enhancer-promoter interactions (EPIs)
466 and gene expression, such as CTCF and cohesin (RAD21 and SMC3), without using
467 chromatin interaction data. DNA motif sequences were searched from open chromatin
468 regions of monocytes, T cells, HMEC and NPC. Total 96 biased [64 forward-reverse
469 (FR) and 52 reverse-forward (RF)] orientations of DNA motif sequences of TFs were
470 found in the four cell types in common, whereas non-orientation (i.e. without
471 considering orientation) of DNA motif sequences of TFs were not found to affect the
472 expression level of putative transcriptional target genes, suggesting that an

473 enhancer-promoter association (EPA) shortened at the FR or RF orientation of DNA
474 motif sites of a TF includes more accurate prediction of EPIs, which were supported by
475 the comparisons with chromatin interaction and gene expression data. FF and RR
476 orientations of DNA sites of TFs were not examined in this analysis, since CTCF has
477 FR or RF orientation of its DNA sites. Among 96 biased orientations of DNA motif
478 sequences of TFs, 42, except for CTCF, RAD21 and SMC3, showed a different
479 distribution of expression level of putative transcriptional target genes in comparison
480 with CTCF under a statistical threshold; four of the 42 were reported to be associated
481 with chromatin interactions and most of the others were known to have
482 chromatin-associated functions. The analysis of biased orientations of DNA sites of the
483 42 TFs at pairs of chromatin interaction anchors indicated a weak statistical significance,
484 except for two TFs, suggesting that 40 TFs were difficult to find (Bailey et al. 2015).

485 To predict DNA motif sites of TFs, I tried several tools and methods, and then
486 used protein interaction quantitation (PIQ) tool (Sherwood et al. 2014). PIQ predicts
487 DNA motif sites of TFs by reducing noises of DNase-seq data, and the comparison of
488 the prediction with ChIP-seq data has revealed that PIQ shows a better performance
489 than other tools and digital genome footprinting (DNase-DGF). The prediction of DNA
490 motif sites of TFs using PIQ in this study increased the number of biased orientations of
491 DNA motif sequences of TFs, reducing the number of non-orientation of DNA motif
492 sequences of TFs. At first, the number of predicted DNA motif sites of TFs was
493 expected to increase, relative to using consensus DNA sequences of TFs, which seemed
494 to be stringent to find all DNA motif sites of a TF in the genome. However, the number

495 of DNA motif sites was almost the same or slightly decreased, compared with the
496 analysis of consensus DNA sequences, suggesting that PIQ increases true positive
497 predictions of DNA motif sites, reducing false positive predictions.

498 When forming a homodimer or heterodimer with another TF, TFs may bind to
499 genome DNA with a specific orientation of their DNA binding sequences (Fig. 5). The
500 approach in this study would be able to identify TFs forming heterodimer. If the DNA
501 motif site of only the mate to a pair of TFs was found in an EPA, the EPA was shortened
502 at one side of the DNA motif site, and transcriptional target genes were predicted based
503 on the EPA shortened at the side. In this analysis, the mate to both heterodimer and
504 homodimer of TFs can be tested to examine the effect on the expression level of
505 transcriptional target genes predicted based on the EPA shortened at one side. Biased
506 orientations of DNA motif sequences may also be found in forward-forward or
507 reverse-reverse orientation.

508 TFs that are multi-functional transcriptional regulator such as YY1 may not
509 indicate a strong directional bias of their DNA motif sites. Though some proportion of
510 DNA motif sites of a TF involved in chromatin interaction and insulator function lie in
511 biased orientation in the genome, the other DNA motif sites of the TF may not form a
512 dimer with another TF, and not be present in a specific orientation in the genome. Thus,
513 overall proportion of biased orientations of DNA motif sites of a TF may be reduced, so
514 that a statistical test cannot detect it.

515 The approach in this study was tested to search for repeat sequences serving as
516 insulator sites and affecting the expression level of putative transcriptional target genes

517 of TFs. A small number of repeat sequences were identified and included MIR, which
518 acts as insulators (Wang et al. 2015).
519 CTCF and cohesin-binding sites are frequently mutated in cancer (Katainen et al.
520 2015). Some biased orientations of DNA motif sequences would be involved in
521 chromatin interactions and might be associated with diseases including cancer.
522 The analyses taken here suggest that biased orientations of DNA motifs of
523 discovered TFs affect enhancer-promoter interactions and the expression level of
524 putative transcriptional target genes of general TFs bound in enhancers and promoters.
525 This gives insight into the nature of TFs and contributes to the researches of functions
526 and molecular mechanisms of chromatin interactions, regulation of chromatin, histone
527 modifications, transcriptional regulation and transcriptional cascades.
528

529 ## Methods

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

531 To identify transcription factor binding sites (TFBSs) from open and closed
532 chromatin regions, TRANSFAC (2019.2), JASPAR (2018), UniPROBE (2018),
533 high-throughput SELEX, transcription factor binding sequences of ENCODE ChIP-seq
534 data, and HOCOMOCO version 9 and 11 were used to predict insulator sites
535 (Wingender et al. 1996; Newburger and Bulyk 2009; Portales-Casamar et al. 2010; Xie
536 et al. 2010; Zhao and Stormo 2011; Jolma et al. 2013; Kheradpour and Kellis 2014;
537 Kulakovskiy et al. 2018). TRANSFAC (2019.2) was used to analyze enhancer-promoter
538 interactions, because these data were sufficient to identify biased orientations of DNA

539 motif sequences of insulator TFs with less computational time, reducing the number of
540 non-orientation of DNA motif sequences of TFs. To reduce false positive prediction of
541 DNA motif sites of TFs in open chromatin regions, DNA motif search was performed
542 by protein interaction quantitation (PIQ) tool with 12,249 position frequency matrices
543 converted from DNA motifs of vertebrate TFs in the above databases and DNase-seq
544 data from GEO and ENCODE database (GSM1024791 $CD14^+$ monocytes; GSM665812
545 $CD4^+$ T cell; GSM736634 Human mammary epithelial cell, HMEC; GSM878615
546 Neural progenitor cell, NPC; ENCFF263GMV, ENCFF886XEV Th17; GSM1024741
547 Treg; GSM736620 GM12878) (Sherwood et al. 2014). To find open chromatin regions
548 associated with enhancer (and promoter) activity, narrow peaks of ChIP-seq
549 experimental data of H3K27ac were predicted using macs2 callpeak (GEO:
550 GSM773004) (Zhang et al. 2008). To predict DNA motif sites of TFs from closed
551 chromatin regions, another method was employed. Position weight matrices of
552 vertebrate transcription factor binding sequences were converted into DNA consensus
553 sequences, which consist of alphabet of 15 characters (the four bases A, C, G, T, the six
554 two-fold degenerate IUPAC codes R=[AG], Y=[CT], K=[GT], M=[AC], S=[GC],
555 W=[AT], the four three-fold degenerate IUPAC codes B=[CGT], D=[AGT], H=[ACT],
556 V=[ACG], and the four-fold degenerate character N=[ATGC]) using convert matrix in
557 RSAT website (van Helden 2003). Transcription factor binding sequences were
558 searched from narrow peaks of DNase-seq data in repeat-masked hg19 genome
559 sequences using Match tool in TRANSFAC with similarity score cutoff of 1 after
560 transforming the DNA consensus sequences of TFs into position frequency matrixes for

561 Match search (Kel et al. 2003). Repeat-masked hg19 genome sequences were
562 downloaded from UCSC genome browser (http://genome.ucsc.edu/,
563 http://hgdownload.soe.ucsc.edu/goldenPath/hg19/bigZips/hg19.fa.masked.gz). TFs
564 corresponding to transcription factor binding sequences were searched computationally
565 and manually by comparing their names and gene symbols of HGNC (HUGO Gene
566 Nomenclature Committee) -approved gene nomenclature and 31,848 UCSC known
567 canonical transcripts
568 (http://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/knownCanonical.txt.gz),
569 because transcription factor binding sequences were not linked to transcript IDs such as
570 UCSC, RefSeq and Ensembl transcripts.

571 Target genes of a TF were assigned when its TFBS was found in DNase-seq
572 narrow peaks in promoter or extended regions of enhancer-promoter association of
573 genes (EPA). Promoter and extended regions were defined as follows: promoter regions
574 were those that were within distance of ±5 kb from transcriptional start sites (TSS).
575 Promoter and extended regions were defined as per the following association rule,
576 which is the same as that defined in Figure 3A of a previous study (McLean et al. 2010):
577 the single nearest gene association rule, which extends the regulatory domain to the
578 midpoint between the TSS of the gene and that of the nearest gene upstream and
579 downstream without the limitation of extension length. The limitation of extension
580 length was ＜1 Mb in this study. Extended regions for an EPA were shortened at the
581 DNA motif sites of a putative insulator TF that were the closest to a TSS, and
582 transcriptional target genes were predicted based on the shortened EPA and DNA motif

583 sites of a general TF in open chromatin regions. Promoter and extended regions of an
584 EPA were shortened at the forward–reverse (FR) orientation of DNA binding sites of a
585 putative insulator TF. When forward or reverse orientation of DNA binding sites were
586 continuously located in the genome several times, the most internal (i.e. closest to a
587 TSS) forward–reverse orientation of DNA binding sites were selected. The genomic
588 positions of genes were identified using 'knownGene.txt.gz' file in UCSC
589 bioinformatics sites (Karolchik et al. 2014). The file 'knownCanonical.txt.gz' was also
590 utilized to select representative transcripts among various alternate forms for assigning
591 promoter and extended regions of an EPA. From the list of transcription factor binding
592 sequences and transcriptional target genes, redundant transcription factor binding
593 sequences were removed by comparing the putative target genes of a transcription
594 factor binding sequence and its corresponding TF; if identical, one of the transcription
595 factor binding sequences was chosen. When the number of transcriptional target genes
596 predicted based on a transcription factor binding sequence was less than five, the
597 transcription factor binding sequence was omitted.

598 For gene expression data, RNA-seq reads mapped onto the human hg19 genome
599 sequence were obtained from UCSF-UBC human reference epigenome mapping project
600 RNA-seq reads with poly-A of naive $CD4^+$ T cells (GEO: GSM669617). For monocytes,
601 Blueprint RNA-seq FPKM data
602 ('C0010KB1.transcript_quantification.rsem_grape2_crg.GRCh38.20150622.results')
603 were downloaded from Blueprint DCC portal
604 (http://dcc.blueprint-epigenome.eu/#/files). RNA-seq reads with poly-A of human

605 mammary epithelial cells (HMEC) in Encyclopedia of DNA Elements at UCSC
606 (http://genome.ucsc.edu/encode/,
607 'wgEncodeCshlLongRnaSeqHmecCellPapAlnRep1.bam' file), H1 derived neural
608 progenitor cell (NPC) (GEO: GSM915326, ENCODE: ENCFF529SIO), Th17
609 (GSM2859479, NCBI SRA: SRR6298326), Treg (GSM2859476, SRR6298323) and
610 GM12878 (GSE78551, ENCFF297QCE) were analyzed. RNA-seq reads were aligned
611 in repeat-masked hg19 genome sequences using BWA with default parameters (Li and
612 Durbin 2009). FPKMs of the RNA-seq data were calculated by RSeQC (Wang et al.
613 2012b). Based on log2-transformed FPKM, UCSC transcripts were arranged in
614 descending order of expression level, and top 4,000 transcripts were selected in each
615 cell type.

616 The expression level of transcriptional target genes predicted based on EPAs
617 shortened at the DNA motif sites of a TF was compared with the expression level of
618 transcriptional target genes predicted based on closed chromatin regions in promoters.
619 For each DNA motif sequence shortening EPAs, transcriptional target genes were
620 predicted based on DNA binding motif sequences of vertebrate TFs in TRANSFAC
621 database, and the ratio of expression level of putative transcriptional target genes of
622 each TF was calculated between an EPA and closed chromatin regions in promoters.
623 The distribution of the ratios of all TFs was compared among forward-reverse (FR),
624 reverse-forward (RF) and non-orientation (i.e. without considering orientation) of a
625 DNA motif sequence shortening an EPA using Mann-Whitney test, two-sided (p-value <
626 10^{-7}). Other parameters for the analysis were set as follows: the number of

627 transcriptional target genes of a TF was ≥50, the number of general TFs (not insulator
628 TFs) to predict transcriptional target genes was ≥50, and the number of DNA motif
629 sites of a TF (insulator TF) shortening an EPA was ≥100.
630
631 **Comparison with chromatin interaction data**
632 For comparison of EPIs with chromatin interactions (HiChIP) in $CD4^+$ T, Th17,
633 Treg and GM12878 cells,
634 'GSM2705049_Naive_HiChIP_H3K27ac_B2T1_allValidPairs.txt',
635 'GSM2705050_Naive_HiChIP_H3K27ac_B2T2_allValidPairs.txt',
636 'GSM2705051_Naive_HiChIP_H3K27ac_B3T1_allValidPairs.txt',
637 'GSM2705054_Th17_HiChIP_H3K27ac_B1T2_allValidPairs.txt',
638 'GSM2705055_Th17_HiChIP_H3K27ac_B2T1_allValidPairs.txt',
639 'GSM2705056_Th17_HiChIP_H3K27ac_B3T1_allValidPairs.txt',
640 'GSM2705059_Treg_HiChIP_H3K27ac_B3T1_allValidPairs.txt',
641 'GSM2705041_GM_HiChIP_H3K27ac_B1_allValidPairs.txt',
642 'GSM2705042_GM_HiChIP_H3K27ac_B2_allValidPairs.txt' files were downloaded
643 from Gene Expression Omnibus (GEO) database. The resolutions of chromatin
644 interactions and EPIs were adjusted to 5 kb before their comparison. Chromatin
645 interactions with ≥6,000 and ≥1,000 counts for each interaction were analyzed in the
646 comparison in $CD4^+$ T cells. According to the unique numbers of chromatin interactions
647 in replications in Th17, Treg and GM12878 cells, chromatin interactions with ≥1,000
648 (B1T2 in Th17), ≥2,000 (B2T1 and B3T1 in Th17), ≥6,000 (B3T1 in Treg), ≥4,000

649 (B1 and B2 in GM12878) counts for each interaction were analyzed in the comparison.

650 Enhancer-promoter interactions (EPIs) were predicted based on three types of
651 EPAs in monocytes: (i) EPA shortened at the FR or RF orientation of DNA motif sites
652 of a TF, (ii) EPA shortened at the non-orientation (i.e. without considering orientation)
653 of DNA motif sites of a TF, and (iii) EPA without being shortened by a DNA motif site.
654 EPIs predicted based on the three types of EPAs in common were removed. EPIs
655 predicted based on EPA (i) were removed from EPIs predicted based on EPA (ii). EPIs
656 predicted based on EPA (i) and (ii) were removed from EPIs predicted based on EPA
657 (iii). The resolution of HiChIP chromatin interaction data was 1-5 kb, so EPIs were
658 adjusted to 5 kb before their comparison. The number and ratio of EPIs overlapped with
659 chromatin interactions were compared two times between EPI (i) and (iii), and EPI (i)
660 and (ii) (binomial distribution, p-value < 0.025 for each test, two-sided, 95% confidence
661 interval).

662 For comparison of EPIs with chromatin interactions (in situ Hi-C) in HMEC, the
663 genomic positions of chromatin interactions in 'HMEC_4DNFI97O9SAZ.pairs.gz' file
664 in 4D nucleome data portal (http://data.4dnucleome.org) were converted into hg19
665 version of human genome using liftOver tool
666 (http://hgdownload.soe.ucsc.edu/downloads.html#liftover). The resolutions of
667 chromatin interaction data and EPIs were adjusted to 3 kb before their comparison. All
668 chromatin interactions were used in the comparison. For comparison of EPIs with
669 chromatin interactions (promoter capture Hi-C) in NPC, the resolutions of chromatin
670 interaction data in 'GSE86189_npc.po.all.txt.bz2' file (GEO: GSE86189) and EPIs were

671 adjusted to 5 kb before their comparison. All chromatin interactions were used in the
672 comparison.

673 Putative target genes for the comparison of EPIs and chromatin interactions were
674 selected from top 4,000 transcripts in term of the expression level. The expression level
675 of putative target genes of EPIs overlapped with chromatin interactions was compared
676 with EPIs not overlapped with them. When the same transcriptional target gene of a TF
677 in an enhancer was predicted based on both an EPI overlapped with a chromatin
678 interaction and an EPI not overlapped with, the target gene was eliminated. The
679 distribution of expression level of putative target genes was compared using
680 Mann-Whitney test, two-sided (p-value < 0.05).

681 Micro-C chromatin interaction data in human fibroblast (4D nucleome data portal
682 ID: 4DNFICOEXGPJ) was converted into hg19 version of human genome using
683 liftOver tool and was adjusted to 200 b resolution. The same genomic locations of pairs
684 of chromatin interaction anchors were merged.

685

686 **Correlation of expression level of gene pairs separated by DNA motif sites of a TF**

687 To examine the correlation of expression level of genes pairs separated by DNA
688 motif sites of a TF, gene expression data among 53 tissues were obtained from
689 'GTEx_Analysis_2016-01-15_v7_RNASeQCv1.1.8_gene_median_tpm.gct.gz' file in
690 GTEx database (https://gtexportal.org/home/). The correlation of log2-transformed gene
691 expression level was analyzed based on protein-coding transcripts of RefSeq (RefSeq
692 ID with 'NM_'). The same genomic location of RefSeq transcripts were removed. DNA

motif sites overlapped with RefSeq transcripts were eliminated. Genes closest to a DNA motif sites and another gene were analyzed using BEDOPS closest-features (Neph et al. 2012). Mann-Whitney test was performed using R.

Acknowledgements

The supercomputing resource was provided by Human Genome Center of the Institute of Medical Science at the University of Tokyo. Publication charges for this article were funded by JSPS KAKENHI Grant Number 16K00387. This work was partially supported by JST CREST Grant Number JPMJCR15G1, Japan.

703 **Figures**

704 Web links to original figures.

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

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

707 76_2014_BFnrg3682_MOESM18_ESM.ppt

708 Graphical Abstract (de Wit et al. 2015)

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

710 92c5eba1d/fx1.jpg

711

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

713 2010)

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

715 87_2010_BFnbt1630_MOESM11_ESM.pdf

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

717 DNA Looping (Guo et al. 2015)

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

719 2af081251/gr4.jpg

720

721 **Figure 1.** Chromatin interaction and enhancer-promoter association. (A) Forward–

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

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

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

725 2014; de Wit et al. 2015; Guo et al. 2015). (B) Computationally-defined regulatory
726 domains for enhancer-promoter association (McLean et al. 2010). The *single nearest*
727 *gene* association rule extends the regulatory domain to the midpoint between this gene's
728 TSS and the nearest gene's TSS both upstream and downstream. The enhancer-promoter
729 association was shortened at the forward-reverse orientation of DNA binding motif sites
730 of a TF in this study (e.g. CTCF in the figure).
731

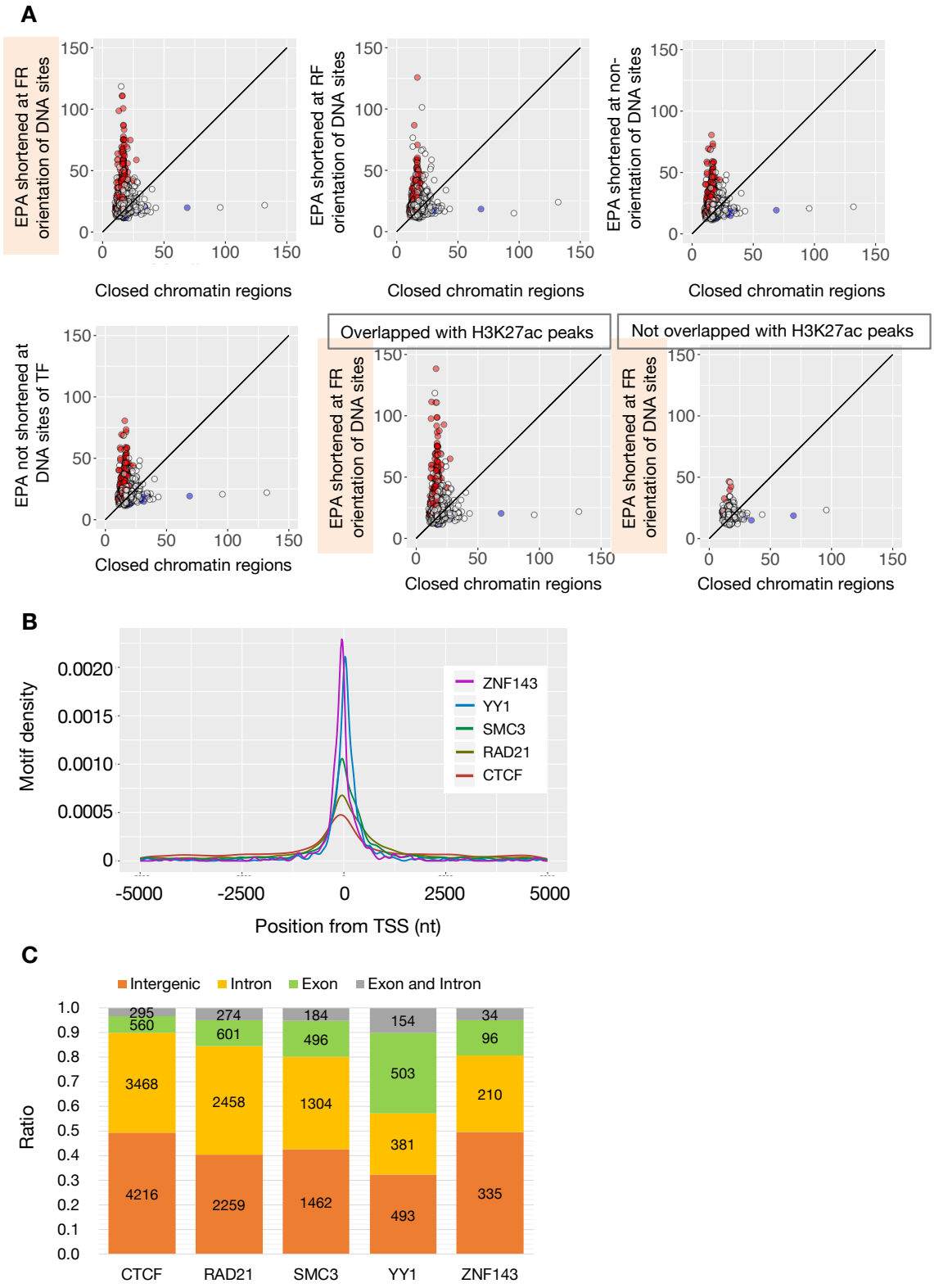


Figure 2. Identification of DNA motif sites of TFs in open chromatin regions and properties of DNA motif sites of TFs. (A) Comparison of expression level of putative

735 transcriptional target genes. Each dot represents the median expression level (FPKM) of
736 target genes predicted based on DNA motif sites of a transcription factor in closed
737 chromatin regions in promoters (X-axis) and an enhancer-promoter association (EPA)
738 shortened at the forward-reverse orientation of DNA motif sites of CTCF (Y-axis) in T
739 cell (upper left graph). The distribution of expression level was changed according to
740 four criteria of EPAs (four graphs) (Osato 2018). Red dots indicate the median
741 expression level of target genes was significantly higher in an EPA than closed
742 chromatin regions, and blue dots indicate the median expression level of target genes
743 was significantly lower in an EPA than closed chromatin regions (Mann-Whitney test).
744 The expression level of target genes predicted based on an EPA tend to be higher than
745 closed chromatin regions, implying that TFs serve as activators of target genes. DNA
746 motif sites of TFs overlapped with H3K27ac ChIP-seq peaks were associated with a
747 significant difference of expression level of target genes, and DNA motif sites not
748 overlapped with H3K27ac peaks were not associated with the difference (lower right
749 two graphs). This suggests that the significant difference of expression level is
750 associated with enhancer (and promoter) signals of H3K27ac. (B) Distribution of DNA
751 motif sites of TFs near TSS. ZNF143 and YY1 tend to be located near TSS, and CTCF
752 tend to be observed more in distant regions than around TSS in monocyte. (C)
753 Distribution of DNA motif sites of TFs in the human genome in monocytes. Most of
754 DNA motif sites were found in intergenic and intron regions. YY1 is known to be at
755 TSS in the edge of exons of genes, and is involved in transcriptional activation,
756 repression and initiation.

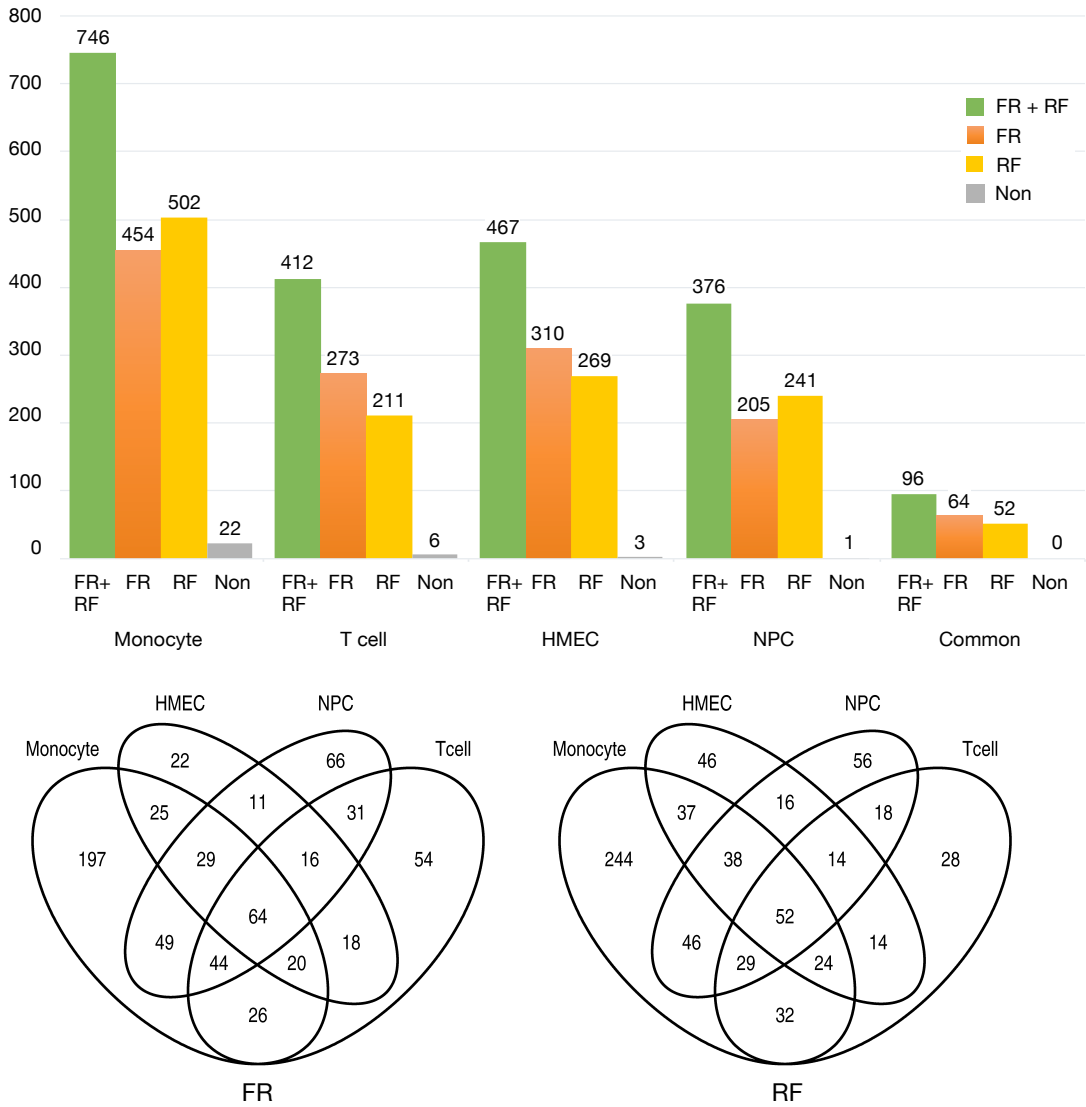
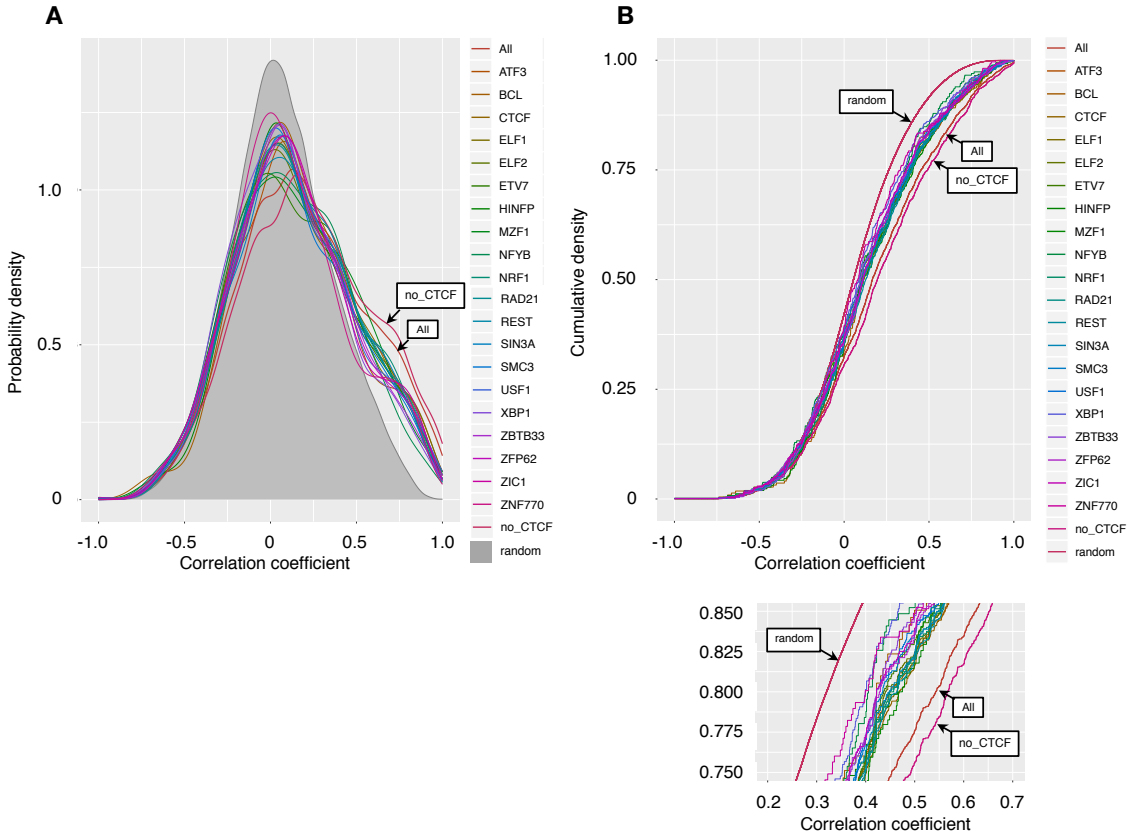


Figure 3. Biased orientations of DNA motif sequences of TFs affect the transcription of genes by the putative insulator function of the DNA motif sites. Total 96 (64 FR and 52 RF) of biased orientations of DNA binding motif sequences of TFs affected significantly the expression level of putative transcriptional target genes in monocytes, T cells, HMEC and NPC in common, whereas non-orientation (i.e. without considering

764 orientation) of DNA binding motif sequence of a TF did not affect the expression level

765 significantly.

766



767

768 **Figure 4.** Genes separated by predicted DNA binding sites of biased orientations of

769 DNA motifs of TFs are less correlated in gene expression. Top 20 TFs are selected

770 according to ascending order of median correlation coefficient. Correlation coefficient

771 between neighboring gene pairs is shown in terms of probability density (A) and

772 cumulative distribution (B). Brown line (All), correlation between all neighboring

773 genes; gray shading in (A) and red line (random) in (B), correlation between randomly

774 chosen gene pairs; Purple line (no_CTCF), correlation between gene pairs not separated

775 by predicted CTCF sites.

776

777 Web links to original figures.

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

779 (Bailey et al. 2015).

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

781

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

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

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

785 2015).

786

787

Tables

Table 1. Biased orientations of DNA binding motif sequences of TFs in monocytes, T cells, HMEC and NPC in common. TF: DNA binding motif sequences of transcription factors.

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

796 **Table 2.** Comparison of biased orientation of a TF and CTCF in terms of expression
797 level of putative transcriptional target genes. The merger of DNA motif sites of a TF
798 and CTCF showed a higher expression level of putative transcriptional target genes than
799 DNA motif sites of only CTCF. In each cell type, a rank was divided by the lowest rank
800 for normalization. Rank total: the sum of the normalized ranks. Known TFs associated
801 with chromatin interactions are indicated in bold.

Rank	TF	Rank total	monocyte	T cell	HMEC	NPC
1	**YY1**	0.58	0.28	0.11	0.11	0.08
2	NR3C1	0.60	0.06	0.06	0.33	0.15
3	MYC	0.93	0.06	0.39	0.33	0.15
4	YY2	1.08	0.56	0.33	0.11	0.08
5	TATA	1.19	0.78	0.22	0.11	0.08
6	MAX	1.32	0.06	0.28	0.22	0.77
7	MZF1	1.56	0.33	0.39	0.22	0.62
8	SREBF2	1.64	0.39	0.33	0.22	0.69
9	ATF3	1.71	0.06	0.67	0.22	0.77
10	CREB3L1	1.74	0.28	0.89	0.11	0.46
11	ETS1	1.78	0.50	0.33	0.33	0.62
12	ETS	1.84	0.28	0.83	0.11	0.62
13	RUNX3	1.99	0.28	0.89	0.67	0.15
14	PLAG1	2.00	0.17	0.72	0.11	1.00
15	TP73	2.12	0.44	1.00	0.44	0.23
16	MYB	2.14	0.17	0.83	0.44	0.69
17	USF1	2.15	0.94	0.83	0.22	0.15
18	ARNT	2.16	0.39	0.94	0.44	0.38
19	**CTCF**	2.17	0.33	0.78	0.44	0.62
20	SP2	2.25	0.83	0.89	0.22	0.31
21	EGR2	2.26	0.94	0.17	0.22	0.92
22	ZBTB33	2.28	0.06	1.00	0.22	1.00
23	YBX1	2.32	0.56	0.78	0.22	0.77
24	XBP1	2.37	0.94	0.56	0.33	0.54
25	PBX3	2.40	0.50	0.89	0.78	0.23
26	**ZNF143**	2.44	0.39	0.94	0.33	0.77
27	NFYB	2.45	0.33	0.72	0.78	0.62
28	TFAP2C	2.48	0.28	0.94	0.33	0.92
28	NFY	2.48	0.72	0.78	0.44	0.54
29	ZBTB7A	2.56	0.56	0.67	0.33	1.00
30	FOS	2.62	1.00	0.89	0.11	0.62
31	**RAD21**	2.66	0.50	0.61	0.78	0.77
31	AHR	2.66	1.00	0.94	0.33	0.38
32	KLF4	2.69	0.72	0.94	0.33	0.69
33	**SMC3**	2.77	0.56	0.89	0.56	0.77
34	MITF	2.83	0.67	0.72	0.44	1.00
35	ELK3	2.89	0.72	0.83	0.33	1.00
35	NONO	2.89	0.89	0.78	0.22	1.00
36	ERG	2.94	0.78	0.83	0.33	1.00
37	FLI1	3.00	0.94	0.83	0.22	1.00
38	ZNF770	3.07	0.72	0.83	0.67	0.85
39	E2F	3.15	1.00	0.89	0.33	0.92
40	HIC1	3.17	0.89	0.94	0.33	1.00
41	BHLHE41	3.33	0.94	0.94	0.44	1.00
41	PLAGL1	3.33	0.56	1.00	0.78	1.00

802

803

804

805 **Table 3.** Comparison of enhancer-promoter interactions (EPIs) with chromatin
806 interactions in T cells. The upper tables show the number of biased orientations of DNA
807 motifs, where a significantly higher ratio of EPIs that were predicted based on an
808 enhancer-promoter association (EPA) (i) overlapped with HiChIP chromatin
809 interactions than the other types of EPA (ii) and (iii) in T cells. The middle tables show

133 FR and 90 RF orientations of DNA motifs found in common among B2T1, B2T2 and B3T1 replications in T cells in the upper table. The lower tables show that among 64 FR and 52 RF orientations of DNA motifs found in common in monocytes, T cell, HMEC and NPC, 44 (69%) FR and 28 (54%) RF were matched with the analysis of HiChIP chromatin interactions according to three types of EPAs. TF: DNA binding motif sequence of a transcription factor. Score: $-\log_{10}$ (p-value). Score 1: Comparison of an EPA shortened at the FR or RF orientation of DNA motif sites [EPA (i)] with an EPA not shortened [EPA (iii)]. Score 2: Comparison of an EPA shortened at the FR or RF orientation of DNA motif sites [EPA (i)] with an EPA shortened at the non-orientation of DNA motif sites [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

825

826

827 FR and RF orientations of DNA motifs found in common among three replications in T

828 cell.

829 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

830

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

FR and RF orientations of DNA motifs found in common in the four cell types.

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 4. Correlation of expression level of gene pairs separated by DNA motif sites of a TF. Pairs of genes separated by 64 FR and 52 RF orientations of DNA motif sites of TFs 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 nearby genes among 53 tissues. Score: $-\log_{10}$ (p-value) of Mann-Whitney test. Median correlation: median correlation coefficient of expression level of nearby gene pairs separated by DNA motif sites. The median correlation coefficient of expression level of all pairs of nearby genes is 0.17. # DNA site: number of DNA motif sites of a

849 TF separating gene pairs with expression data. All DNA motif sites of a TF were used,

850 except for DNA motif sites separating pairs of ubiquitously expressed genes with

851 coefficient of variance <90.

852

853 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

854

855 **Table 5.** Top 30 of co-location and overlapping biased orientations of DNA binding
856 motif sequences of TFs in monocytes. Co-location and overlapping DNA motif
857 sequences of CTCF with another biased orientation of DNA motif sequence are shown
858 in separate tables. Motif 1,2: DNA binding motif sequences of TFs. No.: the number of
859 genomic locations with both Motif 1 and Motif 2 in upstream and downstream of genes
860 in an EPA (< 1 Mb distance from TSS of genes). Total number of genes is 30,523.

861

862 Forward-reverse (FR)

Motif 1	Motif 2	No.	Mean of distances bet. DNA sites (nt)	Median of distances bet. DNA sites (nt)	Mean of distances from TSS (nt)	Median of distances from TSS (nt)	Interquartile range of distance from TSS (nt)
FOS	JUN	3332	0.00	0	43202.95	18085	5524 - 48583
CTCF	**SMC3**	2876	0.01	0	34606.92	13716	3073 - 39471
FOSL2	JUN	2807	0.00	0	40389.84	17365	5424 - 45251
FOSB	FOSL2	2644	0.00	0	44302.93	17867	5633 - 47209
EGR1	EGR2	2397	0.00	0	14578.21	1251	247 - 11473
CTCF	**RAD21**	2352	0.00	0	33492.68	12913	2824 - 37672
RAD21	**SMC3**	2343	0.02	0	33572.80	14302	3856 - 37159
ATF3	FOSL2	2322	0.00	0	42784.04	17841	5366 - 48236
FOSL2	JUNB	2305	0.02	0	42884.36	17867	5706 - 46515
ETS2	ETSLIKE	2205	0.00	0	37660.66	15765	4146 - 41462
FOS	JUNB	2151	0.00	0	38403.20	16367	4849 - 44736
IRF	IRF4	1962	0.00	0	40794.41	15872	3969 - 45946
FOSL2	SMARCC1	1945	0.00	0	42429.69	17369	5094 - 46586
EGR1	WT1	1941	0.16	0	12485.23	427	93 - 5394
FOSB	JUN	1808	0.02	0	36648.74	16705	5156 - 40309
ATF3	JUN	1713	0.02	0	38101.24	17119	5294 - 43559
JUN	TFAP2A	1628	0.00	0	42262.97	18017	5089 - 47094
FOSL2	TFAP2A	1628	0.00	0	36606.38	16201	4949 - 41948
BACH2	SMARCC1	1462	0.16	0	38673.02	14778	4060 - 41262
ELF1	ETS1	1459	0.00	0	37056.72	14346	3370 - 40339
CHD2	E2F	1429	0.00	0	19220.38	1579	329 - 16257
JUN	NR3C1	1262	0.00	0	35052.17	14418	3844 - 38534
USF1	USF2	1260	0.00	0	25188.91	4726	413 - 24699
BACH2	JUNB	1238	0.15	0	36252.37	15302	4462 - 39393
BACH2	FOSB	1184	0.22	0	37190.80	14778	4038 - 41262
ETSLIKE	GABPA	1164	0.00	0	32352.24	11683	2356 - 36919
CTCF	CTCFL	1151	0.00	0	30252.04	11606	2229 - 34539
NR5A1	NR5A2	1133	0.00	0	41171.86	12243	2488 - 43136
ATF3	BACH2	1094	0.18	0	35659.16	13605	3821 - 39756
ETS2	GABPA	1092	0.06	0	27250.24	12184	2858 - 32196

863

864

865

866

Reverse-forward (RF)

Motif 1	Motif 2	No.	Mean of distances bet. DNA sites (nt)	Median of distances bet. DNA sites (nt)	Mean of distances from TSS (nt)	Median of distances from TSS (nt)	Interquartile range of distance from TSS (nt)
EGR1	EGR3	3012	0.00	0	17005.00	709	149 - 10393
NR1I2	NR1I3	2932	0.00	0	49622.84	20548	5924 - 58436
ZIC1	ZIC3	2871	0.07	0	21511.64	3092	434 - 19128
EGR	WT1	2612	0.04	0	9949.38	396	102 - 4266
FOS	FOS:JUN	2512	0.01	0	37200.80	15466	4438 - 42373
EGR	EGR3	2293	0.17	0	11634.92	532	135 - 6044
MAFG	NFE2L1:MAFG	2272	0.03	0	34641.56	16917	5284 - 42097
KLF10	SP1	2233	0.00	0	9778.94	297	73 - 3077
EGR	EGR2	2160	0.16	0	10502.76	490	129 - 5227
KLF3	SP3	2036	0.00	0	6974.78	246	63 - 1447
KLF1	SP3	1771	0.00	0	7748.42	264	66 - 1829
KLF10	SP3	1734	0.00	0	7275.48	222	58 - 1320
KLF7	SP4	1639	0.00	0	12521.62	309	58 - 4894
KLF1	SP1	1574	0.05	0	6780.46	245	61 - 1438
AP1	JUNB	1572	0.05	0	34158.07	15451	4746 - 40315
EGR1	WT1	1543	0.31	0	12742.28	444	111 - 5553
EGR1	SP1:SP3	1393	0.08	0	7789.54	307	91 - 1886
FOS:JUN	JUNB	1386	0.14	0	33012.80	14245	3900 - 37988
SP1	SP3	1372	0.00	0	8845.31	293	77 - 1816
EHF	ELF1	1366	0.02	0	41217.39	13357	2842 - 43178
CTCF	RXRA	1362	0.00	0	29981.33	10937	2100 - 31407
AP1	FOS	1327	0.03	0	33225.52	13455	3503 - 37399
ATF1	ATF3	1215	0.00	0	23125.06	3798	342 - 21091
ELK1	ZNF200	1156	0.05	0	22200.49	2883	181 - 17610
SP1	SP4	1137	0.00	0	7371.85	215	63 - 1232
EHF	ELF3	1134	0.00	0	36296.46	10912	596 - 37318
MEIS2	TGIF1	1018	0.00	0	47677.15	17654	3273 - 49233
KLF3	SP1	1017	0.06	0	6514.33	226	60 - 994
SPI1	ZNF200	913	0.06	0	10962.81	425	85 - 6046
AP1	FOS:JUN	860	0.09	0	50412.92	24217	7302 - 60018

Co-location and overlapping DNA motifs of CTCF with another DNA motifs.

FR orientation of CTCF

Motif 1	Motif 2	No.	Mean of distances bet. DNA sites (nt)	Median of distances bet. DNA sites (nt)	Mean of distances from TSS (nt)	Median of distances from TSS (nt)	Interquartile range of distance from TSS (nt)
CTCF	SMC3	2876	0.01	0	34606.92	13716	3073 - 39471
CTCF	RAD21	2352	0.00	0	33492.68	12913	2824 - 37672
CTCF	CTCFL	1151	0.00	0	30252.04	11606	2229 - 34539
CTCF	ZBTB7A	470	0.27	0	23227.96	7065	745 - 23496
CTCF	E2F	123	0.81	0	7913.36	222	61 - 1997
CTCF	WT1	120	0.00	0	3543.72	114	44 - 375
CTCF	MYC	110	2.30	0	31289.45	11948	2887 - 33972

RF orientation of CTCF

Motif 1	Motif 2	No.	Mean of distances bet. DNA sites (nt)	Median of distances bet. DNA sites (nt)	Mean of distances from TSS (nt)	Median of distances from TSS (nt)	Interquartile range of distance from TSS (nt)
CTCF	RXRA	1362	0.00	0	29981.33	10937	2100 - 31407
CTCF	CTCFL	429	0.00	0	25171.22	9834	1859 - 24741
CTCF	RAD21	337	0.00	0	38829.91	17973	6117 - 45749
CTCF	ZIC1	126	1.40	0	35480.65	10286	2316 - 29380
CTCF	ZIC3	108	0.00	0	30219.32	10904	4463 - 31037

876

877 **References**

878

879 Ackerman SL, Minden AG, Williams GT, Bobonis C, Yeung CY. 1991. Functional
880 significance of an overlapping consensus binding motif for Sp1 and Zif268 in
881 the murine adenosine deaminase gene promoter. *Proc Natl Acad Sci U S A* **88**:
882 7523-7527.
883 Ansari AZ, Peterson-Kaufman KJ. 2011. A partner evokes latent differences between
884 Hox proteins. *Cell* **147**: 1220-1221.
885 Bailey SD, Zhang X, Desai K, Aid M, Corradin O, Cowper-Sal Lari R, Akhtar-Zaidi B,
886 Scacheri PC, Haibe-Kains B, Lupien M. 2015. ZNF143 provides sequence
887 specificity to secure chromatin interactions at gene promoters. *Nat Commun* **2**:
888 6186.
889 Barutcu AR, Lajoie BR, Fritz AJ, McCord RP, Nickerson JA, van Wijnen AJ, Lian JB,
890 Stein JL, Dekker J, Stein GS et al. 2016. SMARCA4 regulates gene expression
891 and higher-order chromatin structure in proliferating mammary epithelial cells.
892 *Genome research* **26**: 1188-1201.
893 Blais A, Dynlacht BD. 2007. E2F-associated chromatin modifiers and cell cycle control.
894 *Curr Opin Cell Biol* **19**: 658-662.
895 Blattler A, Yao L, Wang Y, Ye Z, Jin VX, Farnham PJ. 2013. ZBTB33 binds
896 unmethylated regions of the genome associated with actively expressed genes.
897 *Epigenetics Chromatin* **6**: 13.
898 Chatterjee R, Zhao J, He X, Shlyakhtenko A, Mann I, Waterfall JJ, Meltzer P,
899 Sathyanarayana BK, FitzGerald PC, Vinson C. 2012. Overlapping ETS and CRE
900 Motifs ((G/C)CGGAAGTGACGTCA) preferentially bound by GABPalpha and
901 CREB proteins. *G3 (Bethesda)* **2**: 1243-1256.
902 Chen M, Manley JL. 2009. Mechanisms of alternative splicing regulation: insights from
903 molecular and genomics approaches. *Nat Rev Mol Cell Biol* **10**: 741-754.
904 Das D, Clark TA, Schweitzer A, Yamamoto M, Marr H, Arribere J, Minovitsky S,
905 Poliakov A, Dubchak I, Blume JE et al. 2007. A correlation with exon
906 expression approach to identify cis-regulatory elements for tissue-specific
907 alternative splicing. *Nucleic acids research* **35**: 4845-4857.

908 de Wit E, Vos ES, Holwerda SJ, Valdes-Quezada C, Verstegen MJ, Teunissen H,
909 Splinter E, Wijchers PJ, Krijger PH, de Laat W. 2015. CTCF Binding Polarity
910 Determines Chromatin Looping. *Molecular cell* **60**: 676-684.
911 Di Giammartino DC, Kloetgen A, Polyzos A, Liu Y, Kim D, Murphy D, Abuhashem A,
912 Cavaliere P, Aronson B, Shah V et al. 2019. KLF4 is involved in the
913 organization and regulation of pluripotency-associated three-dimensional
914 enhancer networks. *Nat Cell Biol* **21**: 1179-1190.
915 Frob F, Wegner M. 2020. The role of chromatin remodeling complexes in Schwann cell
916 development. *Glia* **68**: 1596-1603.
917 Guo Y, Xu Q, Canzio D, Shou J, Li J, Gorkin DU, Jung I, Wu H, Zhai Y, Tang Y et al.
918 2015. CRISPR Inversion of CTCF Sites Alters Genome Topology and
919 Enhancer/Promoter Function. *Cell* **162**: 900-910.
920 He X, Syed KS, Tillo D, Mann I, Weirauch MT, Vinson C. 2015. GABPalpha Binding
921 to Overlapping ETS and CRE DNA Motifs Is Enhanced by CREB1: Custom
922 DNA Microarrays. *G3 (Bethesda)* **5**: 1909-1918.
923 Hsu SC, Gilgenast TG, Bartman CR, Edwards CR, Stonestrom AJ, Huang P, Emerson
924 DJ, Evans P, Werner MT, Keller CA et al. 2017. The BET Protein BRD2
925 Cooperates with CTCF to Enforce Transcriptional and Architectural Boundaries.
926 *Molecular cell* **66**: 102-116 e107.
927 Huang S, Li X, Yusufzai TM, Qiu Y, Felsenfeld G. 2007. USF1 recruits histone
928 modification complexes and is critical for maintenance of a chromatin barrier.
929 *Mol Cell Biol* **27**: 7991-8002.
930 Jolma A, Yan J, Whitington T, Toivonen J, Nitta KR, Rastas P, Morgunova E, Enge M,
931 Taipale M, Wei G et al. 2013. DNA-binding specificities of human transcription
932 factors. *Cell* **152**: 327-339.
933 Jolma A, Yin Y, Nitta KR, Dave K, Popov A, Taipale M, Enge M, Kivioja T, Morgunova
934 E, Taipale J. 2015. DNA-dependent formation of transcription factor pairs alters
935 their binding specificity. *Nature* **527**: 384-388.
936 Jung I, Schmitt A, Diao Y, Lee AJ, Liu T, Yang D, Tan C, Eom J, Chan M, Chee S et al.
937 2019. A compendium of promoter-centered long-range chromatin interactions in
938 the human genome. *Nature genetics* **51**: 1442-1449.
939 Karolchik D, Barber GP, Casper J, Clawson H, Cline MS, Diekhans M, Dreszer TR,
940 Fujita PA, Guruvadoo L, Haeussler M et al. 2014. The UCSC Genome Browser

941 database: 2014 update. *Nucleic acids research* **42**: D764-770.
942 Katainen R, Dave K, Pitkanen E, Palin K, Kivioja T, Valimaki N, Gylfe AE, Ristolainen
943 H, Hanninen UA, Cajuso T et al. 2015. CTCF/cohesin-binding sites are
944 frequently mutated in cancer. *Nature genetics* **47**: 818-821.
945 Kel AE, Gossling E, Reuter I, Cheremushkin E, Kel-Margoulis OV, Wingender E. 2003.
946 MATCH: A tool for searching transcription factor binding sites in DNA
947 sequences. *Nucleic acids research* **31**: 3576-3579.
948 Kheradpour P, Kellis M. 2014. Systematic discovery and characterization of regulatory
949 motifs in ENCODE TF binding experiments. *Nucleic acids research* **42**:
950 2976-2987.
951 Kim DE, Procopio MG, Ghosh S, Jo SH, Goruppi S, Magliozzi F, Bordignon P, Neel V,
952 Angelino P, Dotto GP. 2017. Convergent roles of ATF3 and CSL in chromatin
953 control of cancer-associated fibroblast activation. *J Exp Med* **214**: 2349-2368.
954 Kin K, Chen X, Gonzalez-Garay M, Fakhouri WD. 2016. The effect of non-coding
955 DNA variations on P53 and cMYC competitive inhibition at cis-overlapping
956 motifs. *Hum Mol Genet* **25**: 1517-1527.
957 Knoepfler PS, Zhang XY, Cheng PF, Gafken PR, McMahon SB, Eisenman RN. 2006.
958 Myc influences global chromatin structure. *EMBO J* **25**: 2723-2734.
959 Krietenstein N, Abraham S, Venev SV, Abdennur N, Gibcus J, Hsieh TS, Parsi KM,
960 Yang L, Maehr R, Mirny LA et al. 2020. Ultrastructural Details of Mammalian
961 Chromosome Architecture. *Molecular cell* **78**: 554-565 e557.
962 Kulakovskiy IV, Vorontsov IE, Yevshin IS, Sharipov RN, Fedorova AD, Rumynskiy EI,
963 Medvedeva YA, Magana-Mora A, Bajic VB, Papatsenko DA et al. 2018.
964 HOCOMOCO: towards a complete collection of transcription factor binding
965 models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic acids*
966 *research* **46**: D252-D259.
967 Laurette P, Strub T, Koludrovic D, Keime C, Le Gras S, Seberg H, Van Otterloo E,
968 Imrichova H, Siddaway R, Aerts S et al. 2015. Transcription factor MITF and
969 remodeller BRG1 define chromatin organisation at regulatory elements in
970 melanoma cells. *eLife* **4**.
971 Lee JW, Kim DM, Jang JW, Park TG, Song SH, Lee YS, Chi XZ, Park IY, Hyun JW, Ito
972 Y et al. 2019. RUNX3 regulates cell cycle-dependent chromatin dynamics by
973 functioning as a pioneer factor of the restriction-point. *Nat Commun* **10**: 1897.

974 Levy C, Fisher DE. 2011. Dual roles of lineage restricted transcription factors: the case
975 of MITF in melanocytes. *Transcription* **2**: 19-22.
976 Li F, Yuan Q, Di W, Xia X, Liu Z, Mao N, Li L, Li C, He J, Li Y et al. 2020. ERG
977 orchestrates chromatin interactions to drive prostate cell fate reprogramming. *J*
978 *Clin Invest* doi:10.1172/JCI137967.
979 Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler
980 transform. *Bioinformatics (Oxford, England)* **25**: 1754-1760.
981 Li Y, Schulz VP, Deng C, Li G, Shen Y, Tusi BK, Ma G, Stees J, Qiu Y, Steiner LA et al.
982 2016. Setd1a and NURF mediate chromatin dynamics and gene regulation
983 during erythroid lineage commitment and differentiation. *Nucleic acids research*
984 **44**: 7173-7188.
985 Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A,
986 Amit I, Lajoie BR, Sabo PJ, Dorschner MO et al. 2009. Comprehensive
987 mapping of long-range interactions reveals folding principles of the human
988 genome. *Science (New York, NY)* **326**: 289-293.
989 Ma C, Karwacki-Neisius V, Tang H, Li W, Shi Z, Hu H, Xu W, Wang Z, Kong L, Lv R
990 et al. 2016. Nono, a Bivalent Domain Factor, Regulates Erk Signaling and
991 Mouse Embryonic Stem Cell Pluripotency. *Cell Rep* **17**: 997-1007.
992 Mager GM, Ward RM, Srinivasan R, Jang SW, Wrabetz L, Svaren J. 2008. Active gene
993 repression by the Egr2.NAB complex during peripheral nerve myelination. *J*
994 *Biol Chem* **283**: 18187-18197.
995 McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM,
996 Bejerano G. 2010. GREAT improves functional interpretation of cis-regulatory
997 regions. *Nature biotechnology* **28**: 495-501.
998 Merabet S, Mann RS. 2016. To Be Specific or Not: The Critical Relationship Between
999 Hox And TALE Proteins. *Trends Genet* **32**: 334-347.
1000 Mumbach MR, Satpathy AT, Boyle EA, Dai C, Gowen BG, Cho SW, Nguyen ML,
1001 Rubin AJ, Granja JM, Kazane KR et al. 2017. Enhancer connectome in primary
1002 human cells identifies target genes of disease-associated DNA elements. *Nature*
1003 *genetics* **49**: 1602-1612.
1004 Neph S, Kuehn MS, Reynolds AP, Haugen E, Thurman RE, Johnson AK, Rynes E,
1005 Maurano MT, Vierstra J, Thomas S et al. 2012. BEDOPS: high-performance
1006 genomic feature operations. *Bioinformatics (Oxford, England)* **28**: 1919-1920.

1007 Newburger DE, Bulyk ML. 2009. UniPROBE: an online database of protein binding
1008 microarray data on protein-DNA interactions. *Nucleic acids research* **37**:
1009 D77-82.
1010 Oldfield AJ, Yang P, Conway AE, Cinghu S, Freudenberg JM, Yellaboina S, Jothi R.
1011 2014. Histone-fold domain protein NF-Y promotes chromatin accessibility for
1012 cell type-specific master transcription factors. *Molecular cell* **55**: 708-722.
1013 Osato N. 2018. Characteristics of functional enrichment and gene expression level of
1014 human putative transcriptional target genes. *BMC Genomics* **19**: 957.
1015 Portales-Casamar E, Thongjuea S, Kwon AT, Arenillas D, Zhao X, Valen E, Yusuf D,
1016 Lenhard B, Wasserman WW, Sandelin A. 2010. JASPAR 2010: the greatly
1017 expanded open-access database of transcription factor binding profiles. *Nucleic*
1018 *acids research* **38**: D105-110.
1019 Purmann A, Toedling J, Schueler M, Carninci P, Lehrach H, Hayashizaki Y, Huber W,
1020 Sperling S. 2007. Genomic organization of transcriptomes in mammals:
1021 Coregulation and cofunctionality. *Genomics* **89**: 580-587.
1022 Ramos Pittol JM, Oruba A, Mittler G, Saccani S, van Essen D. 2018. Zbtb7a is a
1023 transducer for the control of promoter accessibility by NF-kappa B and multiple
1024 other transcription factors. *PLoS Biol* **16**: e2004526.
1025 Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn
1026 AL, Machol I, Omer AD, Lander ES et al. 2014. A 3D map of the human
1027 genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**:
1028 1665-1680.
1029 Riggi N, Knoechel B, Gillespie SM, Rheinbay E, Boulay G, Suva ML, Rossetti NE,
1030 Boonseng WE, Oksuz O, Cook EB et al. 2014. EWS-FLI1 utilizes divergent
1031 chromatin remodeling mechanisms to directly activate or repress enhancer
1032 elements in Ewing sarcoma. *Cancer Cell* **26**: 668-681.
1033 Sabo A, Amati B. 2014. Genome recognition by MYC. *Cold Spring Harb Perspect Med*
1034 **4**.
1035 Schreiber J, Libbrecht M, Bilmes J, Noble W. 2017. Nucleotide sequence and DNaseI
1036 sensitivity are predictive of 3D chromatin architecture. *bioRxiv*.
1037 Sherwood RI, Hashimoto T, O'Donnell CW, Lewis S, Barkal AA, van Hoff JP, Karun V,
1038 Jaakkola T, Gifford DK. 2014. Discovery of directional and nondirectional
1039 pioneer transcription factors by modeling DNase profile magnitude and shape.

1040 *Nature biotechnology* **32**: 171-178.
1041 Shlyueva D, Stampfel G, Stark A. 2014. Transcriptional enhancers: from properties to
1042 genome-wide predictions. *Nature reviews Genetics* **15**: 272-286.
1043 Shrivastava A, Calame K. 1994. An analysis of genes regulated by the multi-functional
1044 transcriptional regulator Yin Yang-1. *Nucleic acids research* **22**: 5151-5155.
1045 Slattery M, Riley T, Liu P, Abe N, Gomez-Alcala P, Dror I, Zhou T, Rohs R, Honig B,
1046 Bussemaker HJ et al. 2011. Cofactor binding evokes latent differences in DNA
1047 binding specificity between Hox proteins. *Cell* **147**: 1270-1282.
1048 Subtil-Rodriguez A, Vazquez-Chavez E, Ceballos-Chavez M, Rodriguez-Paredes M,
1049 Martin-Subero JI, Esteller M, Reyes JC. 2014. The chromatin remodeller CHD8
1050 is required for E2F-dependent transcription activation of S-phase genes. *Nucleic*
1051 *acids research* **42**: 2185-2196.
1052 Swanson HI. 2002. DNA binding and protein interactions of the AHR/ARNT
1053 heterodimer that facilitate gene activation. *Chem Biol Interact* **141**: 63-76.
1054 Tabuchi TM, Deplancke B, Osato N, Zhu LJ, Barrasa MI, Harrison MM, Horvitz HR,
1055 Walhout AJ, Hagstrom KA. 2011. Chromosome-biased binding and gene
1056 regulation by the Caenorhabditis elegans DRM complex. *PLoS Genet* **7**:
1057 e1002074.
1058 Taubert S, Gorrini C, Frank SR, Parisi T, Fuchs M, Chan HM, Livingston DM, Amati B.
1059 2004. E2F-dependent histone acetylation and recruitment of the Tip60
1060 acetyltransferase complex to chromatin in late G1. *Mol Cell Biol* **24**: 4546-4556.
1061 van Helden J. 2003. Regulatory sequence analysis tools. *Nucleic acids research* **31**:
1062 3593-3596.
1063 Volkel S, Stielow B, Finkernagel F, Stiewe T, Nist A, Suske G. 2015. Zinc finger
1064 independent genome-wide binding of Sp2 potentiates recruitment of histone-fold
1065 protein Nf-y distinguishing it from Sp1 and Sp3. *PLoS Genet* **11**: e1005102.
1066 Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF,
1067 Schroth GP, Burge CB. 2008. Alternative isoform regulation in human tissue
1068 transcriptomes. *Nature* **456**: 470-476.
1069 Wang J, Vicente-Garcia C, Seruggia D, Molto E, Fernandez-Minan A, Neto A, Lee E,
1070 Gomez-Skarmeta JL, Montoliu L, Lunyak VV et al. 2015. MIR retrotransposon
1071 sequences provide insulators to the human genome. *Proc Natl Acad Sci U S A*
1072 **112**: E4428-4437.

1073 Wang J, Zhuang J, Iyer S, Lin X, Whitfield TW, Greven MC, Pierce BG, Dong X,
1074 Kundaje A, Cheng Y et al. 2012a. Sequence features and chromatin structure
1075 around the genomic regions bound by 119 human transcription factors. *Genome*
1076 *research* **22**: 1798-1812.
1077 Wang L, Wang S, Li W. 2012b. RSeQC: quality control of RNA-seq experiments.
1078 *Bioinformatics (Oxford, England)* **28**: 2184-2185.
1079 Wei Z, Gao F, Kim S, Yang H, Lyu J, An W, Wang K, Lu W. 2013. Klf4 organizes
1080 long-range chromosomal interactions with the oct4 locus in reprogramming and
1081 pluripotency. *Cell Stem Cell* **13**: 36-47.
1082 Weintraub AS, Li CH, Zamudio AV, Sigova AA, Hannett NM, Day DS, Abraham BJ,
1083 Cohen MA, Nabet B, Buckley DL et al. 2017. YY1 Is a Structural Regulator of
1084 Enhancer-Promoter Loops. *Cell* **171**: 1573-1588 e1528.
1085 Wen Z, Huang ZT, Zhang R, Peng C. 2018. ZNF143 is a regulator of chromatin loop.
1086 *Cell Biol Toxicol* **34**: 471-478.
1087 Wingender E, Dietze P, Karas H, Knuppel R. 1996. TRANSFAC: a database on
1088 transcription factors and their DNA binding sites. *Nucleic acids research* **24**:
1089 238-241.
1090 Wust HM, Wegener A, Frob F, Hartwig AC, Wegwitz F, Kari V, Schimmel M, Tamm
1091 ER, Johnsen SA, Wegner M et al. 2020. Egr2-guided histone H2B
1092 monoubiquitination is required for peripheral nervous system myelination.
1093 *Nucleic acids research* doi:10.1093/nar/gkaa606.
1094 Xi H, Yu Y, Fu Y, Foley J, Halees A, Weng Z. 2007. Analysis of overrepresented motifs
1095 in human core promoters reveals dual regulatory roles of YY1. *Genome research*
1096 **17**: 798-806.
1097 Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, Lander ES, Kellis M.
1098 2005. Systematic discovery of regulatory motifs in human promoters and 3'
1099 UTRs by comparison of several mammals. *Nature* **434**: 338-345.
1100 Xie X, Mikkelsen TS, Gnirke A, Lindblad-Toh K, Kellis M, Lander ES. 2007.
1101 Systematic discovery of regulatory motifs in conserved regions of the human
1102 genome, including thousands of CTCF insulator sites. *Proc Natl Acad Sci U S A*
1103 **104**: 7145-7150.
1104 Xie Z, Hu S, Blackshaw S, Zhu H, Qian J. 2010. hPDI: a database of experimental
1105 human protein-DNA interactions. *Bioinformatics (Oxford, England)* **26**:

1106 287-289.
1107 Yang H, Schramek D, Adam RC, Keyes BE, Wang P, Zheng D, Fuchs E. 2015. ETS
1108 family transcriptional regulators drive chromatin dynamics and malignancy in
1109 squamous cell carcinomas. *eLife* **4**: e10870.
1110 Yardimci GG, Ozadam H, Sauria MEG, Ursu O, Yan KK, Yang T, Chakraborty A, Kaul
1111 A, Lajoie BR, Song F et al. 2019. Measuring the reproducibility and quality of
1112 Hi-C data. *Genome Biol* **20**: 57.
1113 Yoon SO, Chikaraishi DM. 1992. Tissue-specific transcription of the rat tyrosine
1114 hydroxylase gene requires synergy between an AP-1 motif and an overlapping E
1115 box-containing dyad. *Neuron* **9**: 55-67.
1116 Zhang H, Li F, Jia Y, Xu B, Zhang Y, Li X, Zhang Z. 2017. Characteristic arrangement
1117 of nucleosomes is predictive of chromatin interactions at kilobase resolution.
1118 *Nucleic acids research* **45**: 12739-12751.
1119 Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C,
1120 Myers RM, Brown M, Li W et al. 2008. Model-based analysis of ChIP-Seq
1121 (MACS). *Genome Biol* **9**: R137.
1122 Zhang Z, Chng KR, Lingadahalli S, Chen Z, Liu MH, Do HH, Cai S, Rinaldi N, Poh
1123 HM, Li G et al. 2019. An AR-ERG transcriptional signature defined by
1124 long-range chromatin interactomes in prostate cancer cells. *Genome research*
1125 **29**: 223-235.
1126 Zhao Y, Stormo GD. 2011. Quantitative analysis demonstrates most transcription factors
1127 require only simple models of specificity. *Nature biotechnology* **29**: 480-483.
1128