Discovery of biased orientations of human regulatory motifs affecting transcription of genes and including known insulators

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 (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. 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 (TFs), 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 TFs and experimental data of open chromatin regions in monocytes, T, Th17, Treg, GM12878 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). An EPA was shortened at the FR or reverse-forward (RF) orientation of DNA motif sites of a TF, which were supposed to be at chromatin interaction anchors and acted as insulator sites like CTCF. Then, the expression levels of 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, EPIs predicted from EPAs that were shortened at the biased orientations of DNA motif sites, overlapped with chromatin interactions 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 (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 binding sequence of CTCF indicate its orientation bias at chromatin interaction anchors

55 – forward-reverse (FR) orientation is frequently observed (Rao et al. 2014; de Wit et al.
56 2015; Guo et al. 2015). Other DNA binding proteins such as ZNF143, YY1 and
57 SMARCA4 (BRG1) are found to be associated with chromatin interaction 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 interaction
60 and EPI. The DNA binding motif sequences of the transcription factors (TFs) 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 interaction and insulator function may be
82 found from other researches.

83 As computational approaches, machine-learning analyses to predict chromatin
84 interactions were proposed (Schreiber et al. 2017; Zhang et al. 2017). However, they
85 were not intended to find DNA motif sequences of TFs affecting chromatin interaction,
86 EPI and the expression level of transcriptional target genes, which were examined in
87 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). An EPA was shortened at the FR orientation of CTCF binding sites, and
93 transcriptional target genes of each TF bound in enhancers and promoters were
94 predicted based on the EPA. The EPA affected the expression levels of putative
95 transcriptional target genes, relative to the expression levels of transcriptional target
96 genes predicted from promoters (or closed chromatin regions) (Fig. 2A). The expression
97 levels tend to be increased in monocytes and $CD4^+$ T cells, implying that enhancers
98 activated the transcription of genes. The expression levels tend to be decreased in ES

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

116 ## Results

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

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

143 monocytes, T cells, human mammary epithelial cells (HMEC) and neural progenitor
144 cells (NPC). The number of the TFs seemed to be changed, according to the number of
145 DNase-seq reads, since the number of the reads in monocyte is larger than the other
146 three cell types. When DNase-seq reads increase, more DNA motif sites may be
147 predicted to some extent. Total 96 of biased (64 FR and 52 RF) orientations of DNA
148 binding motif sequences of TFs were found in the four cell types in common, whereas
149 any orientation of DNA binding motif sequence was not found (Fig. 3; Table 1;
150 Supplemental material). The FR orientation of DNA motif sequences included CTCF,
151 cohesin (RAD21 and SMC3), YY1 and ZNF143, which are associated with chromatin
152 interaction and EPI.

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

165 DNA binding motif sequences of TFs composing the complex in *C. elegans* (Tabuchi et
166 al. 2011). The difference of binding specificity of pairs of transcription factors was
167 reported (Jolma et al. 2015). Cofactors also contribute to alter binding specificity of
168 transcription factors (Ansari and Peterson-Kaufman 2011; Slattery et al. 2011; Merabet
169 and Mann 2016). These increase the variety and number of DNA binding motif
170 sequences of a TF, which are potentially involved in different biological functions
171 among cell types.
172
173 **Comparison with chromatin interaction data**
174 To examine whether the biased orientations of DNA motif sequences are
175 associated with chromatin interaction, enhancer-promoter interactions (EPIs) predicted
176 based on an enhancer-promoter association (EPA) were compared with chromatin
177 interaction data. HiChIP chromatin interaction data were used for $CD4^+$ T cells
178 (Mumbach et al. 2017). EPIs predicted based on an EPA were compared with three
179 replications (B2T1, B2T2 and B3T1) of HiChIP chromatin interaction data respectively.
180 The resolutions of HiChIP chromatin interaction data and EPIs were adjusted to 5
181 kilobases (kb). EPI were predicted based on three types of EPAs: (i) EPA shortened at
182 the FR or RF orientation of DNA motif sites of a TF acting as insulator such as CTCF,
183 (ii) EPA shortened at the DNA motif sites of a TF without considering their orientation,
184 and (iii) EPA without being shortened by DNA motif sites. Total 201 biased orientations
185 [133 (49%) FR and 90 (43%) RF of 273 FR and 211 RF] of DNA motif sequences in T
186 cells, which included CTCF, cohesin (RAD21 and SMC3), ZNF143 and YY1 in three

187 replications (B2T1, B2T2, and B3T1), showed a significantly higher ratio of EPIs
188 overlapped with HiChIP chromatin interactions with ≥1,000 counts for each interaction
189 in EPA (i) than the other types of EPA (ii) and (iii) in T cells (Table 2). When comparing
190 EPIs predicted based on only EPA (i) and (iii) with the chromatin interactions, total 390
191 biased orientations [261 (96%) FR and 199 (94%) RF of 273 FR and 211 RF] of DNA
192 motif sequences in T cells showed a significantly higher ratio of EPIs overlapped with
193 the chromatin interactions in EPA (i) than EPA (iii) (Table 2; Supplemental material).
194 The difference between EPIs predicted based on EPA (i) and (ii) seemed to be difficult
195 to distinguish using the chromatin interaction data and statistical test in some cases.
196 However, as for the difference between EPIs predicted based on EPA (i) and (iii), a
197 larger number of biased orientations of DNA motif sequences were found to be
198 correlated with chromatin interaction data. Chromatin interaction data were obtained
199 from different samples from DNase-seq open chromatin regions, so individual
200 differences may exist. (Mumbach et al. 2017) suggested that individual differences of
201 chromatin interactions were larger than those of open chromatin regions. Most of biased
202 orientations of DNA motif sequences (95%) were found to be correlated with chromatin
203 interactions, when comparing EPIs predicted based on EPA (i) and (iii) with HiChIP
204 chromatin interactions.

205 Additionally, to confirm these tendencies of comparison of EPIs with HiChIP
206 chromatin interactions, the same analysis was conducted using HiChIP chromatin
207 interaction data in Th17, Treg and GM12878 cells. Total 128 biased orientations [76
208 (38%) FR and 64 (35%) RF of 200 FR and 183 RF] of DNA motif sequences in Th17

209 cells, which included CTCF, cohesin (RAD21 and SMC3), ZNF143 and YY1 in three
210 replications (B1T2, B2T1, and B3T1), showed a significantly higher ratio of EPIs
211 overlapped with HiChIP chromatin interactions in EPA (i) than the other types of EPA
212 (ii) and (iii) (Supplemental material). When comparing EPIs predicted based on only
213 EPA (i) and (iii) with the chromatin interactions, total 282 biased orientations [182
214 (91%) FR and 162 (89%) RF of 200 FR and 183 RF] of DNA motif sequences showed a
215 significantly higher ratio of EPIs overlapped with the chromatin interactions in EPA (i)
216 than EPA (iii) (Supplemental material).

217 For Treg cells, chromatin interactions in one biological replication were
218 overlapped with EPIs, but those in the other biological replications were less overlapped
219 with EPIs, so the one replication (B3T1) was used for this analysis. Total 281 biased
220 orientations [154 (53%) FR and 176 (54%) RF of 290 FR and 323 RF] of DNA motif
221 sequences, which included CTCF, cohesin (RAD21 and SMC3), ZNF143 and YY1,
222 showed a significantly higher ratio of EPIs overlapped with HiChIP chromatin
223 interactions in EPA (i) than the other types of EPA (ii) and (iii) (Supplemental material).
224 When comparing EPIs predicted based on only EPA (i) and (iii) with the chromatin
225 interactions, total 482 biased orientations [283 (98%) FR and 309 (96%) RF of 290 FR
226 and 323 RF] of DNA motif sequences showed a significantly higher ratio of EPIs
227 overlapped with the chromatin interactions in EPA (i) than EPA (iii) (Supplemental
228 material).

229 For GM12878 cells, total 168 biased orientations [96 (38%) FR and 98 (35%) RF
230 of 229 FR and 237 RF] of DNA motif sequences, which included CTCF, cohesin

231 (RAD21 and SMC3), ZNF143 and YY1 in two replications (B1 and B2), showed a
232 significantly higher ratio of EPIs overlapped with HiChIP chromatin interactions in EPA
233 (i) than the other types of EPA (ii) and (iii) (Supplemental material). When comparing
234 EPIs predicted based on only EPA (i) and (iii) with the chromatin interactions, total 366
235 biased orientations [218 (95%) FR and 227 (96%) RF of 229 FR and 237 RF] of DNA
236 motif sequences showed a significantly higher ratio of EPIs overlapped with the
237 chromatin interactions in EPA (i) than EPA (iii) (Supplemental material).

238 In situ Hi-C data was available from 4D nucleome data portal in HMEC, so the
239 chromatin interaction data was compared with putative EPIs. Total 222 biased
240 orientations [155 (50%) FR and 108 (40%) RF of 310 FR and 269 RF] of DNA motif
241 sequences, which included CTCF, cohesin (RAD21 and SMC3), ZNF143 and YY1,
242 showed a significantly higher ratio of EPIs overlapped with in situ Hi-C chromatin
243 interactions (≥1 count for each interaction, i.e. all interactions) in EPA (i) than the other
244 types of EPA (ii) and (iii) (Supplemental material). When comparing EPIs predicted
245 based on only EPA (i) and (iii) with the chromatin interactions, total 406 biased
246 orientations [268 (86%) FR and 230 (86%) RF of 310 FR and 269 RF] of DNA motif
247 sequences showed a significantly higher ratio of EPIs overlapped with the chromatin
248 interactions in EPA (i) than EPA (iii) (Supplemental material). There were six
249 replications of in situ Hi-C experimental data of HMEC, but other replications
250 contained much smaller number of chromatin interactions. The HiChIP chromatin
251 interaction data in T cells were expected to be enriched with enhancer-promoter
252 interactions, since it used H3K27ac antibody to collect chromatin interactions. In situ

253 Hi-C data contain all types of chromatin interactions and thus, contain a smaller number
254 of enhancer-promoter interactions, relative to HiChIP data. Chromatin interactions with
255 a small number of counts may be artifacts derived from ligation errors during
256 experiments (Lieberman-Aiden et al. 2009; Yardimci et al. 2019). Therefore, other
257 replications of in situ Hi-C data with a smaller number of chromatin interactions will
258 not be available for this analysis.

259 Promoter capture Hi-C data of NPC was downloaded and analyzed in the same
260 way as T cell and HMEC (Jung et al. 2019). Total 109 biased orientations [61 (28%) FR
261 and 60 (25%) RF of 205 FR and 241 RF] of DNA motif sequences, which included
262 CTCF, cohesin (RAD21 and SMC3), ZNF143 and YY1, showed a significantly higher
263 ratio of EPIs overlapped with promoter capture Hi-C chromatin interactions (≥1 count
264 for each interaction, i.e. all interactions) in EPA (i) than the other types of EPA (ii) and
265 (iii) (Supplemental material). When comparing EPIs predicted based on only EPA (i)
266 and (iii) with the chromatin interactions, total 186 biased orientations [99 (48%) FR and
267 103 (43%) RF of 205 FR and 241 RF] of DNA motif sequences showed a significantly
268 higher ratio of EPIs overlapped with the chromatin interactions in EPA (i) than EPA (iii)
269 (Supplemental material). All three types of chromatin interaction data (HiChIP, in situ
270 Hi-C and promoter capture Hi-C) indicated that EPIs predicted based on EPA (i) that
271 were shortened at DNA motif sites of known TFs associated with chromatin interactions
272 (CTCF, RAD21, SMC3, ZNF143 and YY1), overlapped with more chromatin
273 interactions than the other EPA (ii) and (iii) in six cell types (T cells, Th17, Treg,
274 GM12878, HMEC and NPC). For monocyte, Hi-C data was available, but it was not

275 analyzed, due to the low resolution of chromatin interaction data (50 kb).

276 There may be a possibility that the biased orientations of DNA motif sites of
277 known TFs associated with chromatin interactions such as CTCF, RAD21, SMC3, YY1
278 and ZNF143 are located with the biased orientations of DNA motif sites of other TFs,
279 and thus, they showed a significantly higher ratio of EPIs overlapped with chromatin
280 interactions in EPA (i) than the other types of EPA (ii) and (iii). However, among 273
281 FR and 211 RF orientations of DNA motifs of TFs not including known TFs such as
282 CTCF, RAD21, SMC3, YY1 and ZNF143 in T cells, the DNA motif sites of only 15
283 (6%) FR and 9 (4%) RF orientations of DNA motifs of TFs overlapped with $\geq$ 60% of
284 DNA motifs sites of biased orientations of DNA motifs of either of the known TFs at 5
285 kb resolution.

286 Moreover, to examine the enhancer activity of EPIs, the distribution of expression
287 level of putative target genes of EPIs was compared between EPIs overlapped with
288 HiChIP chromatin interactions and EPIs not overlapped with them. Though the target
289 genes of EPIs were selected from top 4,000 transcripts (genes) in terms of the
290 expression level of all transcripts (genes) excluding transcripts not expressed in T, Th17,
291 Treg and GM12878 cells respectively, target genes of EPIs overlapped with chromatin
292 interactions showed a significantly higher expression level than EPIs not overlapped
293 with them, suggesting that EPIs overlapped with chromatin interactions activated the
294 expression of target genes in the four cell types. Almost all (99% - 100%) FR and RF
295 orientations of DNA motifs showed a significantly higher expression level of putative
296 target genes of EPIs overlapped with chromatin interactions than EPIs not overlapped in

297 the four cell types. When a biased orientation of DNA motif showed a significantly
298 higher expression level of putative target genes of EPIs overlapped with chromatin
299 interactions than EPIs not overlapped, '1' was marked with in the tables of comparison
300 between EPIs and HiChIP chromatin interactions in Supplemental material. DNA motifs
301 showing a significantly lower expression level or not showing a significant difference of
302 expression level were not observed in this analysis. HiChIP data were produced using
303 H3K27ac antibody, so chromatin interactions acting as repressor would not be identified.
304 For in situ Hi-C and promoter capture Hi-C data in HMEC and NPC, the significant
305 difference of expression level of putative target genes of EPIs was observed in a small
306 percentage (less than 1%).
307
308 **Correlation of expression level of gene pairs separated by DNA motif sites of a TF**
309 To examine the effect of biased orientations of DNA motif sequences on gene
310 expression, I compared the expression level of protein-coding genes the most closely
311 located to upstream and downstream of biased orientations of DNA motif sites in the
312 human genome. Closely located genes including divergent gene pairs are expected to
313 show correlation of gene expression among tissues (Purmann et al. 2007; Xie et al.
314 2007). When a biased orientation of DNA motif sites act as insulator (e.g. CTCF), the
315 correlation of gene expression level would be reduced. Among 96 biased (64 FR and 52
316 RF) orientations of DNA motifs of a TF found in common in monocytes, T cells,
317 HMEC and NPC, ubiquitously expressed genes among 53 tissues and DNA motif sites
318 near the genes were eliminated with coefficient of variance < 90, and the DNA motif

319 sites of a TF with ≥50 genomic locations were selected (see Methods). After this
320 filtering, 43 (72%) FR and 40 (80%) RF of 60 FR and 50 RF showed a distribution of
321 significantly lower correlation of gene expression of the closest genes of the DNA motif
322 sites of a TF than the correlation of all pairs of neighbor genes with intergene distance
323 <1 megabases (Mb) (Mann-Whitney test, p-value < 0.05) (Table 3). The FR and RF
324 orientations of TFs included known TFs associated with chromatin interactions such as
325 CTCF, RAD21, SMC3, YY1 and ZNF143. Top 20 of the FR and RF orientations of
326 DNA motifs were selected in ascending order of the median correlation coefficient of
327 expression level of the closest gene pairs of the DNA motif sites of a TF. The 20 TFs
328 included known TFs associated with chromatin interactions such as CTCF, RAD21 and
329 SMC3 (Fig. 4) (Mann-Whitney test, p-value < 0.01). Instead of using all DNA motif
330 sites of TFs, when DNA motif sites of TFs were limited to sites examined for biased
331 orientations of DNA motifs in EPAs, 33 (60%) FR and 36 (77%) RF of 55 FR and 47
332 RF showed a distribution of significantly lower correlation of expression level of gene
333 pairs separated by the sites (Supplemental material). With intergene distance <20 kb and
334 the DNA motif sites of a TF with ≥50 genomic locations, 21 (34%) FR and 8 (15%)
335 RF of 62 FR and 52 RF showed a distribution of significantly lower correlation of gene
336 expression of the closest genes of the DNA motif sites of a TF than the correlation of all
337 pairs of neighbor genes (Supplemental material).

338 Some DNA motif sites of biased orientations of DNA motifs of TFs were found
339 from genomic regions not including known DNA motif sites of TFs associated with
340 chromatin interactions such as CTCF, RAD21, SMC3, YY1 and ZNF143, and were

341 located in genomic regions with significantly lower correlation of expression level of
342 gene pairs. The DNA motifs of TFs with a large number of DNA motif sites tend to
343 show significantly lower correlation of gene expression (Median numbers of sites of
344 statistically significant DNA motifs were 656 and 479 for FR and RF respectively, and
345 median numbers of sites of not significant DNA motifs were 122 and 115 for FR and RF
346 respectively. Mann-Whitney test *p*-value = 0 and < 10^{-10} for FR and RF respectively).
347 Many DNA motifs of TFs showed a significantly lower correlation of gene expression,
348 so there may be a possibility that when we take a small number of gene pairs randomly
349 (about hundreds as the same as the number of DNA motif sites of a TF), the correlation
350 of gene pairs might be significantly low occasionally. However, the possibility that
351 happen was lower than the expectation from *p*-value threshold by computer simulation.
352

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

354 To examine the association of biased orientations of DNA binding motif
355 sequences of TFs, co-location of the DNA binding motif sequences within 200bp to
356 each other, which is close to the length of core DNA around octamer of core histone
357 (146bp), that is, the length of the minimum open chromatin region, was analyzed in
358 monocytes, T cells, HMEC, and NPC. The numbers of the same pairs of DNA binding
359 motif sequences in upstream and downstream of genes in EPAs were counted, and the
360 pairs of DNA binding motif sequences found in ≥100 genomic regions were listed
361 (Table 4; Supplemental material). Majority of pairs of DNA motif sites overlapped with
362 more than 1 base, but the same tendency was found in the analysis of DNA motif sites

363 of all TFs (Supplemental material). As already known, CTCF was found with cohesin
364 RAD21 and SMC3 (Table 4). Top 30 pairs of FR and RF orientations of DNA motifs
365 overlapped or co-localized were shown (Table 4). Total numbers of unique pairs of
366 DNA motif sites of TFs overlapped or co-localized were 404 (FR) and 662 (RF) in
367 monocytes, consisting of 177 (39% of all FR) and 218 (43% of all RF) unique DNA
368 motifs. Only ten overlapping or co-location of DNA motifs of TFs were found in the
369 four cell types in common, including pairs of CTCF, RAD21, SMC3 and YY1. This
370 implied that most of overlapping and co-location of DNA motifs of TFs act in
371 cell-type-specific manner (Supplemental material). Pairs of biased orientations of DNA
372 binding motifs of TFs tend to overlap, and relatively a small number of motifs were
373 co-localized in upstream and downstream of genes. Overlapping DNA motifs act as
374 DNA binding sites of several TF simultaneously or competitively (Ackerman et al.
375 1991; Yoon and Chikaraishi 1992; Ansari and Peterson-Kaufman 2011; Slattery et al.
376 2011; Chatterjee et al. 2012; He et al. 2015; Kin et al. 2016; Merabet and Mann 2016).
377 CTCF and cohesin have also been reported to act coordinately with another TF and a
378 protein such as Klf4 and Brd2 (Wei et al. 2013; Hsu et al. 2017). Klf4 recruits cohesin
379 to the Oct4 enhancer region. BRD2 supports boundary (insulator) activity of CTCF. The
380 analysis of 457 ChIP-seq data sets on 119 human TFs showed that secondary motifs
381 were detected in addition to the canonical motifs of the TFs, indicating tethered binding
382 and co-binding between multiple TFs (Wang et al. 2012a). The authors observed
383 significant position and orientation preferences between many co-binding TFs.
384 Overlapping DNA motif sites of TFs found in this study may not be artifact, but

385 biologically functional. Though repeat sequences in human genome were masked, more
386 than a hundred of pairs of DNA motif sites overlapped at hundreds or thousands
387 upstream and downstream regions of genes.
388

389 **Discussion**

390 To find DNA motif sequences of transcription factors (TFs) affecting the
391 expression level of human putative transcriptional target genes, DNA motif sequences
392 were searched from open chromatin regions of monocytes, T cells, HMEC and NPC.
393 Total 96 biased [64 forward-reverse (FR) and 52 reverse-forward (RF)] orientations of
394 DNA motif sequences of TF were found in the four cell types in common, whereas any
395 orientation (i.e. without considering orientation) of DNA motif sequence of TF was not
396 found to affect the expression level of putative transcriptional target genes, suggesting
397 that enhancer-promoter association (EPA) shortened at the FR or RF orientation of DNA
398 motif sites of a TF is effective for the prediction of enhancer-promoter interactions
399 (EPIs).

400 At first, EPIs were compared with chromatin interactions (Hi-C) in monocytes.
401 Using open chromatin regions overlapped with peaks of ChIP-seq experiment of histone
402 modification marks of an enhancer (H3K27ac), the ratio of EPIs not overlapped with
403 chromatin interactions was reduced. (Phanstiel et al. 2017) also reported that there was
404 an especially strong enrichment for loops with H3K27 acetylation peaks at both ends
405 (Fisher's Exact Test, $p = 1.4 \times 10^{-27}$). However, the total number of EPIs overlapped
406 with chromatin interactions was also reduced using H3K27ac peaks, so more chromatin

407 interaction data would be needed to obtain reliable results in this analysis. As an issue of
408 experimental data, data for chromatin interactions and open chromatin regions were
409 obtained from different samples and donors, so individual differences would exist. The
410 resolution of chromatin interaction data used in monocytes was about 50 kb, thus the
411 number of unique chromatin interactions was relatively small (72,284 at 50 kb
412 resolution with a cutoff score of CHiCAGO tool > 1 and 16,501 with a cutoff score of
413 CHiCAGO tool > 5) ('PCHiC_peak_matrix_cutoff0.txt.gz' file was downloaded from
414 'Promoter Capture Hi-C in 17 human primary blood cell types' website
415 https://osf.io/u8tzp/files/). To examine the difference of the numbers of EPIs overlapped
416 with chromatin interactions, according to the three types of EPAs, the unique and total
417 number of chromatin interactions should be large enough.

418 As HiChIP chromatin interaction data were available in $CD4^+$ T, Th17, Treg and
419 GM12878 cells, biased orientations of DNA motif sequences of TFs were examined in
420 the four cell types. The resolutions of chromatin interactions and EPIs were adjusted to
421 5 kb by fragmentation of genome sequences. The number of unique HiChIP cis (i.e. in
422 the same chromatin) chromatin interactions was 19,926,360 at 5 kb resolution in B2T1
423 replication of T cells, 666,149 at 5 kb resolution with chromatin interactions (≥1,000
424 counts for each interaction) and 78,209 at 5 kb resolution with chromatin interactions
425 (≥6,000 counts for each interaction) (Supplemental material). As expected, the number
426 of EPIs overlapped with chromatin interactions was increased, and 133 FR and 90 RF
427 orientations of DNA motif sequences of TFs showed a statistical significance in EPIs
428 predicted based on an EPA shortened at the DNA motif sites of TFs, compared with the

429 other types of EPA or EPA not shortened. The numbers of unique HiChIP cis chromatin
430 interactions in other cell types (Th17, Treg and GM12878) were shown in Supplemental
431 material. The numbers of unique cis chromatin interactions at 5 kb resolution were
432 different among replications in Th17 cells, so the numbers of interactions were adjusted
433 between the numbers with ≥1,000 and ≥6,000 counts for each interaction in T cells.
434 For Treg, chromatin interactions in one biological replication overlapped with EPIs, but
435 those in the other biological replications were less overlapped with EPIs, so the one
436 replication was used for this analysis. HiChIP analysis used H3K27ac antibody to
437 capture chromatin interactions and identified twenty millions of chromatin interactions,
438 but the number of peaks of ChIP-seq experimental data of H3K27ac is usually less than
439 a hundred thousand. There seems to be much difference of the number of H3K27ac sites
440 between HiChIP and ChIP-seq, so ChIP-seq peaks of H3K27ac may not cover all
441 H3K27ac sites. For HMEC, in situ Hi-C chromatin interaction data were available, the
442 number of unique in situ Hi-C cis chromatin interactions was 121,873,642 at 3 kb
443 resolution. All chromatin interaction data were used for this analysis to compare EPIs
444 with chromatin interactions. For NPC, the number of unique promoter capture Hi-C cis
445 chromatin interactions was 9,436,689 at 5 kb resolution. All chromatin interaction data
446 were utilized for analysis.

447 When forming a homodimer or heterodimer with another TF, TFs may bind to
448 genome DNA with a specific orientation of their DNA binding sequences (Fig. 5). From
449 the analysis of biased orientations of DNA motif sequences of TFs, TFs forming
450 heterodimer would also be found. If the DNA binding motif sequence of only the mate

451 to a pair of TFs was found in an EPA, the EPA was shortened at one side, which is the
452 DNA binding motif site of the mate to the pair, and transcriptional target genes were
453 predicted using the EPA shortened at the side. In this analysis, the mate to both
454 heterodimer and homodimer of TFs can be used to examine the effect on the expression
455 level of transcriptional target genes predicted based on the EPA shortened at one side.
456 Biased orientations of DNA motif sequences may also be found in forward-forward or
457 reverse-reverse orientation.

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

473 is short (5 or 6 bp), it may be strict to find perfect matches of long consensus DNA
474 sequences in genome sequences, not allowing a mismatch of the DNA sequence.

475 Secondly, protein interaction quantitation (PIQ) tool was used to find DNA motif
476 sites of TFs (Sherwood et al. 2014). PIQ predicts DNA motif sites of TFs by reducing
477 noises of DNase-seq data, and the comparison of result of prediction with ChIP-seq data
478 revealed that PIQ has a better performance than other tools and digital genome
479 footprinting (DNase-DGF). The results of prediction of DNA motif sites of TFs using
480 PIQ in this study showed an increase of biased orientations of DNA motif sequences of
481 TFs, reducing the number of any orientation of DNA motif sequences of TFs. The
482 number of DNA motif sites of TFs was expected to increase, relative to the above
483 method to use consensus DNA sequences of TFs. However, the number of DNA motif
484 sites was almost the same or slightly decreased, compared with the result of consensus
485 DNA sequences, suggesting that PIQ increases true positive predictions of DNA motif
486 sites, reducing false positive predictions.

487 To estimate enhancer activities of DNA motifs of TFs, the expression level of
488 putative transcriptional target genes of a TF was compared between DNA motif sites of
489 TFs in distal open chromatin regions (enhancers) and closed chromatin regions in
490 promoters. Previously, the expression level was compared between distal open
491 chromatin regions and open chromatin regions in promoters. However, it would be easy
492 to understand the comparison of DNA motif sites of a TF between distal open chromatin
493 regions and closed chromatin regions to estimate the enhancer activity of TFs. The
494 number of DNA motif sites in closed chromatin regions was large, so the number in

495 closed chromatin regions in promoter was used.

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

500 The analyses in this study revealed novel properties of DNA binding motif
501 sequences of TFs involved in chromatin interaction, insulator function and potentially
502 forming a homodimer, heterodimer or complex with other TFs, affecting the
503 transcriptional regulation of genes.

504

505 ## Methods

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

507 To identify transcription factor binding sites (TFBSs) from open and closed
508 chromatin regions, TRANSFAC (2019.2), JASPAR (2018), UniPROBE (2018),
509 high-throughput SELEX, transcription factor binding sequences of ENCODE ChIP-seq
510 data, and HOCOMOCO version 9 and 11 were used to predict insulator sites
511 (Wingender et al. 1996; Newburger and Bulyk 2009; Portales-Casamar et al. 2010; Xie
512 et al. 2010; Zhao and Stormo 2011; Jolma et al. 2013; Kheradpour and Kellis 2014;
513 Kulakovskiy et al. 2018). TRANSFAC (2019.2) was used to analyze enhancer-promoter
514 interactions, since these data were sufficient to identify biased orientations of DNA
515 motif sequences of TFs with less computational time, reducing the number of any
516 orientation of DNA motif sequences of TFs. To reduce false positive prediction of DNA

517 motif sites of TFs in open chromatin regions, protein interaction quantitation (PIQ) tool
518 was used with 12,249 position frequency matrices converted from DNA motifs of
519 vertebrate TFs in the above databases and DNase-seq data from GEO and Encode
520 database (GSM1024791 $CD14^+$ monocytes; GSM665812 $CD4^+$ T cell; GSM736634
521 Human mammary epithelial cell, HMEC; GSM878615 Neural progenitor cell, NPC;
522 ENCFF263GMV, ENCFF886XEV Th17; GSM1024741 Treg; GSM736620 GM12878)
523 (Sherwood et al. 2014). To find open chromatin regions associated with enhancer (and
524 promoter) activity, ChIP-seq experimental data of H3K27ac was downloaded
525 (GSM773004). Narrow peaks of ChIP-seq data were predicted using macs2 callpeak
526 (Zhang et al. 2008). To predict DNA motif sites of TFs from closed chromatin regions,
527 another method was used. Position weight matrices of vertebrate transcription factor
528 binding sequences were converted into DNA consensus sequences considering
529 degenerated sites using convert matrix in RSAT website (van Helden 2003).
530 Transcription factor binding sequences of vertebrate TFs were used for further analyses.
531 Transcription factor binding sequences were searched from narrow peaks of DNase-seq
532 data in repeat-masked hg19 genome sequences using Match tool in TRANSFAC with
533 similarity score cutoff of 1 after transforming the DNA consensus sequences of TFs into
534 position frequency matrixes for Match search (Kel et al. 2003). Repeat-masked hg19
535 genome sequences were downloaded from UCSC genome browser
536 (http://genome.ucsc.edu/,
537 http://hgdownload.soe.ucsc.edu/goldenPath/hg19/bigZips/hg19.fa.masked.gz). TFs
538 corresponding to transcription factor binding sequences were searched computationally

539 and manually by comparing their names and gene symbols of HGNC (HUGO Gene
540 Nomenclature Committee) -approved gene nomenclature and 31,848 UCSC known
541 canonical transcripts
542 (http://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/knownCanonical.txt.gz), as
543 transcription factor binding sequences were not linked to transcript IDs such as UCSC,
544 RefSeq and Ensembl transcripts.

545 Target genes of a TF were assigned when its TFBS was found in DNase-seq
546 narrow peaks in promoter or extended regions for enhancer-promoter association of
547 genes (EPA). Promoter and extended regions were defined as follows: promoter regions
548 were those that were within distance of ±5 kb from transcriptional start sites (TSS).
549 Promoter and extended regions were defined as per the following association rule,
550 which is the same as that defined in Figure 3A of a previous study (McLean et al. 2010):
551 the single nearest gene association rule, which extends the regulatory domain to the
552 midpoint between the TSS of the gene and that of the nearest gene upstream and
553 downstream without the limitation of extension length. Extended regions for EPA were
554 shortened at the DNA binding sites of a TF that was the closest to a TSS, and
555 transcriptional target genes were predicted from the shortened enhancer regions using
556 TFBS. Furthermore, promoter and extended regions for EPA were shortened at the
557 forward–reverse (FR) orientation of DNA binding sites of a TF. When forward or
558 reverse orientation of DNA binding sites were continuously located in genome
559 sequences several times, the most internal (i.e. closest to a TSS) forward–reverse
560 orientation of DNA binding sites were selected. The genomic positions of genes were

561 identified using ‘knownGene.txt.gz’ file in UCSC bioinformatics sites (Karolchik et al.
562 2014). The file ‘knownCanonical.txt.gz’ was also utilized for choosing representative
563 transcripts among various alternate forms for assigning promoter and extended regions
564 for EPA. From the list of transcription factor binding sequences and transcriptional
565 target genes, redundant transcription factor binding sequences were removed by
566 comparing the target genes of a transcription factor binding sequence and its
567 corresponding TF; if identical, one of the transcription factor binding sequences was
568 used. When the number of transcriptional target genes predicted from a transcription
569 factor binding sequence was less than five, the transcription factor binding sequence
570 was omitted.

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

575 For gene expression data, RNA-seq reads mapped onto human hg19 genome
576 sequences were obtained from UCSF-UBC human reference epigenome mapping
577 project RNA-seq reads with poly-A of naive $CD4^+$ T cells (GEO: GSM669617).
578 RNA-seq reads were aligned in repeat-masked hg19 genome sequences using BWA
579 with default parameters (Li and Durbin 2009). FPKMs of the RNA-seq data were
580 calculated using RSeQC (Wang et al. 2012b). For monocytes, Blueprint RNA-seq
581 FPKM data
582 (‘C0010KB1.transcript_quantification.rsem_grape2_crg.GRCh38.20150622.results’)

583 were downloaded from Blueprint DCC portal
584 (http://dcc.blueprint-epigenome.eu/#/files). RNA-seq reads with poly-A of human
585 mammary epithelial cells (HMEC) in Encyclopedia of DNA Elements at UCSC
586 (http://genome.ucsc.edu/encode/,
587 'wgEncodeCshlLongRnaSeqHmecCellPapAlnRep1.bam' file), H1 derived neural
588 progenitor cell (NPC) (GEO: GSM915326, Encode: ENCFF529SIO), Th17
589 (GSM2859479, NCBI SRA: SRR6298326), Treg (GSM2859476, SRR6298323), and
590 GM12878 (GSE78551, ENCFF297QCE) were used. Based on log2-transformed FPKM,
591 UCSC transcripts were arranged in descending order of expression level, and top 4,000
592 transcripts were selected in each cell type.

593 The expression level of transcriptional target genes predicted based on EPAs
594 shortened at the DNA motif sites of a TF was compared with the expression level of
595 transcriptional target genes predicted from closed chromatin regions in promoters. For
596 each DNA motif sequence shortening EPAs, transcriptional target genes were predicted
597 using DNA binding motif sequences of vertebrate TFs in TRANSFAC database, and the
598 ratio of expression level of putative transcriptional target genes of each TF was
599 calculated between EPAs and closed chromatin regions in promoters. The distribution of
600 the ratios of all TF was compared among forward-reverse (FR), reverse-forward (RF),
601 and any orientation (i.e. without considering orientation) of a DNA motif sites
602 shortening EPAs using Mann-Whitney test, two-sided (p-value $< 10^{-7}$). Other
603 parameters for the analysis were that the number of transcriptional target genes of a TF
604 was $\geq$50, and the number of TFs (not insulator TFs) to predict transcriptional target

605 genes was ≥50. The number of DNA motif sites of TFs (insulator TFs) shortening
606 EPAs was ≥100.
607
608 **Comparison with chromatin interaction data**
609 For comparison of EPIs with chromatin interactions (HiChIP) in $CD4^+$ T, Th17,
610 Treg and GM12878 cells,
611 'GSM2705049_Naive_HiChIP_H3K27ac_B2T1_allValidPairs.txt',
612 'GSM2705050_Naive_HiChIP_H3K27ac_B2T2_allValidPairs.txt',
613 'GSM2705051_Naive_HiChIP_H3K27ac_B3T1_allValidPairs.txt',
614 'GSM2705054_Th17_HiChIP_H3K27ac_B1T2_allValidPairs.txt',
615 'GSM2705055_Th17_HiChIP_H3K27ac_B2T1_allValidPairs.txt',
616 'GSM2705056_Th17_HiChIP_H3K27ac_B3T1_allValidPairs.txt',
617 'GSM2705059_Treg_HiChIP_H3K27ac_B3T1_allValidPairs.txt',
618 'GSM2705041_GM_HiChIP_H3K27ac_B1_allValidPairs.txt',
619 'GSM2705042_GM_HiChIP_H3K27ac_B2_allValidPairs.txt' files were downloaded
620 from Gene Expression Omnibus (GEO) database. The resolutions of chromatin
621 interactions and EPIs were adjusted to 5 kb before their comparison. Chromatin
622 interactions with ≥6,000 and ≥1,000 counts for each interaction were used in $CD4^+$ T
623 cells. According to the unique numbers of chromatin interactions in replications in Th17,
624 Treg and GM12878 cells, chromatin interactions with ≥1,000 (B1T2 in Th17),
625 ≥2,000 (B2T1 and B3T1 in Th17), ≥6,000 (B3T1 in Treg), ≥4,000 (B1 and B2 in
626 GM12878) counts for each interaction were analyzed.

627 Enhancer-promoter interactions (EPIs) were predicted using three types of EPAs
628 in monocytes: (i) EPA shortened at the FR or RF orientation of DNA motif sites of a TF,
629 (ii) EPA shortened at the any orientation (i.e. without considering orientation) of DNA
630 motif sites of a TF, and (iii) EPA without being shortened by a DNA motif site. EPIs
631 predicted in the three types of EPAs in common were removed. EPIs predicted from
632 EPA (i) were removed from EPIs predicted from EPA (ii). EPIs predicted from EPA (i)
633 and (ii) were removed from EPIs predicted from EPA (iii). The resolution of HiChIP
634 chromatin interaction data was 1-5 kb, so EPIs were adjusted to 5 kb before their
635 comparison. The number and ratio of EPIs overlapped with chromatin interactions were
636 compared two times between EPIs (i) and (iii), and EPIs (i) and (ii) (binomial
637 distribution, p-value < 0.025 for each test, two-sided, 95% confidence interval).
638 For comparison of EPIs with chromatin interactions (in situ Hi-C) in HMEC,
639 'HMEC_4DNFI97O9SAZ.pairs.gz' file was downloaded from 4D nucleome data portal
640 (http://data.4dnucleome.org). The genomic positions of chromatin interactions were
641 converted into hg19 version of human genome using liftOver tool
642 (http://hgdownload.soe.ucsc.edu/downloads.html#liftover). The resolutions of
643 chromatin interaction data and EPIs were adjusted to 3 kb before their comparison. All
644 chromatin interactions were used for analysis. For comparison of EPIs with chromatin
645 interactions (promoter capture Hi-C) in NPC, 'GSE86189_npc.po.all.txt.bz2' file was
646 downloaded from GEO database (GSE86189). The resolutions of chromatin interaction
647 data and EPIs were adjusted to 5 kb before their comparison. All chromatin interactions
648 were used for analysis.

649 Putative target genes for the comparison of EPIs and chromatin interactions were
650 selected from top 4,000 transcripts in term of the expression level. The expression level
651 of putative target genes of EPIs overlapped with chromatin interactions was compared
652 with EPIs not overlapped with them. When a putative transcriptional target gene of a TF
653 in an enhancer was found in both an EPI overlapped with a chromatin interaction and an
654 EPI not overlapped with, the target gene was eliminated. The distribution of expression
655 level of putative target genes was compared using Mann-Whitney test, two-sided
656 (p-value < 0.05).
657
658 **Correlation of expression level of gene pairs separated by DNA motif sites of a TF**
659 To examine the correlation of expression level of genes pairs separated by DNA
660 motif sites of a TF, gene expression data among 53 tissues were obtained from
661 'GTEx_Analysis_2016-01-15_v7_RNASeQCv1.1.8_gene_median_tpm.gct.gz' file in
662 GTEx database (https://gtexportal.org/home/). The correlation of log2-transformed gene
663 expression level was analyzed based on protein-coding transcripts of RefSeq (RefSeq
664 ID with 'NM_'). The same genomic location of RefSeq transcripts were removed. DNA
665 motif sites overlapped with RefSeq transcripts were eliminated. Genes closest to DNA
666 motif sites and other genes were analyzed using BEDOPS closest-features (Neph et al.
667 2012). Mann-Whitney test was performed using R.
668
669 **Co-location of biased orientations of DNA motif sites of TFs**
670 Co-location of biased orientations of DNA binding motif sites of TFs was

671 examined. The number of genomic regions with the same pair of DNA motif sites of
672 biased orientations of DNA motifs of TFs in upstream and downstream of genes in
673 EPAs was counted, and the same pair of DNA motif sites of TFs found in ≥100
674 genomic regions was listed.
675

676 **Acknowledgements**

677 The supercomputing resource was provided by Human Genome Center of the Institute
678 of Medical Science at the University of Tokyo. Computations were partially performed
679 on the NIG supercomputer at ROIS National Institute of Genetics. Publication charges
680 for this article were funded by JSPS KAKENHI Grant Number 16K00387. This
681 research was partially supported by the Platform Project for Supporting in Drug
682 Discovery and Life Science Research（Platform for Dynamic Approaches to Living
683 System）from Japan Agency for Medical Research and Development (AMED). This
684 research was partially supported by Development of Fundamental Technologies for
685 Diagnosis and Therapy Based upon Epigenome Analysis from Japan Agency for
686 Medical Research and Development (AMED). This work was partially supported by
687 JST CREST Grant Number JPMJCR15G1, Japan.
688

689 **Figures**

690 Web links to original figures.

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

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

693 76_2014_BFnrg3682_MOESM18_ESM.ppt

694 Graphical Abstract (de Wit et al. 2015)

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

696 92c5eba1d/fx1.jpg

697

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

699 2010)

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

701 87_2010_BFnbt1630_MOESM11_ESM.pdf

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

703 DNA Looping (Guo et al. 2015)

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

705 2af081251/gr4.jpg

706

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

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

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

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

711 2014; de Wit et al. 2015; Guo et al. 2015). (B) Computationally-defined regulatory
712 domains for enhancer-promoter association (McLean et al. 2010). The *single nearest*
713 *gene* association rule extends the regulatory domain to the midpoint between this gene's
714 TSS and the nearest gene's TSS both upstream and downstream. Enhancer-promoter
715 association was shortened at the forward-reverse orientation of DNA binding motif sites
716 of a transcription factor (e.g. CTCF in the figure).
717

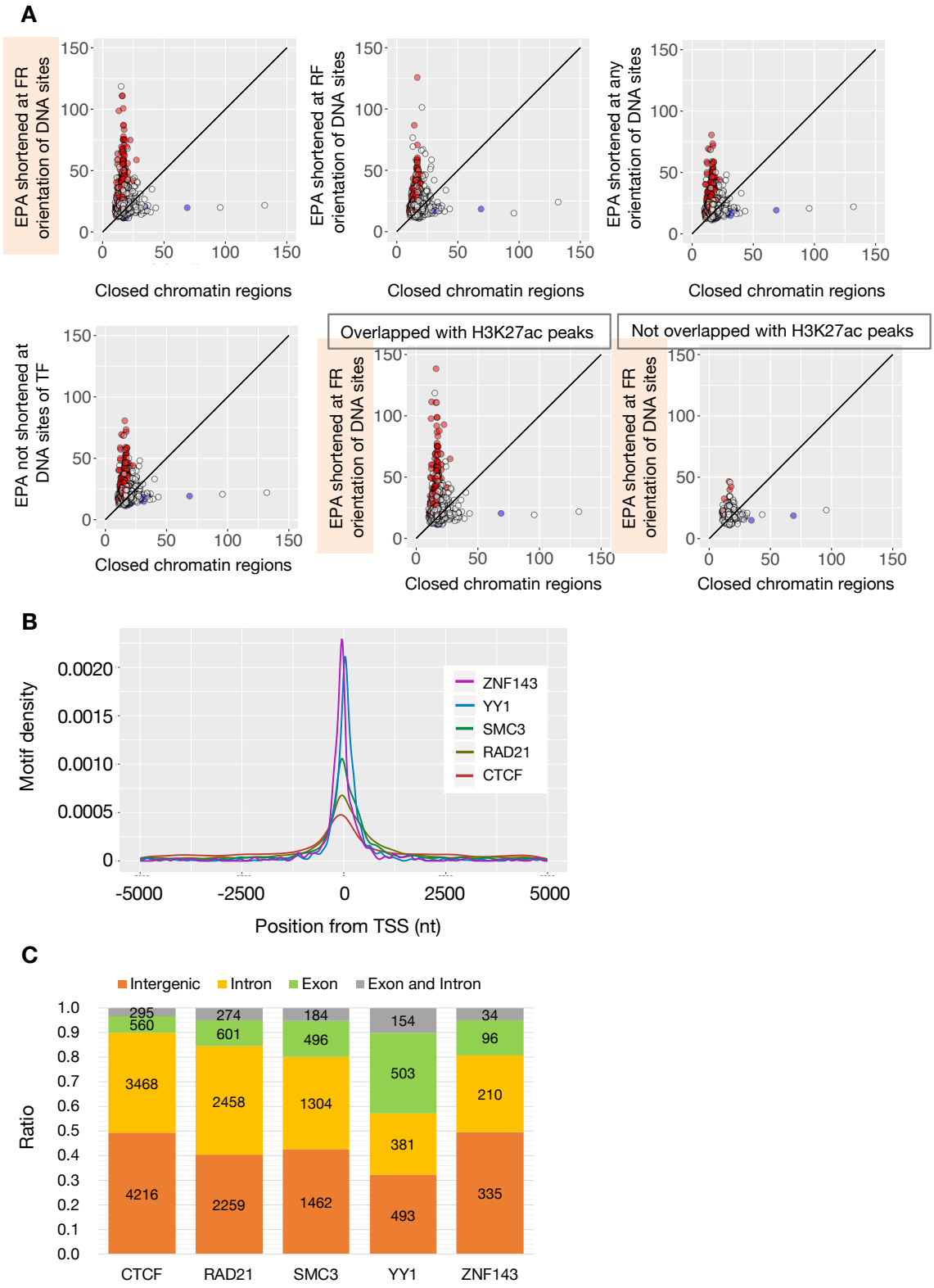

Figure 2. Identification of DNA motif sites of TFs from open chromatin regions and properties of DNA motif sites. (A) Comparison of expression level of putative

721 transcriptional target genes. Each dot shows the median expression level (FPKM) of
722 target genes of the same transcription factor binding sites in closed chromatin regions
723 (or promoters) (X-axis) and enhancer-promoter association (EPA) shortened at the
724 forward-reverse orientation of DNA motif sites of CTCF (Y-axis) in T cell (upper left
725 graph). The distribution of expression level was changed according to four criteria of
726 EPAs (four graphs) (see Method) (Osato 2018). Red dots show the median expression
727 level of target genes was higher significantly (Mann-Whitney test) in EPA than closed
728 chromatin regions, and blue dots show the median expression level of target genes was
729 lower significantly in EPA than closed chromatin regions. The expression level of target
730 genes predicted in EPA tend to be higher than closed chromatin regions, implying that
731 TFs acted as activators of target genes. DNA motif sites of TFs overlapped with
732 H3K27ac ChIP-seq peaks showed significant differences of expression level of target
733 genes, and DNA motif sites not overlapped with H3K27ac peaks did not show
734 significantly (lower right two graphs). This suggested that the significant differences of
735 expression level were associated with enhancer (and promoter) signals of H3K27ac. (B)
736 Distribution of DNA motif sites of TFs near TSS. ZNF143 and YY1 tend to be located
737 near TSS, and CTCF tend to be observed more from distant regions than around TSS in
738 monocyte. (C) Distribution of DNA motif sites of TFs in the human genome. Most of
739 DNA motif sites were found from intergenic and intron regions. YY1 is known to be at
740 TSS in the edge of exons of genes, and is associated with transcriptional activation,
741 repression and initiation. YY1 is multi-functional transcriptional regulator (Shrivastava
742 and Calame 1994; Xi et al. 2007)

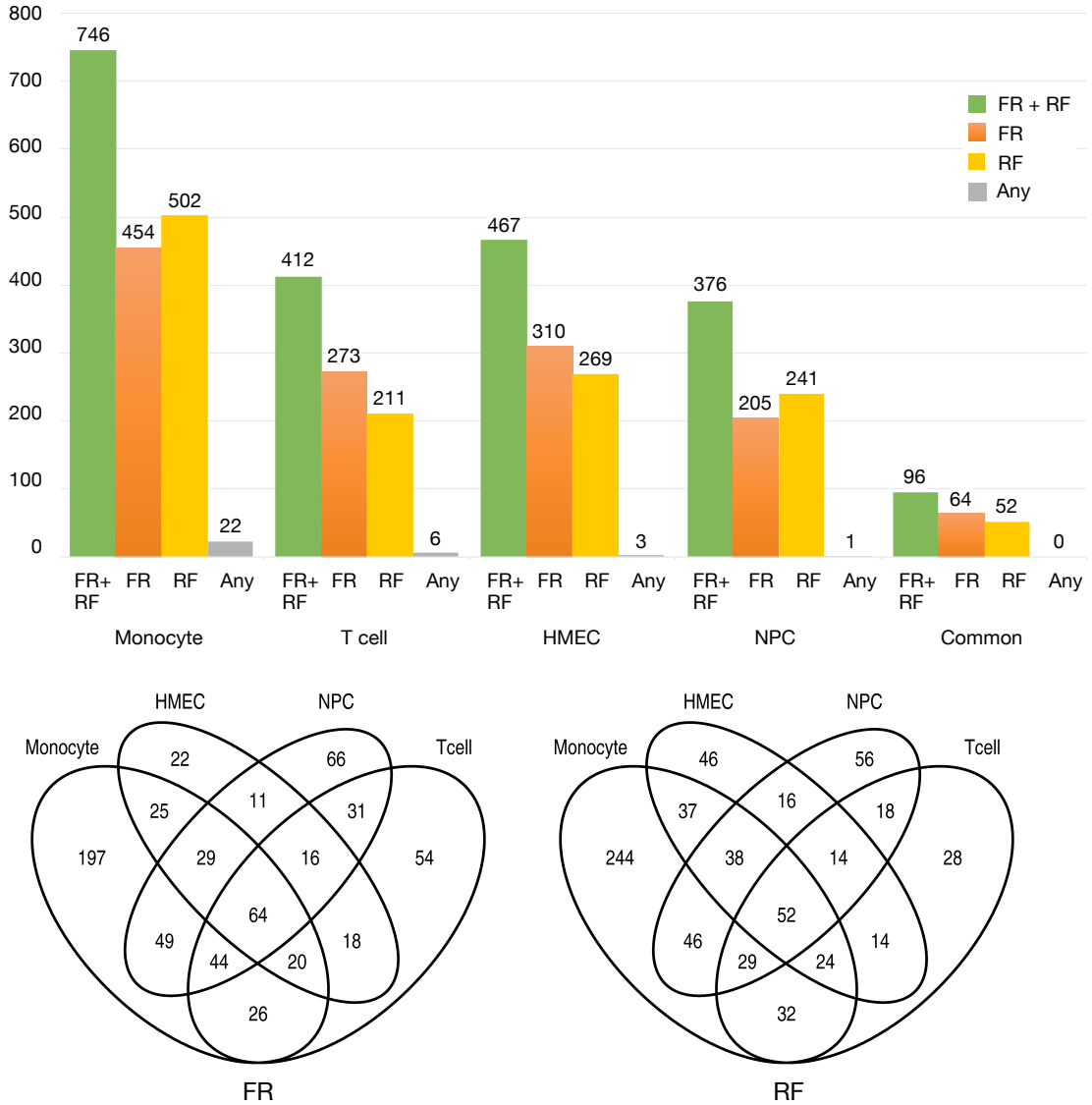

Figure 3. Biased orientations of DNA motif sequences of TFs affecting the transcription of genes. Total 96 (64 FR and 52 RF) of biased orientations of DNA binding motif sequences of TFs were found to affect significantly the expression level of putative transcriptional target genes in monocytes, T cells,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

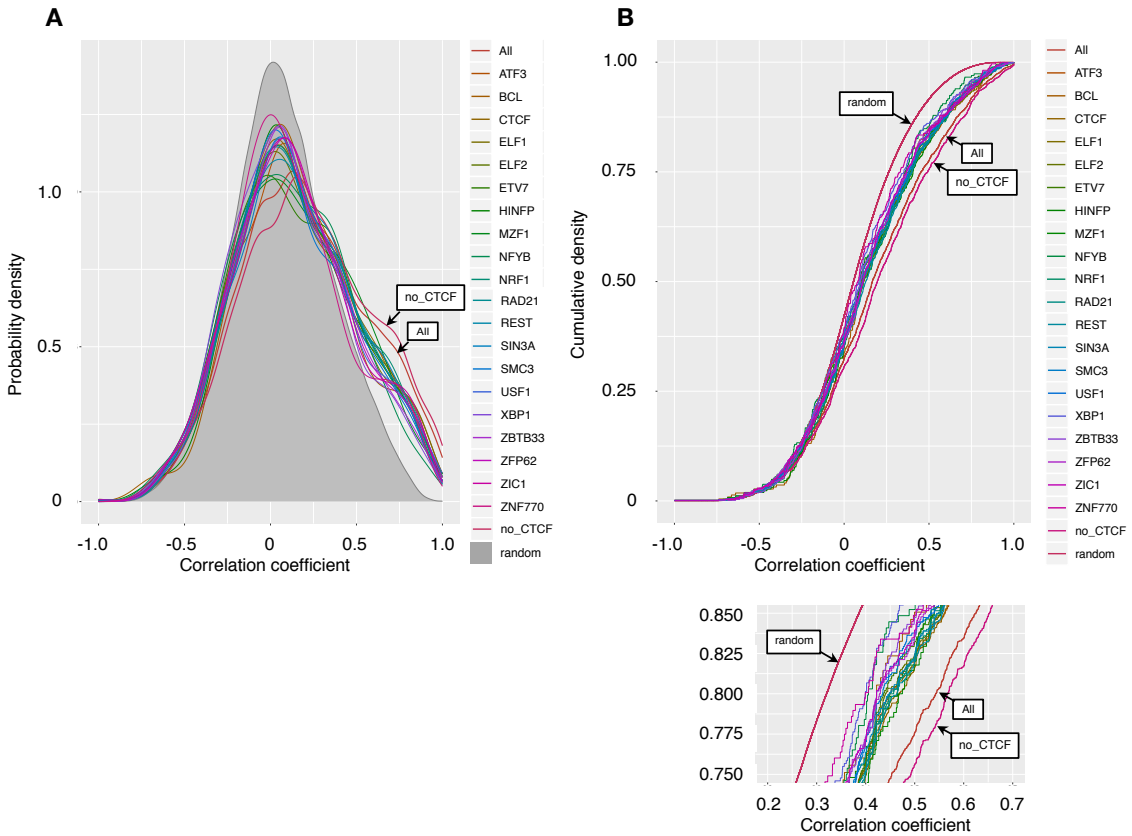

Figure 4. Genes separated by predicted DNA binding sites of biased orientations of DNA motifs of TFs are less correlated in gene expression. Top 20 TFs are shown according to ascending order of median correlation coefficient. Correlation coefficient between neighboring gene pairs is shown in terms of 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.

763 Web links to original figures.

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

765 (Bailey et al. 2015).

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

767

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

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

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

771 2015).

772

773

Tables

Table 1. Biased orientations of DNA binding motif sequences of TFs 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

781 **Table 2.** Comparison of enhancer-promoter interactions (EPIs) with chromatin
782 interactions in T cells. The upper tables show the number of biased orientations of DNA
783 motifs, where a significantly higher ratio of EPIs predicted based on enhancer-promoter
784 association (EPA) (i) overlapped with HiChIP chromatin interactions than the other
785 types of EPA (ii) and (iii). The middle tables show 133 FR and 90 RF orientations of
786 DNA motifs in T cells found in common among B2T1, B2T2 and B3T1 replications in
787 the upper table. The lower tables show that among 64 FR and 52 RF biased orientations
788 of DNA motifs found in common in monocytes, T cell, HMEC and NPC, 44 (69%) FR
789 and 28 (54%) RF were matched with the analysis of HiChIP chromatin interactions for
790 three types of EPAs. TF: DNA binding motif sequence of a transcription factor. Score: –
791 $\log_{10}$ (*p*-value). Score 1: Comparison of EPAs shortened at the FR or RF orientation of
792 DNA motif sites [EPA (i)] with EPA not shortened [EPA (iii)]. Score 2: Comparison of
793 EPAs shortened at the FR or RF orientation of DNA motif sites [EPA (i)] with EPAs
794 shortened at the any 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

795

Forward-reverse (FR)

TF	Score 1	Score 2
AHR	38.81	13.41
ARNTL	155.78	1.96
ATF1	323.31	15.00
ATF3	323.31	35.22
BCL	153.72	3.52
CREB	90.91	2.32
CREB3	323.31	38.74
CREB3L1	323.31	27.96
CTCF	323.31	3.50
CTCFL	174.90	2.09
CXXC1	69.92	3.30
E2	323.31	20.15
E2F	122.12	21.87
E2F1	43.22	2.85
EGR1	79.09	3.79
EGR2	176.69	4.53
EGR3	85.25	34.18
EHF	323.31	8.66
ELF1	203.38	3.86
ELF2	227.62	7.83
ELF5	323.31	8.76
ELK1:TBX21	126.46	8.41
ELK1:TEF	323.31	7.99
ELK4	117.83	4.81
ENO1	80.42	3.13
EPAS1	65.43	7.76
ERF	207.67	6.25
ERG	199.81	25.70
ETS	196.80	1.63
ETS1	308.00	3.92

TF	Score 1	Score 2
ETS2	208.78	13.99
ETV1	153.32	7.02
ETV3	140.11	14.58
ETV5:EVX1	145.48	17.33
ETV5:HES7	8.76	8.74
ETV5:TCF3	203.52	5.90
FOXN4	73.43	5.59
FOXO1:ELK3	323.31	43.51
GABPA	197.57	20.39
GABPA:GABPB1	142.28	6.10
GCM2	13.68	18.80
GDNF	286.73	3.15
GMEB1	136.68	10.56
HIC1	223.55	4.56
HIF1A	241.47	1.71
HOXB2:ETV1	102.22	2.42
ID2	323.31	14.52
INSM1	193.85	15.87
JUN:JUNB	275.78	6.48
JUNB:JUN	202.73	5.07
KLF	252.10	8.35
KLF15	323.31	22.56
KLF17	87.51	14.77
KLF2	90.74	15.34
KLF4	209.71	9.38
KLF7	229.95	11.96
LHX8	244.38	2.95
MAX	323.31	8.35
MAZ	109.42	3.37
MBD2	8.41	1.95

TF	Score 1	Score 2
MNT	215.63	8.00
MTF1	323.31	14.21
MXI1	323.31	70.98
MYC	84.79	24.93
MYCN	165.16	25.87
MZF1	81.83	2.29
NFKB1	279.24	14.27
NFKB1:NFKB1	323.31	13.79
NFY	323.31	3.42
NONO	323.31	9.24
NR2C2	323.31	3.64
NRF1	45.13	30.75
OVOL2	323.31	31.61
PATZ1	101.85	9.30
PLAG1	177.90	11.02
PURA	196.04	17.98
R	22.49	3.99
RAD21	258.21	4.81
RBAK	267.43	11.25
RELA	323.31	4.68
RFX1	159.78	17.05
RFX4	323.31	35.28
RFX5	323.31	26.52
SAP1	124.23	14.63
SMC3	222.09	1.76
SP1	235.48	29.09
SP2	94.11	9.54
SP4	14.30	6.31
SPDEF	179.39	16.81
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	TF	Score 1	Score 2	TF	Score 1	Score 2
ARNT	323.31	14.12	MAX	304.34	8.64	TFE3	95.23	10.59
ARNT2	41.41	42.78	MLXIPL	64.44	10.74	THAP1	127.72	5.07
ASCL2	138.15	6.41	MUSCLE	23.05	3.47	**YY1**	323.31	9.92
ATF3	323.31	14.64	MYB	323.31	17.68	YY2	47.49	2.61
ATF4	323.31	25.14	MYBL1	316.23	23.27	ZBTB33	90.29	7.78
ATF7	79.30	4.96	MYC	323.31	12.23	ZBTB45	323.31	57.08
BCLAF1	148.24	8.52	MZF1	200.49	22.73	ZBTB7A	89.19	16.19
BHLHE40	323.31	12.42	NFY	209.97	3.22	ZBTB7B	43.90	23.59
CDCA7L	84.72	8.42	NFYA	323.31	11.75	ZFP64	147.00	13.59
CREB3	206.55	11.07	NHLH1	105.19	4.42	ZIC3	42.65	15.39
CREB3L1	323.31	7.63	NR1H3	323.31	9.38	ZNF100	27.67	4.75
CREB3L2	67.51	7.98	NR3C1	43.16	8.31	**ZNF143**	63.03	9.69
CTCF	240.61	4.56	OSR2	111.77	14.65	ZNF148	213.77	18.93
E2F	78.16	7.07	RFX1	323.31	22.99	ZNF16	323.31	47.43
E2F1	124.10	15.73	RXRA	146.30	15.11	ZNF37A	310.67	14.79
E2F3	65.67	8.85	SALL1	277.84	1.91	ZNF410	140.45	13.57
E2F6	165.75	30.69	SCRT2	301.48	9.60	ZNF419	247.63	11.25
EFNA2	255.01	10.95	SMAD6	259.24	7.43	ZNF429	87.05	32.39
ELK3	143.48	7.57	SP100	20.75	7.99	ZNF436	323.31	11.48
EP300	121.01	10.90	SP2	294.86	2.59	ZNF468	21.12	7.67
ETV5:CLOCK	42.10	7.29	SP4	110.94	19.63	ZNF557	299.44	9.58
GLIS2	37.40	17.34	SPDEF	185.85	14.35	ZNF575	39.45	14.38
GMEB2	323.31	7.81	SPZ1	323.31	9.87	ZNF586	260.73	6.18
HES1	88.18	9.46	SREBF1	67.27	15.74	ZNF611	279.80	15.65
HES7	166.52	14.60	SREBF2	297.27	1.62	ZNF616	189.53	30.98
HIVEP1	323.31	4.90	TAF1	252.27	17.47	ZNF692	202.47	2.33
HOXB2:ELK1	217.94	11.92	TATA	2.16	4.02	ZNF727	214.01	17.22
IRC900814	323.31	4.89	TFAP2A	20.17	29.15	ZNF770	300.52	4.71
KLF4	213.35	9.49	TFAP2B	29.46	12.62	ZNF773	150.45	8.84
KLF8	323.31	17.42	TFAP2C	112.10	51.21	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 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 TF separating gene pairs with expression data. Ubiquitously expressed genes with

817 coefficient of variance < 90 and DNA motif sites nearby were not used for statistical test,
818 so the number of DNA motif sites was relatively small. The results using all genes were
819 shown in Supplemental material. All DNA motif sites of a TF were used including DNA
820 motif sites examined for the biased orientations of DNA motif sites of a TF in EPAs.
821 The results using only the biased orientations of DNA motif sites of a TF were shown in
822 Supplemental material.
823

824 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

825

826 **Table 4.** Top 30 of co-location and overlapping biased orientations of DNA binding
827 motif sequences of TFs in monocytes. Co-location and overlapping DNA motif
828 sequences of CTCF with another biased orientation of DNA motif sequence were shown
829 in separate tables. Motif 1,2: DNA binding motif sequences of TFs. No.: the number of
830 genomic locations with both Motif 1 and Motif 2 in upstream and downstream of genes
831 in EPAs (< 1 Mb distance from TSS of genes). Total number of genes is 30,523.

832

833 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

834

835

836

837

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

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

847 **References**

848

849 Ackerman SL, Minden AG, Williams GT, Bobonis C, Yeung CY. 1991. Functional
850 significance of an overlapping consensus binding motif for Sp1 and Zif268 in
851 the murine adenosine deaminase gene promoter. *Proc Natl Acad Sci U S A* **88**:
852 7523-7527.
853 Ansari AZ, Peterson-Kaufman KJ. 2011. A partner evokes latent differences between
854 Hox proteins. *Cell* **147**: 1220-1221.
855 Bailey SD, Zhang X, Desai K, Aid M, Corradin O, Cowper-Sal Lari R, Akhtar-Zaidi B,
856 Scacheri PC, Haibe-Kains B, Lupien M. 2015. ZNF143 provides sequence
857 specificity to secure chromatin interactions at gene promoters. *Nat Commun* **2**:
858 6186.
859 Barutcu AR, Lajoie BR, Fritz AJ, McCord RP, Nickerson JA, van Wijnen AJ, Lian JB,
860 Stein JL, Dekker J, Stein GS et al. 2016. SMARCA4 regulates gene expression
861 and higher-order chromatin structure in proliferating mammary epithelial cells.
862 *Genome research* **26**: 1188-1201.
863 Chatterjee R, Zhao J, He X, Shlyakhtenko A, Mann I, Waterfall JJ, Meltzer P,
864 Sathyanarayana BK, FitzGerald PC, Vinson C. 2012. Overlapping ETS and CRE
865 Motifs ((G/C)CGGAAGTGACGTCA) preferentially bound by GABPalpha and
866 CREB proteins. *G3 (Bethesda)* **2**: 1243-1256.
867 Chen M, Manley JL. 2009. Mechanisms of alternative splicing regulation: insights from
868 molecular and genomics approaches. *Nat Rev Mol Cell Biol* **10**: 741-754.
869 Das D, Clark TA, Schweitzer A, Yamamoto M, Marr H, Arribere J, Minovitsky S,
870 Poliakov A, Dubchak I, Blume JE et al. 2007. A correlation with exon
871 expression approach to identify cis-regulatory elements for tissue-specific
872 alternative splicing. *Nucleic acids research* **35**: 4845-4857.
873 de Wit E, Vos ES, Holwerda SJ, Valdes-Quezada C, Verstegen MJ, Teunissen H,
874 Splinter E, Wijchers PJ, Krijger PH, de Laat W. 2015. CTCF Binding Polarity
875 Determines Chromatin Looping. *Molecular cell* **60**: 676-684.
876 Guo Y, Xu Q, Canzio D, Shou J, Li J, Gorkin DU, Jung I, Wu H, Zhai Y, Tang Y et al.
877 2015. CRISPR Inversion of CTCF Sites Alters Genome Topology and
878 Enhancer/Promoter Function. *Cell* **162**: 900-910.

879 He X, Syed KS, Tillo D, Mann I, Weirauch MT, Vinson C. 2015. GABPalpha Binding
880 to Overlapping ETS and CRE DNA Motifs Is Enhanced by CREB1: Custom
881 DNA Microarrays. *G3 (Bethesda)* **5**: 1909-1918.
882 Hsu SC, Gilgenast TG, Bartman CR, Edwards CR, Stonestrom AJ, Huang P, Emerson
883 DJ, Evans P, Werner MT, Keller CA et al. 2017. The BET Protein BRD2
884 Cooperates with CTCF to Enforce Transcriptional and Architectural Boundaries.
885 *Molecular cell* **66**: 102-116 e107.
886 Ji X, Dadon DB, Abraham BJ, Lee TI, Jaenisch R, Bradner JE, Young RA. 2015.
887 Chromatin proteomic profiling reveals novel proteins associated with
888 histone-marked genomic regions. *Proc Natl Acad Sci U S A* **112**: 3841-3846.
889 Jolma A, Yan J, Whitington T, Toivonen J, Nitta KR, Rastas P, Morgunova E, Enge M,
890 Taipale M, Wei G et al. 2013. DNA-binding specificities of human transcription
891 factors. *Cell* **152**: 327-339.
892 Jolma A, Yin Y, Nitta KR, Dave K, Popov A, Taipale M, Enge M, Kivioja T, Morgunova
893 E, Taipale J. 2015. DNA-dependent formation of transcription factor pairs alters
894 their binding specificity. *Nature* **527**: 384-388.
895 Jung I, Schmitt A, Diao Y, Lee AJ, Liu T, Yang D, Tan C, Eom J, Chan M, Chee S et al.
896 2019. A compendium of promoter-centered long-range chromatin interactions in
897 the human genome. *Nature genetics* **51**: 1442-1449.
898 Karolchik D, Barber GP, Casper J, Clawson H, Cline MS, Diekhans M, Dreszer TR,
899 Fujita PA, Guruvadoo L, Haeussler M et al. 2014. The UCSC Genome Browser
900 database: 2014 update. *Nucleic acids research* **42**: D764-770.
901 Katainen R, Dave K, Pitkanen E, Palin K, Kivioja T, Valimaki N, Gylfe AE, Ristolainen
902 H, Hanninen UA, Cajuso T et al. 2015. CTCF/cohesin-binding sites are
903 frequently mutated in cancer. *Nature genetics* **47**: 818-821.
904 Kel AE, Gossling E, Reuter I, Cheremushkin E, Kel-Margoulis OV, Wingender E. 2003.
905 MATCH: A tool for searching transcription factor binding sites in DNA
906 sequences. *Nucleic acids research* **31**: 3576-3579.
907 Kheradpour P, Kellis M. 2014. Systematic discovery and characterization of regulatory
908 motifs in ENCODE TF binding experiments. *Nucleic acids research* **42**:
909 2976-2987.
910 Kin K, Chen X, Gonzalez-Garay M, Fakhouri WD. 2016. The effect of non-coding
911 DNA variations on P53 and cMYC competitive inhibition at cis-overlapping

912 motifs. *Hum Mol Genet* **25**: 1517-1527.
913 Kulakovskiy IV, Vorontsov IE, Yevshin IS, Sharipov RN, Fedorova AD, Rumynskiy EI,
914 Medvedeva YA, Magana-Mora A, Bajic VB, Papatsenko DA et al. 2018.
915 HOCOMOCO: towards a complete collection of transcription factor binding
916 models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic acids*
917 *research* **46**: D252-D259.
918 Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler
919 transform. *Bioinformatics (Oxford, England)* **25**: 1754-1760.
920 Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A,
921 Amit I, Lajoie BR, Sabo PJ, Dorschner MO et al. 2009. Comprehensive
922 mapping of long-range interactions reveals folding principles of the human
923 genome. *Science (New York, NY)* **326**: 289-293.
924 McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM,
925 Bejerano G. 2010. GREAT improves functional interpretation of cis-regulatory
926 regions. *Nature biotechnology* **28**: 495-501.
927 Merabet S, Mann RS. 2016. To Be Specific or Not: The Critical Relationship Between
928 Hox And TALE Proteins. *Trends Genet* **32**: 334-347.
929 Mumbach MR, Satpathy AT, Boyle EA, Dai C, Gowen BG, Cho SW, Nguyen ML,
930 Rubin AJ, Granja JM, Kazane KR et al. 2017. Enhancer connectome in primary
931 human cells identifies target genes of disease-associated DNA elements. *Nature*
932 *genetics* **49**: 1602-1612.
933 Neph S, Kuehn MS, Reynolds AP, Haugen E, Thurman RE, Johnson AK, Rynes E,
934 Maurano MT, Vierstra J, Thomas S et al. 2012. BEDOPS: high-performance
935 genomic feature operations. *Bioinformatics (Oxford, England)* **28**: 1919-1920.
936 Newburger DE, Bulyk ML. 2009. UniPROBE: an online database of protein binding
937 microarray data on protein-DNA interactions. *Nucleic acids research* **37**:
938 D77-82.
939 Osato N. 2018. Characteristics of functional enrichment and gene expression level of
940 human putative transcriptional target genes. *BMC Genomics* **19**: 957.
941 Phanstiel DH, Van Bortle K, Spacek D, Hess GT, Shamim MS, Machol I, Love MI,
942 Aiden EL, Bassik MC, Snyder MP. 2017. Static and Dynamic DNA Loops form
943 AP-1-Bound Activation Hubs during Macrophage Development. *Molecular cell*
944 **67**: 1037-1048.e1036.

945 Portales-Casamar E, Thongjuea S, Kwon AT, Arenillas D, Zhao X, Valen E, Yusuf D,
946 Lenhard B, Wasserman WW, Sandelin A. 2010. JASPAR 2010: the greatly
947 expanded open-access database of transcription factor binding profiles. *Nucleic*
948 *acids research* **38**: D105-110.
949 Purmann A, Toedling J, Schueler M, Carninci P, Lehrach H, Hayashizaki Y, Huber W,
950 Sperling S. 2007. Genomic organization of transcriptomes in mammals:
951 Coregulation and cofunctionality. *Genomics* **89**: 580-587.
952 Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn
953 AL, Machol I, Omer AD, Lander ES et al. 2014. A 3D map of the human
954 genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**:
955 1665-1680.
956 Schreiber J, Libbrecht M, Bilmes J, Noble W. 2017. Nucleotide sequence and DNaseI
957 sensitivity are predictive of 3D chromatin architecture. *bioRxiv*.
958 Sherwood RI, Hashimoto T, O'Donnell CW, Lewis S, Barkal AA, van Hoff JP, Karun V,
959 Jaakkola T, Gifford DK. 2014. Discovery of directional and nondirectional
960 pioneer transcription factors by modeling DNase profile magnitude and shape.
961 *Nature biotechnology* **32**: 171-178.
962 Shlyueva D, Stampfel G, Stark A. 2014. Transcriptional enhancers: from properties to
963 genome-wide predictions. *Nature reviews Genetics* **15**: 272-286.
964 Shrivastava A, Calame K. 1994. An analysis of genes regulated by the multi-functional
965 transcriptional regulator Yin Yang-1. *Nucleic acids research* **22**: 5151-5155.
966 Slattery M, Riley T, Liu P, Abe N, Gomez-Alcala P, Dror I, Zhou T, Rohs R, Honig B,
967 Bussemaker HJ et al. 2011. Cofactor binding evokes latent differences in DNA
968 binding specificity between Hox proteins. *Cell* **147**: 1270-1282.
969 Tabuchi TM, Deplancke B, Osato N, Zhu LJ, Barrasa MI, Harrison MM, Horvitz HR,
970 Walhout AJ, Hagstrom KA. 2011. Chromosome-biased binding and gene
971 regulation by the Caenorhabditis elegans DRM complex. *PLoS Genet* **7**:
972 e1002074.
973 van Helden J. 2003. Regulatory sequence analysis tools. *Nucleic acids research* **31**:
974 3593-3596.
975 Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF,
976 Schroth GP, Burge CB. 2008. Alternative isoform regulation in human tissue
977 transcriptomes. *Nature* **456**: 470-476.

978 Wang J, Zhuang J, Iyer S, Lin X, Whitfield TW, Greven MC, Pierce BG, Dong X,
979 Kundaje A, Cheng Y et al. 2012a. Sequence features and chromatin structure
980 around the genomic regions bound by 119 human transcription factors. *Genome*
981 *research* **22**: 1798-1812.
982 Wang L, Wang S, Li W. 2012b. RSeQC: quality control of RNA-seq experiments.
983 *Bioinformatics (Oxford, England)* **28**: 2184-2185.
984 Wei Z, Gao F, Kim S, Yang H, Lyu J, An W, Wang K, Lu W. 2013. Klf4 organizes
985 long-range chromosomal interactions with the oct4 locus in reprogramming and
986 pluripotency. *Cell Stem Cell* **13**: 36-47.
987 Weintraub AS, Li CH, Zamudio AV, Sigova AA, Hannett NM, Day DS, Abraham BJ,
988 Cohen MA, Nabet B, Buckley DL et al. 2017. YY1 Is a Structural Regulator of
989 Enhancer-Promoter Loops. *Cell* **171**: 1573-1588 e1528.
990 Wingender E, Dietze P, Karas H, Knuppel R. 1996. TRANSFAC: a database on
991 transcription factors and their DNA binding sites. *Nucleic acids research* **24**:
992 238-241.
993 Xi H, Yu Y, Fu Y, Foley J, Halees A, Weng Z. 2007. Analysis of overrepresented motifs
994 in human core promoters reveals dual regulatory roles of YY1. *Genome research*
995 **17**: 798-806.
996 Xie X, Mikkelsen TS, Gnirke A, Lindblad-Toh K, Kellis M, Lander ES. 2007.
997 Systematic discovery of regulatory motifs in conserved regions of the human
998 genome, including thousands of CTCF insulator sites. *Proc Natl Acad Sci U S A*
999 **104**: 7145-7150.
1000 Xie Z, Hu S, Blackshaw S, Zhu H, Qian J. 2010. hPDI: a database of experimental
1001 human protein-DNA interactions. *Bioinformatics (Oxford, England)* **26**:
1002 287-289.
1003 Yardimci GG, Ozadam H, Sauria MEG, Ursu O, Yan KK, Yang T, Chakraborty A, Kaul
1004 A, Lajoie BR, Song F et al. 2019. Measuring the reproducibility and quality of
1005 Hi-C data. *Genome Biol* **20**: 57.
1006 Yoon SO, Chikaraishi DM. 1992. Tissue-specific transcription of the rat tyrosine
1007 hydroxylase gene requires synergy between an AP-1 motif and an overlapping E
1008 box-containing dyad. *Neuron* **9**: 55-67.
1009 Zhang H, Li F, Jia Y, Xu B, Zhang Y, Li X, Zhang Z. 2017. Characteristic arrangement
1010 of nucleosomes is predictive of chromatin interactions at kilobase resolution.

1011 *Nucleic acids research* **45**: 12739-12751.
1012 Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C,
1013 Myers RM, Brown M, Li W et al. 2008. Model-based analysis of ChIP-Seq
1014 (MACS). *Genome Biol* **9**: R137.
1015 Zhao Y, Stormo GD. 2011. Quantitative analysis demonstrates most transcription factors
1016 require only simple models of specificity. *Nature biotechnology* **29**: 480-483.
1017