Discovery of biased orientations of human DNA motif sequences affecting enhancer-promoter interactions and transcription of genes

Naoki Osato[1*]

[1]Department of Bioinformatic Engineering, Graduate School of Information Science and Technology, Osaka University, Osaka 565-0871, Japan

[*]Corresponding author

E-mail address: naokiosato11@gmail.com, nosato@ist.osaka-u.ac.jp

Abstract

Chromatin interactions have important roles for enhancer-promoter interactions (EPI) and regulating the transcription of genes. CTCF and cohesin proteins are located at the anchors of chromatin interactions, forming their loop structures. CTCF has insulator function limiting the activity of enhancers into the loops. DNA binding sequences of CTCF indicate their orientation bias at chromatin interaction anchors – forward-reverse (FR) orientation is frequently observed. Other DNA binding proteins such as YY1, ZNF143, and SMARCA4 are also reported to be associated with chromatin interactions. It is still unclear what proteins are associated with chromatin interactions and insulator functions.

To find DNA binding motif sequences of transcription factors (TF), such as CTCF, affecting the interaction between enhancers and promoters of genes and their expression, first I predicted TF bound in enhancers and promoters using DNA motif sequences of TF and experimental data of open chromatin regions in monocytes, T cells, HMEC and NPC, which were obtained from public and commercial databases. Second, transcriptional target genes of each TF were predicted based on enhancer-promoter association (EPA). EPA was shortened at the genomic locations of FR or reverse-forward (RF) orientation of DNA motif sequence of a TF, which were supposed to be at chromatin interaction anchors and acted as insulator sites like CTCF. Then, the expression levels of the transcriptional target genes predicted based on the EPA were compared with those predicted from closed chromatin regions.

Total 96 biased orientations of DNA motifs (64 FR and 52 RF orientations, the

reverse complement sequences of some DNA motifs were also registered in databases, so the total number was smaller than the number of FR and RF) affected the expression level of putative transcriptional target genes significantly in monocytes, T cells, HMEC and NPC in common. For 44 (69%) FR and 28 (54%) RF of 64 FR and 52 RF orientations of DNA motif sequences in $CD4^+$ T cells, EPI predicted from EPA that was shortened at the genomic locations of the biased orientation of DNA motif sequence were overlapped with chromatin interaction data significantly more than other types of EPAs. Moreover, 43 (72%) FR and 40 (80%) RF of 64 FR and 52 RF orientations of DNA motifs found in the four cell types in common showed significantly reduced correlation in expression level of nearby genes separated by the motif sites, implying that the DNA motifs were associated with insulator functions.

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

Background

Chromatin interactions have important roles for enhancer-promoter interactions (EPI) and regulating the transcription of genes. CTCF and cohesin proteins are located at the anchors of chromatin interactions, forming their loop structures. CTCF has insulator function limiting the activity of enhancers into the loops (Fig. 1A). DNA binding sequences of CTCF indicate their orientation bias at chromatin interaction

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

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

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

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

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

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

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

116 ## Results

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

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

121 using TFBS in promoters and enhancers. Enhancer-promoter association (EPA) was
122 shortened at the genomic locations of DNA binding motif sequence of a TF acting as
123 insulator such as CTCF and cohesin (RAD21 and SMC3) (Fig. 1B). To find DNA motif
124 sequences of TF acting as insulator, other than CTCF and cohesin, affecting the
125 expression level of genes, EPA was shortened at the genomic locations of DNA motif
126 sequence of a TF, and transcriptional target genes of each TF bound in (i) distal open
127 chromatin regions (enhancers) and (ii) closed chromatin regions in promoters were
128 predicted based on the EPA, in order to estimate enhancer activity of each TF. The ratio
129 of expression level of putative transcriptional target genes of each TF between (i) and
130 (ii) was calculated. The distribution of the ratios of all TF was compared among
131 forward-reverse (FR), reverse-forward (RF), and any orientation (i.e. without
132 considering orientation) of a DNA motif sequence shortening EPA using Mann-Whitney
133 test, two-sided (p-value $< 10^{-7}$), and the number of TF with significant difference of the
134 distributions among FR, RF and Any was counted (Fig. 2A). DNA sites of TF
135 associated with chromatin interactions were found near TSS (Fig. 2B). ZNF143 is
136 known to bind to promoter of genes establishing looping with distal element bound by
137 CTCF (Fig. 5) (Bailey et al. 2015). YY1 is multi-functional transcriptional regulator and
138 is involved in transcriptional activation, repression and initiation (Shrivastava and
139 Calame 1994; Xi et al. 2007). Most of DNA sites of the TF were found in intergenic and
140 intron regions of human genome (Fig. 2C). Several hundreds of FR and RF orientations
141 of DNA motifs of TF were found in monocytes, T cells, human mammary epithelial
142 cells (HMEC) and neural progenitor cells (NPC). The number of the TF seemed to be

143 changed, according to the number of DNase-seq reads, since the number of the reads of
144 monocyte is larger than other three cell types (i.e. more DNA motif sites would be
145 predicted to some extent, if DNase-seq reads increase). Total 96 of biased (64 FR and
146 52 RF) orientations of DNA binding motif sequences of TF were found in the four cell
147 types in common, whereas any orientation of DNA binding motif sequence was not
148 found (Fig. 3; Table 1; Supplemental material). FR orientation of DNA motif sequences
149 included CTCF, cohesin (RAD21 and SMC3), YY1 and ZNF143 which are associated
150 with chromatin interactions and EPI.

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

165 to alter binding specificity of transcription factors (Ansari and Peterson-Kaufman 2011;
166 Slattery et al. 2011; Merabet and Mann 2016). These increase the variety and number of
167 DNA binding motif sequences of a TF.
168
169 **Comparison with chromatin interaction data**
170 To examine whether the biased orientations of DNA motif sequences is associated
171 with chromatin interactions, enhancer-promoter interactions (EPI) predicted based on
172 enhancer-promoter associations (EPA) were compared with chromatin interaction data.
173 HiChIP chromatin interaction data for $CD4^+$ T cells were used (Mumbach et al. 2017).
174 EPI predicted based on EPA were compared with three replications (B2T1, B2T2, and
175 B3T1) of HiChIP chromatin interaction data in $CD4^+$ T cells respectively. The
176 resolutions of HiChIP chromatin interaction data and EPI were adjusted to 5kb. EPI
177 were predicted based on three types of EPA: (i) EPA shortened at the genomic locations
178 of FR or RF orientation of DNA motif sequence of a TF acting as insulator sites such as
179 CTCF, (ii) EPA shortened at the genomic locations of DNA motif sequence of a TF
180 without considering their orientation, and (iii) EPA without being shortened by the
181 genomic locations of a DNA motif sequence. Total 201 biased orientations [133 (49%)
182 FR and 90 (43%) RF of 273 FR and 211 RF] of DNA motif sequences in T cells, which
183 included CTCF, cohesin (RAD21 and SMC3), ZNF143, and YY1 in three replications
184 (B2T1, B2T2, and B3T1) showed a significantly higher ratio of EPI overlapped with
185 HiChIP chromatin interactions (more than 1,000 counts for each interaction) in EPA (i)
186 than the other types of EPA (ii) and (iii) in T cells (Table 2). When comparing EPI

187 predicted based on only EPA (i) and (iii) with the chromatin interactions, total 390
188 biased orientations [261 (96%) FR and 199 (94%) RF of 273 FR and 211 RF] of DNA
189 motif sequences in T cells showed a significantly higher ratio of EPI predicted based on
190 EPA (i) overlapped with the chromatin interactions than EPI predicted based on EPA
191 (iii) (Table 2; Supplemental material). The difference between EPI predicted based on
192 EPA (i) and (ii) seemed to be difficult to distinguish using the chromatin interaction data
193 and the statistical test in some cases. However, as for the difference between EPI
194 predicted based on EPA (i) and (iii), a larger number of biased orientations of DNA
195 motif sequences was found to be correlated with chromatin interaction data. Chromatin
196 interaction data were obtained from different samples from DNase-seq, open chromatin
197 regions, so individual differences may exist. (Mumbach et al. 2017) suggested that
198 individual differences of chromatin interactions were larger than those of open
199 chromatin regions. Most of biased orientations of DNA motif sequences (95%) were
200 found to be correlated with chromatin interactions, when comparing EPI predicted
201 based on EPA (i) and (iii) with HiChIP chromatin interactions.

202 For HMEC, in situ Hi-C data was available from 4D nucleome data portal, so the
203 chromatin interaction data was compared with putative EPI. Total 222 biased
204 orientations [155 (50%) FR and 108 (40%) RF of 310 FR and 269 RF] of DNA motif
205 sequences in HMEC, which included CTCF, cohesin (RAD21 and SMC3), ZNF143,
206 and YY1 showed a significantly higher ratio of EPI overlapped with in situ Hi-C
207 chromatin interactions (more than one count for each interaction, i.e. all interactions) in
208 EPA (i) than the other types of EPA (ii) and (iii) in HMEC (Supplemental material).

209 When comparing EPI predicted based on only EPA (i) and (iii) with the chromatin
210 interactions, total 406 biased orientations [268 (86%) FR and 230 (86%) RF of 310 FR
211 and 269 RF] of DNA motif sequences in HMEC showed a significantly higher ratio of
212 EPI predicted based on EPA (i) overlapped with the chromatin interactions than EPI
213 predicted based on EPA (iii) (Supplemental material). There were six replications of
214 experimental data of in situ Hi-C data of HMEC, but other replications included much
215 smaller number of chromatin interaction data. The HiChIP chromatin interaction data in
216 T cells was expected to be enriched with enhancer-promoter interactions, since it used
217 H3K27ac antibody to collect chromatin interactions. In situ Hi-C data include all types
218 of chromatin interactions, so in situ Hi-C data seemed to include much smaller number
219 of enhancer-promoter interactions, compared with the HiChIP data. Therefore, other
220 replications of in situ Hi-C data with a smaller number of chromatin interactions will
221 not be available for this analysis.

222 Promoter capture Hi-C data of NPC was downloaded and analyzed in the same
223 way as T cell and HMEC ('GSE86189_npc.po.all.txt.bz2' file) (Jung et al. 2019). Total
224 109 biased orientations [61 (28%) FR and 60 (25%) RF of 205 FR and 241 RF] of DNA
225 motif sequences in NPC, which included CTCF, cohesin (RAD21 and SMC3), ZNF143,
226 and YY1 showed a significantly higher ratio of EPI overlapped with promoter capture
227 Hi-C chromatin interactions (more than one count for each interaction, i.e. all
228 interactions) in EPA (i) than the other types of EPA (ii) and (iii) in NPC (Supplemental
229 material). When comparing EPI predicted based on only EPA (i) and (iii) with the
230 chromatin interactions, total 186 biased orientations [99 (48%) FR and 103 (43%) RF of

231 205 FR and 241 RF] of DNA motif sequences in NPC showed a significantly higher
232 ratio of EPI predicted based on EPA (i) overlapped with the chromatin interactions than
233 EPI predicted based on EPA (iii) (Supplemental material). All three types of chromatin
234 interaction data (HiChIP, in situ Hi-C and promoter capture Hi-C) identified known TF
235 associated with chromatin interactions (CTCF, RAD21, SMC3, ZNF143 and YY1) in
236 three cell types (T cells, HMEC and NPC). For monocyte, Hi-C data was available, but
237 it was not analyzed, due to the low resolution of chromatin interactions (50kb). The
238 ratio of biased orientations of DNA motifs of TF, where EPI were significantly more
239 overlapped with chromatin interactions in EPA (i) than EPA (ii) and (iii), was the
240 highest in H3K27ac HiChIP chromatin interaction data, which were enriched with
241 enhancer-promoter interactions. By adjusting the parameters of the comparison, the
242 ratio may be more increased for the comparison with HiChIP chromatin interaction
243 data.

244 Moreover, to examine the enhancer activity of EPI, the distribution of expression
245 level of putative target genes of EPI was compared between EPI overlapped with
246 HiChIP chromatin interactions and EPI not overlapped with them. Though the target
247 genes of EPI were selected from top 4,000 transcripts (genes) in terms of the expression
248 level of all transcripts (genes) excluding transcripts not expressed in T cells, target
249 genes of EPI overlapped with chromatin interactions showed a significantly higher
250 expression level than EPI not overlapped with them, suggesting that EPI overlapped
251 with chromatin interactions activated the expression of target genes in T cells. Almost
252 all (99%) FR and RF orientations of DNA motifs showed a significantly higher

253 expression level of putative target genes of EPI overlapped with chromatin interactions
254 than EPI not overlapped. When a biased orientation of DNA motif showed a
255 significantly higher expression level of putative target genes of EPI overlapped with
256 chromatin interactions than EPI not overlapped, '1' was marked with in the tables of
257 comparison between EPI and HiChIP chromatin interactions in Supplemental material.
258 DNA motifs showing a significantly lower expression level or not showing significant
259 difference of expression level were not observed in this analysis. HiChIP data were
260 produced using H3K27ac antibody, so chromatin interactions acting as repressor would
261 not be identified. For in situ Hi-C and promoter capture Hi-C data in HMEC and NPC,
262 the significant difference of expression level of putative target genes of EPI was
263 observed in a small percentage (less than 1%).
264
265 **Correlation of expression level of gene pairs separated by a DNA motif sequence**
266 To examine the effect of biased orientations of DNA motif sequences on gene
267 expression, I compared the expression level of protein-coding genes the most closely
268 located to upstream and downstream of biased orientations of DNA motif sequences in
269 human genome. Closely located genes are expected to show correlation of gene
270 expression among tissues (Xie et al. 2007). When a biased orientation of DNA motif
271 sequence acts as insulator (e.g. CTCF), the correlation of gene expression level would
272 be reduced. Among 96 biased (64 FR and 52 RF) orientations of DNA motifs found in
273 common in monocytes, T cells, HMEC, and NPC, top 20 were selected in ascending
274 order of the median correlation coefficient of expression level of the closest gene pairs

275 of the DNA motif that was observed in more than 50 genomic locations. The number of
276 genomic locations was relatively small, when ubiquitously expressed genes among 53
277 tissues and DNA motif sites near the genes were eliminated with coefficient of variance
278 < 90 (see Methods). Pairs of genes near all 20 motifs showed a distribution of
279 significantly lower correlation of gene expression level than all pairs of genes, and
280 included known DNA binding motif sequences of TF associated with chromatin
281 interactions such as CTCF, RAD21, and SMC3 (Fig. 4) (Mann-Whitney test, p-value <
282 0.01). Of 96 biased orientations of DNA motifs found in common in the four cell types,
283 43 (72%) FR and 40 (80%) RF of 60 FR and 50 RF with variable expression data of
284 nearby genes among 53 tissues showed a distribution of significantly lower correlation
285 of gene expression of the closest genes of the DNA motifs than the correlation of all
286 pairs of neighbor genes (Mann-Whitney test, p-value < 0.05) (Table 3). The DNA
287 motifs included known DNA binding motif sequences of TF associated with chromatin
288 interactions such as CTCF, RAD21, SMC3, YY1 and ZNF143. Though the number of
289 DNA motif sites was reduced, compared with all DNA motif sites of a TF, DNA motif
290 sites examined for biased orientation in EPA showed a distribution of significantly
291 lower correlation of expression level of gene pairs separated by the sites (33 (60%) FR
292 and 36 (77%) RF of 55 FR and 47 RF orientations of DNA motifs of TF with variable
293 expression data of nearby genes) (Supplemental material). Some DNA motif sites of
294 biased orientations of DNA motifs were found from genomic regions not including
295 known DNA motifs of TF associated with chromatin interactions such as CTCF,
296 RAD21, SMC3, YY1 and ZNF143, and located in significantly lower correlation of

297 expression level of gene pairs. Significantly lower correlation of gene expression tend to
298 be observed in DNA motifs with relatively a large number of genomic sites (Median
299 number of sites of statistically significant DNA motifs is 656 and 479 for FR and RF
300 respectively, and not significant DNA motifs is 122 and 115 for FR and RF respectively.
301 *p*-value = 0 and < 10^{-10} for FR and RF respectively, Mann-Whitney test). Many DNA
302 motifs showed significantly lower correlation of gene expression, so I wondered there
303 may be a possibility that when we take a small number of gene pairs randomly (about
304 hundreds as the same as the number of DNA motif sites), the correlation of gene pairs
305 might be significantly low occasionally. However, the possibility that happen was low
306 and less than expectation from *p*-value threshold by computer simulation.
307
308 **Co-location of biased orientations of DNA motif sequences**
309 To examine the association of biased orientations of DNA binding motif
310 sequences of TF, co-location of the DNA binding motif sequences within 200bp to each
311 other was analyzed in monocytes, T cells, HMEC, and NPC. The number of genes with
312 the same pairs of DNA binding motif sequences in upstream and downstream of the
313 genes in EPA was counted, and the pairs of DNA binding motif sequences found in
314 more than 100 genomic regions were listed (Table 4; Supplemental material). Majority
315 of pairs of DNA motif sites were overlapped with more than 1bp, but the same tendency
316 was found in the analysis of DNA motif sites of all TF (Supplemental material). As
317 already known, CTCF was found with cohesin RAD21 and SMC3 (Table 4). Top 30
318 pairs of FR and RF orientations of DNA motifs overlapped or co-localized were shown

319 (Table 4). Total number of unique pairs of DNA motifs overlapped or co-localized
320 (counted based on gene symbols, not motif IDs) was 404 (FR) and 662 (RF) in
321 monocytes, consisting of 177 (39% of all FR) and 218 (43% of all RF) unique DNA
322 motifs. Only ten overlapping or co-location of DNA motifs of TF were found in the four
323 cell types in common, including pairs of CTCF, RAD21, SMC3 and YY1. This implied
324 that most of overlapping and co-location of DNA motifs of TF act in cell-type-specific
325 manner (Supplemental material). Pairs of biased orientations of DNA binding motifs of
326 TF tend to be overlapped, and relatively a small number of motifs were co-localized in
327 upstream and downstream of genes. Overlapping DNA motifs act as DNA binding sites
328 of several TF simultaneously or competitively (Ackerman et al. 1991; Yoon and
329 Chikaraishi 1992; Ansari and Peterson-Kaufman 2011; Slattery et al. 2011; Chatterjee et
330 al. 2012; He et al. 2015; Kin et al. 2016; Merabet and Mann 2016). CTCF and cohesin
331 have also been reported to act cooperatively with other TF and protein such as Klf4 and
332 Brd2 (Wei et al. 2013; Hsu et al. 2017). Klf4 recruits cohesin to the Oct4 enhancer
333 region. BRD2 supports boundary (insulator) activity of CTCF. The analysis of 457
334 ChIP-seq data sets on 119 human TFs showed that secondary motifs were detected in
335 addition to the canonical motifs of the TFs, indicating tethered binding and co-binding
336 between multiple TFs (Wang et al. 2012a). The authors observed significant position
337 and orientation preferences between many co-binding TFs. Overlapping DNA motif
338 sites of TF found in this study may not be artifact, but biologically functional. Though
339 repeat sequences in human genome were masked, more than a hundred of pairs of DNA
340 motif sites overlapped at hundreds or thousands upstream and downstream regions of

341 genes.

342

343 **Discussion**

344 To find DNA motif sequences of transcription factors (TF) affecting the
345 expression level of human putative transcriptional target genes, DNA motif sequences
346 were searched from open chromatin regions of monocytes, T cells, HMEC and NPC.
347 Total 96 biased [64 forward-reverse (FR) and 52 reverse-forward (RF)] orientations of
348 DNA motif sequences of TF were found in the four cell types in common, whereas any
349 orientation (i.e. without considering orientation) of DNA motif sequence of TF was not
350 found to affect the expression level of putative transcriptional target genes, suggesting
351 that enhancer-promoter association (EPA) shortened at the genomic locations of FR or
352 RF orientation of DNA motif sequence of a TF is effective for the prediction of
353 enhancer-promoter interactions (EPI).

354 At first, EPI were compared with chromatin interactions (Hi-C) in monocytes.
355 Using open chromatin regions overlapped with ChIP-seq experimental data of histone
356 modification of an enhancer mark (H3K27ac), the ratio of EPI not overlapped with
357 Hi-C was reduced. (Phanstiel et al. 2017) also reported that there was an especially
358 strong enrichment for loops with H3K27 acetylation peaks at both ends (Fisher's Exact
359 Test, $p = 1.4 \times 10^{-27}$). However, the total number of EPI overlapped with chromatin
360 interactions was also reduced using H3K27ac peaks, so more chromatin interaction data
361 would be needed to obtain reliable results in this analysis. As an issue of experimental
362 data, data for chromatin interactions and open chromatin regions were obtained from

363 different samples and donors, so individual differences would exist in the data. The
364 resolution of chromatin interaction data used in monocytes was about 50kb, thus the
365 number of chromatin interactions was relatively small (72,284 at 50kb resolution with a
366 cutoff score of CHiCAGO tool > 1 and 16,501 with a cutoff score of CHiCAGO tool >
367 5) ('PCHiC_peak_matrix_cutoff0.txt.gz' file was downloaded from 'Promoter Capture
368 Hi-C in 17 human primary blood cell types' website https://osf.io/u8tzp/files/). To
369 examine the difference of the numbers of EPI overlapped with chromatin interactions,
370 according to the three types of EPAs, the number of chromatin interactions should be
371 large enough.

372 As HiChIP chromatin interaction data were available in $CD4^+$ T cells, biased
373 orientations of DNA motif sequences of TF were examined in T cell. The resolutions of
374 chromatin interactions and EPI were adjusted to 5kb by fragmentation of genome
375 sequences. The number of unique HiChIP cis (i.e. in the same chromatin) chromatin
376 interactions was 19,926,360 at 5kb resolution, 666,149 at 5kb resolution with chromatin
377 interactions (more than 1,000 counts for each interaction), and 78,209 at 5kb resolution
378 with chromatin interactions (more than 6,000 counts for each interaction). As expected,
379 the number of EPI overlapped with chromatin interactions was increased, and 133 FR
380 and 90 RF orientations of DNA motif sequences of TF showed a statistical significance
381 in EPI predicted based on EPA shortened at the genomic locations of the DNA motif
382 sequence, compared with the other types of EPAs or EPA not shortened. HiChIP
383 analysis used H3K27ac antibody to capture chromatin interactions and identified twenty
384 millions of chromatin interactions, but the number of peaks of ChIP-seq data of

385 H3K27ac is usually less than a hundred thousand. There seems to be much difference of
386 the number of H3K27ac sites between HiChIP and ChIP-seq, so ChIP-seq of H3K27ac
387 may not cover all H3K27ac sites. For HMEC, in situ Hi-C chromatin interaction data
388 were available, the number of unique in situ Hi-C cis chromatin interactions was
389 121,873,642 at 3kb resolution. All chromatin interaction data were used for this analysis
390 to compare EPI with chromatin interactions. For NPC, the number of unique promoter
391 capture Hi-C cis chromatin interactions was 9,436,689 at 5kb resolution. All chromatin
392 interaction data were utilized for analysis.

393 When forming a homodimer or heterodimer with another TF, TF may bind to
394 genome DNA with a specific orientation of their DNA binding sequences (Fig. 5). From
395 the analysis of biased orientations of DNA motif sequences of TF, TF forming
396 heterodimer would also be found. If the DNA binding motif sequence of only the mate
397 to a pair of TF was found in EPA, EPA was shortened at one side, which is the genomic
398 location of the DNA binding motif sequence of the mate to the pair, and transcriptional
399 target genes were predicted using the EPA shortened at the side. In this analysis, the
400 mate to both heterodimer and homodimer of TF can be used to examine the effect on the
401 expression level of transcriptional target genes predicted based on the EPA shortened at
402 one side. Biased orientations of DNA motif sequences may also be found in
403 forward-forward or reverse-reverse orientation.

404 In general, DNA motif sequences of TF were searched from genome sequences
405 using position weight matrix or position frequency matrix and allowing mismatch of
406 DNA sequences under a certain threshold. However, the change of the threshold

407 affected the result of analyses and it seems to be difficult to find the best parameter
408 setting to reduce false-positive and false-negative predictions of DNA biding sites of TF.
409 Therefore, firstly, to obtain a robust result in this study, position weight matrixes of
410 DNA motifs were converted into consensus DNA sequences including degenerated sites,
411 which consist of A, T, G, C, R, Y, M, K, S, W, B, H, V, D, and N letters. This does not
412 need parameter setting in the search for DNA motif sequences, and is the same as *k*-mer
413 analysis of DNA motif sequences of TF. Thought the number of genomic regions
414 matched with a DNA motif sequence was decreased, about 200 FR and RF orientations
415 of DNA motif sequences of TF were found, and the number of any orientation of DNA
416 motif sequences was quite small, suggesting that false-positive prediction of DNA
417 biding sites of TF was reduced. However, the length of the consensus DNA sequences
418 of some TF is relatively long. When the core DNA binding sequence of a TF is short (5
419 or 6 bp), it may be strict to find perfect matches of the long consensus DNA sequence in
420 genome sequences, not allowing a mismatch of a DNA sequence.

421 Secondly, protein interaction quantitation (PIQ) tool was used to find DNA motif
422 sites of TF (Sherwood et al. 2014). PIQ predicts DNA motifs sites of TF by reducing
423 noise of DNase-seq data, and the comparison of the result of prediction with ChIP-seq
424 data revealed that PIQ showed a better performance than other tools and digital genome
425 footprinting (DNase-DGF). The results of prediction of DNA motif sites of TF using
426 PIQ in this study showed an increase of biased orientations of DNA motif sequences of
427 TF, reducing the number of any orientation of DNA motif sequences of TF. The number
428 of DNA motif sites of TF was expected to increase, compared with the above method to

429 use consensus DNA sequences of TF. However, the number of DNA motif sites was
430 almost the same or slightly decreased, compared with the result of consensus DNA
431 sequences, suggesting that PIQ increases true positive prediction of DNA motif sites,
432 reducing false positive prediction.

433 To estimate enhancer activities of DNA motifs of TF, the expression level of
434 putative transcriptional target genes of a TF was compared between DNA motifs of TF
435 in distal open chromatin regions (enhancers) and closed chromatin regions in promoters.
436 Previously, the expression level was compared between distal open chromatin regions
437 and open chromatin regions in promoters. However, it would be easy to understand the
438 comparison of DNA motif sites of a TF between distal open chromatin regions and
439 closed chromatin regions to estimate the enhancer activity of TF. The number of DNA
440 motif sites in closed chromatin regions was large, so the number in closed chromatin
441 regions in promoter was used.

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

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

450

451

452 ## Methods

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

454 To identify transcription factor binding sites (TFBS) from open and closed
455 chromatin regions, TRANSFAC (2019.2), JASPAR (2018), UniPROBE (2018),
456 high-throughput SELEX, transcription factor binding sequences of ENCODE ChIP-seq
457 data, and HOCOMOCO version 9 and 11 were used to predict insulator sites
458 (Wingender et al. 1996; Newburger and Bulyk 2009; Portales-Casamar et al. 2010; Xie
459 et al. 2010; Zhao and Stormo 2011; Jolma et al. 2013; Kheradpour and Kellis 2014;
460 Kulakovskiy et al. 2018). TRANSFAC (2019.2) was used to analyze enhancer-promoter
461 interactions, since these data were sufficient to identify biased orientations of DNA
462 motif sequences of TF with less computational time, reducing the number of any
463 orientation of DNA motif sequences of TF. To reduce false positive prediction of DNA
464 motif sites of TF in open chromatin regions, protein interaction quantitation (PIQ) tool
465 was used with 12,249 position frequency matrices converted from DNA motifs of
466 vertebrate TF in the above databases and DNase-seq data from GEO database
467 (GSM1024791 $CD14^+$ monocytes; GSM665812 $CD4^+$ T cell; GSM736634 Human
468 mammary epithelial cell, HMEC; GSM878615 Neural progenitor cell, NPC) (Sherwood
469 et al. 2014). To predict DNA motif sites of TF from closed chromatin regions, another
470 method was used. Position weight matrices of vertebrate transcription factor binding
471 sequences were converted into DNA consensus sequences considering degenerated sites
472 using convert matrix in RSAT website (van Helden 2003). Transcription factor binding

473 sequences of vertebrate TF were used for further analyses. Transcription factor binding
474 sequences were searched from each narrow peak of DNase-seq data in repeat-masked
475 hg19 genome sequences using Match tool in TRANSFAC with similarity score cutoff of
476 1 after transforming the DNA consensus sequences of TF into position frequency
477 matrixes for Match search (Kel et al. 2003). Repeat-masked hg19 genome sequences
478 were downloaded from UCSC genome browser (http://genome.ucsc.edu/,
479 http://hgdownload.soe.ucsc.edu/goldenPath/hg19/bigZips/hg19.fa.masked.gz). TF
480 corresponding to transcription factor binding sequences were searched computationally
481 and manually by comparing their names and gene symbols of HGNC (HUGO Gene
482 Nomenclature Committee) -approved gene nomenclature and 31,848 UCSC known
483 canonical transcripts
484 (http://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/knownCanonical.txt.gz), as
485 transcription factor binding sequences were not linked to transcript IDs such as UCSC,
486 RefSeq, and Ensembl transcripts.

487 Target genes of a TF were assigned when its TFBS was found in DNase-seq
488 narrow peaks in promoter or extended regions for enhancer-promoter association of
489 genes (EPA). Promoter and extended regions were defined as follows: promoter regions
490 were those that were within distance of ±5 kb from transcriptional start sites (TSS).
491 Promoter and extended regions were defined as per the following association rule,
492 which is the same as that defined in Figure 3A of a previous study (McLean et al. 2010):
493 the single nearest gene association rule, which extends the regulatory domain to the
494 midpoint between the TSS of the gene and that of the nearest gene upstream and

495 downstream without the limitation of extension length. Extended regions for EPA were
496 shortened at the genomic locations of DNA binding sites of a TF that was the closest to
497 a transcriptional start site, and transcriptional target genes were predicted from the
498 shortened enhancer regions using TFBS. Furthermore, promoter and extended regions
499 for EPA were shortened at the genomic locations of forward–reverse (FR) orientation of
500 DNA binding sites of a TF. When forward or reverse orientation of DNA binding sites
501 were continuously located in genome sequences several times, the most internal (i.e.
502 closest to a transcript) forward–reverse orientation of DNA binding sites were selected.
503 The genomic positions of genes were identified using 'knownGene.txt.gz' file in UCSC
504 bioinformatics sites (Karolchik et al. 2014). The file 'knownCanonical.txt.gz' was also
505 utilized for choosing representative transcripts among various alternate forms for
506 assigning promoter and extended regions for EPA. From the list of transcription factor
507 binding sequences and transcriptional target genes, redundant transcription factor
508 binding sequences were removed by comparing the target genes of a transcription factor
509 binding sequence and its corresponding TF; if identical, one of the transcription factor
510 binding sequences was used. When the number of transcriptional target genes predicted
511 from a transcription factor binding sequence was less than five, the transcription factor
512 binding sequence was omitted.

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

517 For gene expression data, RNA-seq reads mapped onto human hg19 genome
518 sequences were obtained from UCSF-UBC human reference epigenome mapping
519 project RNA-seq reads with poly-A of naive $CD4^+$ T cells (GEO: GSM669617).
520 FPKMs of the RNA-seq data were calculated using RSeQC (Wang et al. 2012b). For
521 monocytes, Blueprint RNA-seq FPKM data
522 ('C0010KB1.transcript_quantification.rsem_grape2_crg.GRCh38.20150622.results')
523 were downloaded from Blueprint DCC portal
524 (http://dcc.blueprint-epigenome.eu/#/files). RNA-seq reads with poly-A of human
525 mammary epithelial cells (HMEC) in Encyclopedia of DNA Elements at UCSC
526 (http://genome.ucsc.edu/encode/,
527 'wgEncodeCshlLongRnaSeqHmecCellPapAlnRep1.bam' file) and H1 derived neural
528 progenitor cell (NPC) (GEO: GSM915326, Encode: ENCFF529SIO) were used. Based
529 on log2-transformed FPKM, UCSC transcripts were arranged in descending order of
530 expression level, and top 4,000 transcripts were selected in each cell type.
531 The expression level of transcriptional target genes predicted based on EPA
532 shortened at the genomic locations of DNA motif sequence of a TF was compared with
533 the expression level of transcriptional target genes predicted from closed chromatin
534 regions in promoters. For each DNA motif sequence shortening EPA, transcriptional
535 target genes were predicted using DNA binding motif sequences of vertebrate TF in
536 TRANSFAC database, and the ratio of expression level of putative transcriptional target
537 genes of each TF was calculated between EPA and closed chromatin regions in
538 promoters. The distribution of the ratios of all TF was compared among forward-reverse

539 (FR), reverse-forward (RF), and any orientation (i.e. without considering orientation) of
540 a DNA motif sequence shortening EPA using Mann-Whitney test, two-sided (p-value <
541 10^{-7}). Other parameters for the analysis were that the number of transcriptional target
542 genes of a TF was more than 50, and the number of TF (not insulator TF) to predict
543 transcriptional target genes was more than 50. The number of DNA motif sites of TF
544 (insulator TF) shortening EPA was more than 100. When a DNA motif sequence of a TF
545 shortening EPA showed a significant difference of expression level of putative
546 transcriptional target genes among FR, RF or any orientation, the DNA motif sequence
547 was listed.
548
549 **Comparison with chromatin interaction data**
550 For comparison of EPA with chromatin interactions (HiChIP) in $CD4^+$ T cells,
551 'GSM2705049_Naive_HiChIP_H3K27ac_B2T1_allValidPairs.txt',
552 'GSM2705050_Naive_HiChIP_H3K27ac_B2T2_allValidPairs.txt', and
553 'GSM2705051_Naive_HiChIP_H3K27ac_B3T1_allValidPairs.txt' files were
554 downloaded from Gene Expression Omnibus (GEO) database (GSM2705049,
555 GSM2705050 and GSM2705051). The resolutions of chromatin interaction data and
556 EPI were adjusted to 5kb before their comparison. Chromatin interactions with more
557 than 6,000 and 1,000 counts for each interaction were used in this study.
558 Enhancer-promoter interactions (EPI) were predicted using three types of EPAs in
559 monocytes: (i) EPA shortened at the genomic locations of FR or RF orientation of DNA
560 motif sequence of a TF, (ii) EPA shortened at the genomic locations of any orientation

561 (i.e. without considering orientation) of DNA motif sequence of a TF, and (iii) EPA
562 without being shortened by a DNA motif sequence. EPI predicted in the three types of
563 EPAs in common were removed. First, EPI predicted based on EPA (i) were compared
564 with chromatin interactions. The resolution of chromatin interaction data used in this
565 study was 1-5kb, so EPI were adjusted to 5kb before their comparison. The number and
566 ratio of EPI overlapped with chromatin interactions were counted. Second, EPI were
567 predicted based on EPA (ii), and EPI predicted based on EPA (i) were removed from the
568 EPI. The number and ratio of EPI overlapped with chromatin interactions were counted.
569 Third, EPI were predicted based on EPA (iii), and EPI predicted based on EPA (i) and
570 (ii) were removed from the EPI. The number and ratio of EPI overlapped with
571 chromatin interactions were counted. The number and ratio of the EPI were compared
572 two times between EPA (i) and (iii), and EPA (i) and (ii) (binomial distribution, p-value
573 < 0.025 for each test, two-sided, 95% confidence interval).

574 For comparison of EPA with chromatin interactions (in situ Hi-C) in HMEC,
575 'HMEC_4DNFI97O9SAZ.pairs.gz' file was downloaded from 4D nucleome data portal
576 (http://data.4dnucleome.org). The genomic positions of chromatin interactions were
577 converted into hg19 version of human genome using liftOver tool
578 (http://hgdownload.soe.ucsc.edu/downloads.html#liftover). The resolutions of
579 chromatin interaction data and EPI were adjusted to 3kb before their comparison. All
580 chromatin interactions were used for analysis. For comparison of EPA with chromatin
581 interactions (promoter capture Hi-C) in NPC, 'GSE86189_npc.po.all.txt.bz2' file was
582 downloaded from GEO database (GSE86189). The resolutions of chromatin interaction

583 data and EPI were adjusted to 5kb before their comparison. All chromatin interactions
584 were used for analysis.

585 Putative target genes for the analysis of EPI were selected from top 4000
586 transcripts in term of the expression level. The expression level of putative target genes
587 of EPI overlapped with HiChIP chromatin interactions was compared with EPI not
588 overlapped with them. For each FR or RF orientation of DNA motif, EPI were predicted
589 based on EPA and the overlap of EPI with chromatin interactions was examined. When
590 a putative transcriptional target gene of a TF in an enhancer was found in both EPI
591 overlapped with a chromatin interaction and EPI not overlapped with, the target gene
592 was removed. The distribution of expression level of putative target genes was
593 compared using Mann-Whitney test, two-sided (p-value < 0.05).

594

595 **Correlation of expression level of gene pairs separated by a DNA motif sequence**

596 To examine the correlation of expression level of genes pairs separated by a DNA
597 motif sequence, gene expression data among 53 tissues were obtained from
598 'GTEx_Analysis_2016-01-15_v7_RNASeQCv1.1.8_gene_median_tpm.gct.gz' file in
599 GTEx database (https://gtexportal.org/home/). The correlation of log2-transformed gene
600 expression level was analyzed based on protein-coding transcripts of RefSeq (RefSeq
601 ID with 'NM_'). The same genomic location of RefSeq transcripts was removed. DNA
602 motifs overlapped with RefSeq transcripts were eliminated. Genes closest to DNA
603 motifs and other genes were analyzed using BEDOPS closest-features (Neph et al.
604 2012). Mann-Whitney test was performed using R.

605

606 **Co-location of biased orientations of DNA motif sequences**

607 Co-location of biased orientations of DNA binding motif sequences of TF was
608 examined. The number of genes with the same pair of DNA motifs of biased
609 orientations of DNA motifs of TF in upstream and downstream of genes in EPA was
610 counted, and the pairs of DNA motifs found in more than 100 genomic regions were
611 listed.

612

613 ## Acknowledgements

614 The supercomputing resource was provided by Human Genome Center of the Institute
615 of Medical Science at the University of Tokyo. Computations were partially performed
616 on the NIG supercomputer at ROIS National Institute of Genetics. Publication charges
617 for this article were funded by JSPS KAKENHI Grant Number 16K00387. This
618 research was partially supported by the Platform Project for Supporting in Drug
619 Discovery and Life Science Research（Platform for Dynamic Approaches to Living
620 System）from Japan Agency for Medical Research and Development (AMED). This
621 research was partially supported by Development of Fundamental Technologies for
622 Diagnosis and Therapy Based upon Epigenome Analysis from Japan Agency for
623 Medical Research and Development (AMED). This work was partially supported by
624 JST CREST Grant Number JPMJCR15G1, Japan.

625

626 **Figures**

627 Web links to original figures.

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

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

630 76_2014_BFnrg3682_MOESM18_ESM.ppt

631 Graphical Abstract (de Wit et al. 2015)

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

633 92c5eba1d/fx1.jpg

634

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

636 2010)

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

638 87_2010_BFnbt1630_MOESM11_ESM.pdf

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

640 DNA Looping (Guo et al. 2015)

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

642 2af081251/gr4.jpg

643

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

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

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

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

648 2014; de Wit et al. 2015; Guo et al. 2015). (B) Computationally-defined regulatory
649 domains for enhancer-promoter association (McLean et al. 2010). The *single nearest*
650 *gene* association rule extends the regulatory domain to the midpoint between this gene's
651 TSS and the nearest gene's TSS both upstream and downstream. Enhancer-promoter
652 association was shortened at the genomic locations of forward-reverse orientation of
653 DNA binding motif sequence of a transcription factor (e.g. CTCF in the figure).
654

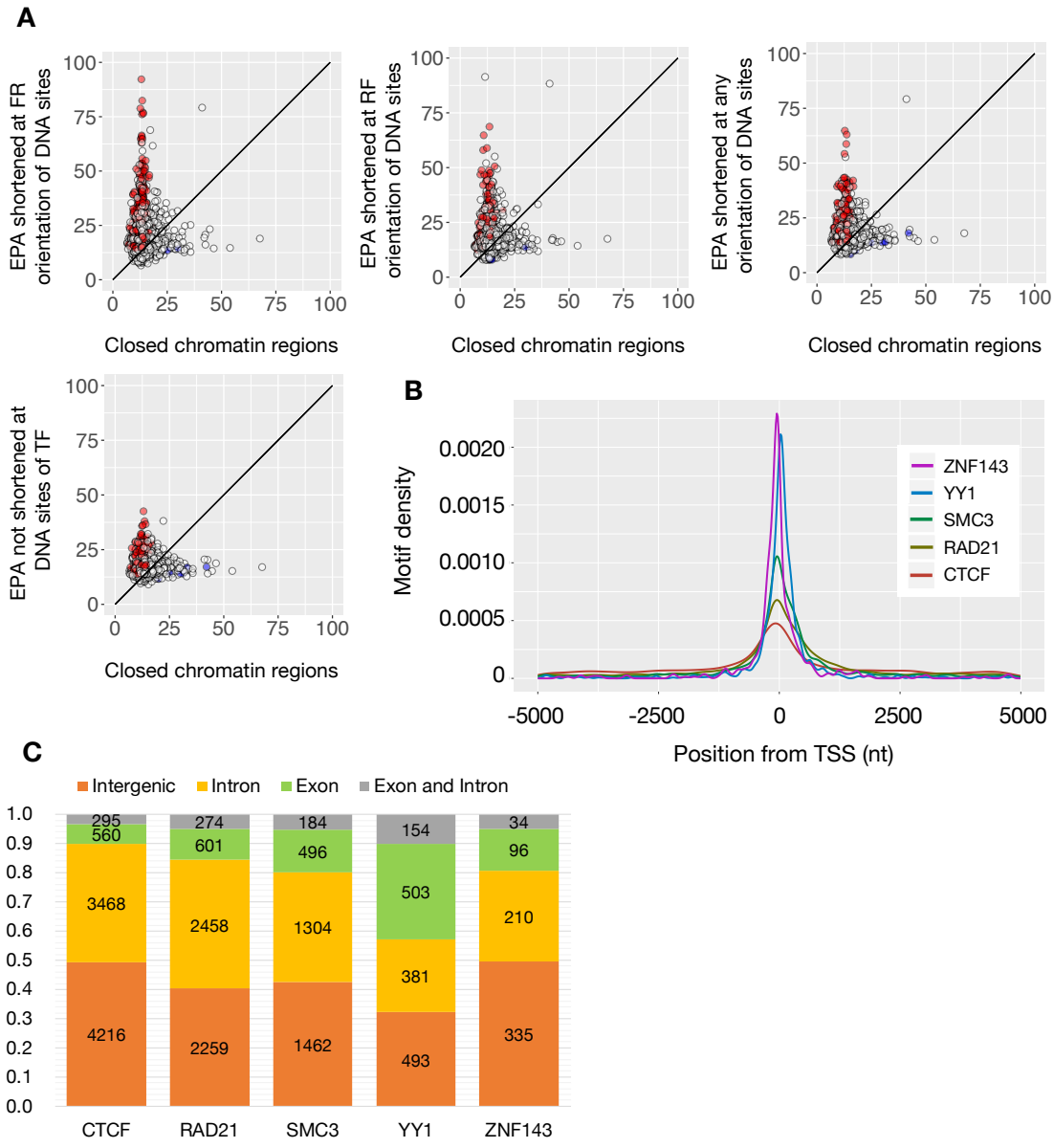

Figure 2. Identification of DNA sites of TF from open chromatin regions and characteristics of DNA sites. (A) Comparison of expression level of putative transcriptional target genes. The median expression levels (FPKM) of target genes of the same transcription factor binding sequences were compared between closed chromatin regions (or promoters) and enhancer-promoter association (EPA) shortened at the genomic locations of forward-reverse orientation of DNA motif sequence of CTCF

662 in monocytes. The distribution of expression levels was changed according to four
663 criteria of EPA (see Method) (Osato 2018). Red dots show the median expression level
664 of target genes was higher significantly (Mann-Whitney test) based on the EPA than
665 closed chromatin regions, and blue dots show the median expression level of target
666 genes was lower significantly based on the EPA than closed chromatin regions. The
667 expression level of target genes predicted based on the EPA tend to be higher than
668 closed chromatin regions, implying that transcription factors acted as activators of target
669 genes. (B) Distribution of DNA sites of TF near TSS. ZNF143 and YY1 tend to locate
670 near TSS, and CTCF tend to be found more from distant regions than around TSS. (C)
671 Distribution of DNA sites of TF in human genome. Most of DNA sites were found from
672 intergenic and intron regions. YY1 are known to be at TSS in the first exon of genes,
673 associated with transcriptional activation, repression and initiation. YY1 is
674 multi-functional transcriptional regulator (Shrivastava and Calame 1994; Xi et al. 2007)
675

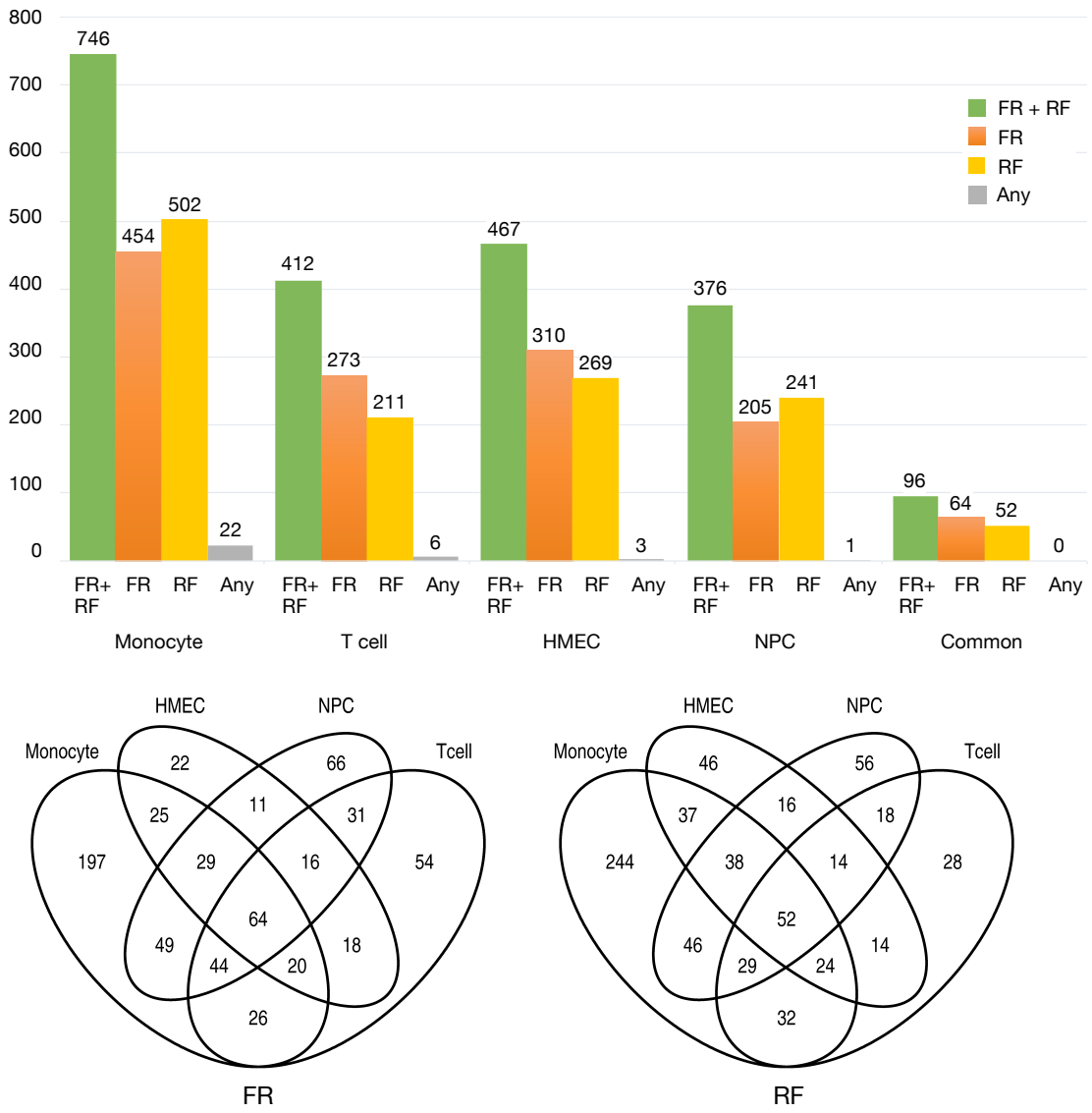

Figure 3. Biased orientations of DNA motif sequences of transcription factors affecting the transcription of genes. Biased orientations of DNA motif sequences of transcription factors. Total 96 (64 FR and 52 RF) of biased orientations of DNA binding motif sequences of transcription factors were found to affect significantly the expression level of putative transcriptional target genes in monocytes, T cells, HMEC and NPC in common, whereas any orientation (i.e. without considering orientation) of DNA binding motif sequence was not found to affect the expression level significantly.

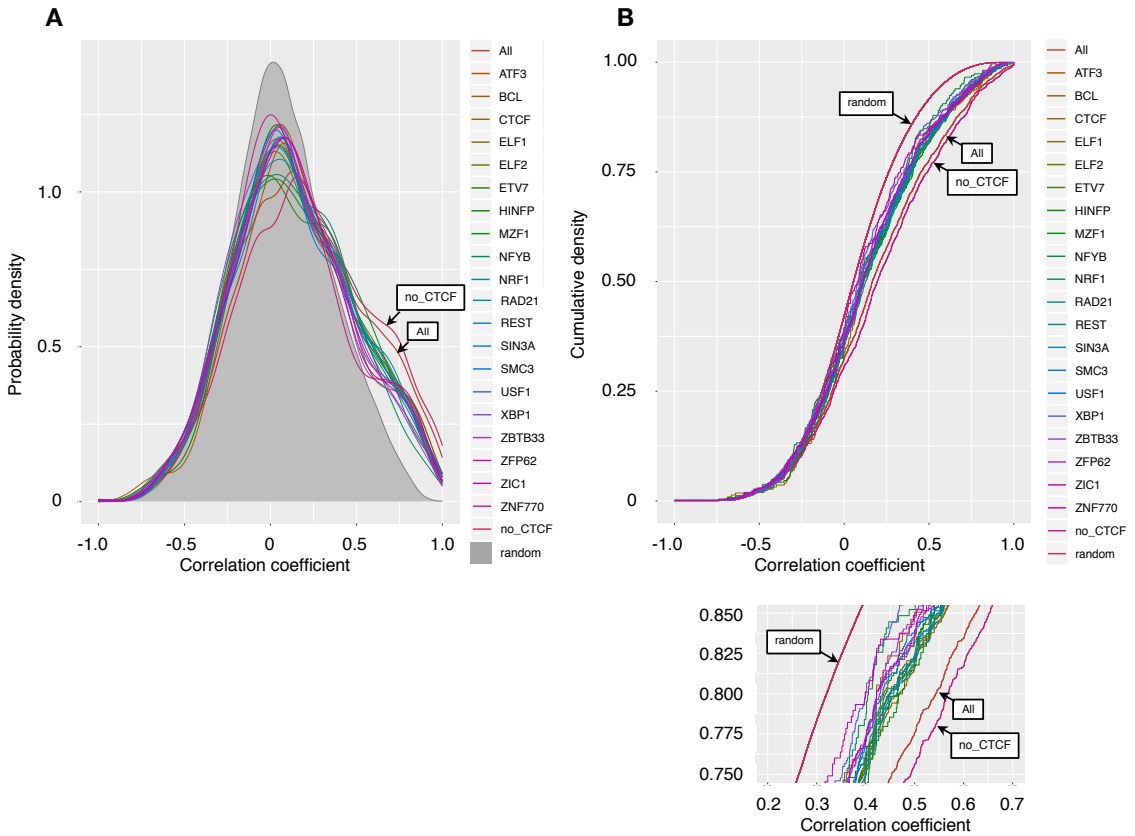

Figure 4. Genes separated by predicted DNA binding sites of biased orientations of DNA motifs of TFs are less correlated in gene expression. Top 20 TF are shown according to ascending order of median correlation coefficient. Correlation coefficient between neighboring gene pairs is shown in terms of probability density (A) and cumulative distribution (B). Brown line (All), correlation between all neighboring genes; gray shading in (A) and red line (random) in (B), correlation between randomly chosen gene pairs; Purple line (no_CTCF), correlation between gene pairs not separated by predicted CTCF sites.

696

697 Web links to original figures.

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

699 (Bailey et al. 2015).

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

701

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

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

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

705 2015).

706

Tables

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

715 **Table 2.** Comparison of enhancer-promoter interactions (EPI) with chromatin
716 interactions in T cells. The upper tables show the number of biased orientations of DNA
717 motifs, where a significantly higher ratio of EPI, which were predicted based on
718 enhancer-promoter association (EPA) (i) overlapped with HiChIP chromatin interaction
719 data than the other types of EPA (ii) and (iii). The middle tables show 133 FR and 90
720 RF orientations of DNA motifs in T cells found in common among B2T1, B2T2 and
721 B3T1 replications in the upper table. The lower table shows that among 64 FR and 52
722 RF biased orientations of DNA motifs found in common in monocytes, T cell, HMEC
723 and NPC, 44 (69%) FR and 28 (54%) RF were matched with the analysis of HiChIP for
724 three types of EPA. TF: DNA binding motif sequence of a transcription factor. Score: –
725 $\log_{10}$ (*p*-value). Score 1: Comparison of EPA shortened at the genomic locations of FR
726 or RF orientation of DNA motif sequence [EPA (i)] with EPA not shortened [EPA (iii)].
727 Score 2: Comparison of EPA shortened at the genomic locations of FR or RF orientation
728 of DNA motif sequence [EPA (i)] with EPA shortened at the genomic locations of any
729 orientation of DNA motif sequence [EPA (ii)].

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

Replication of HiChIP data	Total no. (FR + RF) of DNA motifs	No. of FR DNA motifs	No. of RF DNA motifs
B2T1	226	153	101
B2T2	221	149	98
B3T1	237	160	109
DNA motifs found in common among replications			
B2T1 and B2T2	212	141	95
B2T1 and B3T1	212	143	93
B2T2 and B3T1	208	138	93
B2T1, B2T2 and B3T1	201	133	90

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

Replication of HiChIP data	Total no. (FR + RF) of DNA motifs	No. of FR DNA motifs	No. of RF DNA motifs
B2T1	395	262	205
B2T2	398	264	204
B3T1	394	263	203
DNA motifs found in common among replications			
B2T1 and B2T2	394	262	202
B2T1 and B3T1	391	261	202
B2T2 and B3T1	392	262	200
B2T1, B2T2 and B3T1	390	261	199

Forward-reverse (FR)

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

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

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

Reverse-forward (RF)

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

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

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

Forward-reverse (FR)

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

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

Reverse-forward (RF)

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

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

752 DNA motif site nearby were not used for statistical test, so the number of DNA site is
753 relatively small. The results using all genes were shown in Supplemental material. All
754 DNA motif sites of a TF were used including DNA motif sites examined for biased
755 orientation in EPA. The results using only biased orientation of DNA motif sites
756 examined in this study were shown in Supplemental material.

757

758 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

759

760

761 **Table 4.** Top 30 of co-location and overlapping biased orientations of DNA binding
762 motif sequences of transcription factors in monocytes. Co-location and overlapping
763 DNA motif sequences of CTCF with another biased orientation of DNA motif sequence
764 were shown in a separate table. Motif 1,2: DNA binding motif sequences of
765 transcription factors. No.: the number of genes with both Motif 1 and Motif 2 in
766 upstream and downstream of genes in EPA (< 1 Mb distance from TSS of genes). Total
767 number of genes is 30,523.

768

769 Forward-reverse (FR)

Motif 1	Motif 2	No.	Mean of distances between motifs (nt)	Median of distances between motifs (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

770

771

772

Reverse-forward (RF)

Motif 1	Motif 2	No.	Mean of distances between motifs (nt)	Median of distances between motifs (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

FR orientation of CTCF

Motif 1	Motif 2	No.	Mean of distances between motifs (nt)	Median of distances between motifs (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 between motifs (nt)	Median of distances between motifs (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

782 **References**

783

784 Ackerman SL, Minden AG, Williams GT, Bobonis C, Yeung CY. 1991. Functional
785 significance of an overlapping consensus binding motif for Sp1 and Zif268 in
786 the murine adenosine deaminase gene promoter. *Proc Natl Acad Sci U S A* **88**:
787 7523-7527.
788 Ansari AZ, Peterson-Kaufman KJ. 2011. A partner evokes latent differences between
789 Hox proteins. *Cell* **147**: 1220-1221.
790 Bailey SD, Zhang X, Desai K, Aid M, Corradin O, Cowper-Sal Lari R, Akhtar-Zaidi B,
791 Scacheri PC, Haibe-Kains B, Lupien M. 2015. ZNF143 provides sequence
792 specificity to secure chromatin interactions at gene promoters. *Nat Commun* **2**:
793 6186.
794 Barutcu AR, Lajoie BR, Fritz AJ, McCord RP, Nickerson JA, van Wijnen AJ, Lian JB,
795 Stein JL, Dekker J, Stein GS et al. 2016. SMARCA4 regulates gene expression
796 and higher-order chromatin structure in proliferating mammary epithelial cells.
797 *Genome research* **26**: 1188-1201.
798 Chatterjee R, Zhao J, He X, Shlyakhtenko A, Mann I, Waterfall JJ, Meltzer P,
799 Sathyanarayana BK, FitzGerald PC, Vinson C. 2012. Overlapping ETS and CRE
800 Motifs ((G/C)CGGAAGTGACGTCA) preferentially bound by GABPalpha and
801 CREB proteins. *G3 (Bethesda)* **2**: 1243-1256.
802 Chen M, Manley JL. 2009. Mechanisms of alternative splicing regulation: insights from
803 molecular and genomics approaches. *Nat Rev Mol Cell Biol* **10**: 741-754.
804 Das D, Clark TA, Schweitzer A, Yamamoto M, Marr H, Arribere J, Minovitsky S,
805 Poliakov A, Dubchak I, Blume JE et al. 2007. A correlation with exon
806 expression approach to identify cis-regulatory elements for tissue-specific
807 alternative splicing. *Nucleic acids research* **35**: 4845-4857.
808 de Wit E, Vos ES, Holwerda SJ, Valdes-Quezada C, Verstegen MJ, Teunissen H,
809 Splinter E, Wijchers PJ, Krijger PH, de Laat W. 2015. CTCF Binding Polarity
810 Determines Chromatin Looping. *Molecular cell* **60**: 676-684.
811 Guo Y, Xu Q, Canzio D, Shou J, Li J, Gorkin DU, Jung I, Wu H, Zhai Y, Tang Y et al.
812 2015. CRISPR Inversion of CTCF Sites Alters Genome Topology and
813 Enhancer/Promoter Function. *Cell* **162**: 900-910.

He X, Syed KS, Tillo D, Mann I, Weirauch MT, Vinson C. 2015. GABPalpha Binding to Overlapping ETS and CRE DNA Motifs Is Enhanced by CREB1: Custom DNA Microarrays. *G3 (Bethesda)* **5**: 1909-1918.

Hsu SC, Gilgenast TG, Bartman CR, Edwards CR, Stonestrom AJ, Huang P, Emerson DJ, Evans P, Werner MT, Keller CA et al. 2017. The BET Protein BRD2 Cooperates with CTCF to Enforce Transcriptional and Architectural Boundaries. *Molecular cell* **66**: 102-116 e107.

Ji X, Dadon DB, Abraham BJ, Lee TI, Jaenisch R, Bradner JE, Young RA. 2015. Chromatin proteomic profiling reveals novel proteins associated with histone-marked genomic regions. *Proc Natl Acad Sci U S A* **112**: 3841-3846.

Jolma A, Yan J, Whitington T, Toivonen J, Nitta KR, Rastas P, Morgunova E, Enge M, Taipale M, Wei G et al. 2013. DNA-binding specificities of human transcription factors. *Cell* **152**: 327-339.

Jolma A, Yin Y, Nitta KR, Dave K, Popov A, Taipale M, Enge M, Kivioja T, Morgunova E, Taipale J. 2015. DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature* **527**: 384-388.

Jung I, Schmitt A, Diao Y, Lee AJ, Liu T, Yang D, Tan C, Eom J, Chan M, Chee S et al. 2019. A compendium of promoter-centered long-range chromatin interactions in the human genome. *Nature genetics* **51**: 1442-1449.

Karolchik D, Barber GP, Casper J, Clawson H, Cline MS, Diekhans M, Dreszer TR, Fujita PA, Guruvadoo L, Haeussler M et al. 2014. The UCSC Genome Browser database: 2014 update. *Nucleic acids research* **42**: D764-770.

Katainen R, Dave K, Pitkanen E, Palin K, Kivioja T, Valimaki N, Gylfe AE, Ristolainen H, Hanninen UA, Cajuso T et al. 2015. CTCF/cohesin-binding sites are frequently mutated in cancer. *Nature genetics* **47**: 818-821.

Kel AE, Gossling E, Reuter I, Cheremushkin E, Kel-Margoulis OV, Wingender E. 2003. MATCH: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic acids research* **31**: 3576-3579.

Kheradpour P, Kellis M. 2014. Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic acids research* **42**: 2976-2987.

Kin K, Chen X, Gonzalez-Garay M, Fakhouri WD. 2016. The effect of non-coding DNA variations on P53 and cMYC competitive inhibition at cis-overlapping

847 motifs. *Hum Mol Genet* **25**: 1517-1527.
848 Kulakovskiy IV, Vorontsov IE, Yevshin IS, Sharipov RN, Fedorova AD, Rumynskiy EI,
849 Medvedeva YA, Magana-Mora A, Bajic VB, Papatsenko DA et al. 2018.
850 HOCOMOCO: towards a complete collection of transcription factor binding
851 models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic acids*
852 *research* **46**: D252-D259.
853 McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM,
854 Bejerano G. 2010. GREAT improves functional interpretation of cis-regulatory
855 regions. *Nature biotechnology* **28**: 495-501.
856 Merabet S, Mann RS. 2016. To Be Specific or Not: The Critical Relationship Between
857 Hox And TALE Proteins. *Trends Genet* **32**: 334-347.
858 Mumbach MR, Satpathy AT, Boyle EA, Dai C, Gowen BG, Cho SW, Nguyen ML,
859 Rubin AJ, Granja JM, Kazane KR et al. 2017. Enhancer connectome in primary
860 human cells identifies target genes of disease-associated DNA elements. *Nature*
861 *genetics* **49**: 1602-1612.
862 Neph S, Kuehn MS, Reynolds AP, Haugen E, Thurman RE, Johnson AK, Rynes E,
863 Maurano MT, Vierstra J, Thomas S et al. 2012. BEDOPS: high-performance
864 genomic feature operations. *Bioinformatics (Oxford, England)* **28**: 1919-1920.
865 Newburger DE, Bulyk ML. 2009. UniPROBE: an online database of protein binding
866 microarray data on protein-DNA interactions. *Nucleic acids research* **37**:
867 D77-82.
868 Osato N. 2018. Characteristics of functional enrichment and gene expression level of
869 human putative transcriptional target genes. *BMC Genomics* **19**: 957.
870 Phanstiel DH, Van Bortle K, Spacek D, Hess GT, Shamim MS, Machol I, Love MI,
871 Aiden EL, Bassik MC, Snyder MP. 2017. Static and Dynamic DNA Loops form
872 AP-1-Bound Activation Hubs during Macrophage Development. *Molecular cell*
873 **67**: 1037-1048.e1036.
874 Portales-Casamar E, Thongjuea S, Kwon AT, Arenillas D, Zhao X, Valen E, Yusuf D,
875 Lenhard B, Wasserman WW, Sandelin A. 2010. JASPAR 2010: the greatly
876 expanded open-access database of transcription factor binding profiles. *Nucleic*
877 *acids research* **38**: D105-110.
878 Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn
879 AL, Machol I, Omer AD, Lander ES et al. 2014. A 3D map of the human

880 genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**:
881 1665-1680.
882 Schreiber J, Libbrecht M, Bilmes J, Noble W. 2017. Nucleotide sequence and DNaseI
883 sensitivity are predictive of 3D chromatin architecture. *bioRxiv*.
884 Sherwood RI, Hashimoto T, O'Donnell CW, Lewis S, Barkal AA, van Hoff JP, Karun V,
885 Jaakkola T, Gifford DK. 2014. Discovery of directional and nondirectional
886 pioneer transcription factors by modeling DNase profile magnitude and shape.
887 *Nature biotechnology* **32**: 171-178.
888 Shlyueva D, Stampfel G, Stark A. 2014. Transcriptional enhancers: from properties to
889 genome-wide predictions. *Nature reviews Genetics* **15**: 272-286.
890 Shrivastava A, Calame K. 1994. An analysis of genes regulated by the multi-functional
891 transcriptional regulator Yin Yang-1. *Nucleic acids research* **22**: 5151-5155.
892 Slattery M, Riley T, Liu P, Abe N, Gomez-Alcala P, Dror I, Zhou T, Rohs R, Honig B,
893 Bussemaker HJ et al. 2011. Cofactor binding evokes latent differences in DNA
894 binding specificity between Hox proteins. *Cell* **147**: 1270-1282.
895 Tabuchi TM, Deplancke B, Osato N, Zhu LJ, Barrasa MI, Harrison MM, Horvitz HR,
896 Walhout AJ, Hagstrom KA. 2011. Chromosome-biased binding and gene
897 regulation by the Caenorhabditis elegans DRM complex. *PLoS Genet* **7**:
898 e1002074.
899 van Helden J. 2003. Regulatory sequence analysis tools. *Nucleic acids research* **31**:
900 3593-3596.
901 Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF,
902 Schroth GP, Burge CB. 2008. Alternative isoform regulation in human tissue
903 transcriptomes. *Nature* **456**: 470-476.
904 Wang J, Zhuang J, Iyer S, Lin X, Whitfield TW, Greven MC, Pierce BG, Dong X,
905 Kundaje A, Cheng Y et al. 2012a. Sequence features and chromatin structure
906 around the genomic regions bound by 119 human transcription factors. *Genome*
907 *research* **22**: 1798-1812.
908 Wang L, Wang S, Li W. 2012b. RSeQC: quality control of RNA-seq experiments.
909 *Bioinformatics (Oxford, England)* **28**: 2184-2185.
910 Wei Z, Gao F, Kim S, Yang H, Lyu J, An W, Wang K, Lu W. 2013. Klf4 organizes
911 long-range chromosomal interactions with the oct4 locus in reprogramming and
912 pluripotency. *Cell Stem Cell* **13**: 36-47.

913 Weintraub AS, Li CH, Zamudio AV, Sigova AA, Hannett NM, Day DS, Abraham BJ,
914 Cohen MA, Nabet B, Buckley DL et al. 2017. YY1 Is a Structural Regulator of
915 Enhancer-Promoter Loops. *Cell* **171**: 1573-1588 e1528.
916 Wingender E, Dietze P, Karas H, Knuppel R. 1996. TRANSFAC: a database on
917 transcription factors and their DNA binding sites. *Nucleic acids research* **24**:
918 238-241.
919 Xi H, Yu Y, Fu Y, Foley J, Halees A, Weng Z. 2007. Analysis of overrepresented motifs
920 in human core promoters reveals dual regulatory roles of YY1. *Genome research*
921 **17**: 798-806.
922 Xie X, Mikkelsen TS, Gnirke A, Lindblad-Toh K, Kellis M, Lander ES. 2007.
923 Systematic discovery of regulatory motifs in conserved regions of the human
924 genome, including thousands of CTCF insulator sites. *Proc Natl Acad Sci U S A*
925 **104**: 7145-7150.
926 Xie Z, Hu S, Blackshaw S, Zhu H, Qian J. 2010. hPDI: a database of experimental
927 human protein-DNA interactions. *Bioinformatics (Oxford, England)* **26**:
928 287-289.
929 Yoon SO, Chikaraishi DM. 1992. Tissue-specific transcription of the rat tyrosine
930 hydroxylase gene requires synergy between an AP-1 motif and an overlapping E
931 box-containing dyad. *Neuron* **9**: 55-67.
932 Zhang H, Li F, Jia Y, Xu B, Zhang Y, Li X, Zhang Z. 2017. Characteristic arrangement
933 of nucleosomes is predictive of chromatin interactions at kilobase resolution.
934 *Nucleic acids research* **45**: 12739-12751.
935 Zhao Y, Stormo GD. 2011. Quantitative analysis demonstrates most transcription factors
936 require only simple models of specificity. *Nature biotechnology* **29**: 480-483.
937