Discovery of biased orientations of regulatory motifs affecting transcription of human 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 play essential roles in enhancer-promoter interactions (EPIs) and regulating the transcription of genes. CTCF and cohesin proteins are located at the anchors of chromatin interactions, forming their loop structures. Other DNA binding proteins such as YY1, ZNF143 and SMARCA4 are also reported to be involved in chromatin interactions. However, the overall picture of proteins associated with chromatin interactions and insulator function is still unclear. Here I describe a systematic and comprehensive approach to discover DNA binding motif sequences of transcription factors (TFs), affecting the interaction between enhancers and promoters of genes and their expression by the insulator function of the TFs. This analysis identified 96 biased orientations of DNA motifs of TFs that affected the expression level of putative transcriptional target genes significantly in monocytes, T cells, HMEC and NPC in common, and included known TFs involved in chromatin interactions and insulator function such as CTCF, cohesin (RAD21 and SMC3), YY1 and ZNF143. The 96 biased orientations of DNA motifs contain 64 forward-reverse (FR) and 52 reverse-forward (RF) orientations of DNA motifs (Some TFs have more than one DNA motifs in databases, so the total number is smaller than the sum of FR and RF). The merger of DNA motif sites of biased orientation of a TF and CTCF showed a higher expression level of putative transcriptional target genes than DNA motif sites of only CTCF. To confirm the predicted DNA motifs, first, I compared with chromatin interaction data. Despite the limitation of resolution of chromatin interaction data, for 44 (69%) FR and 28 (54%) RF orientations of DNA motif sequences in $CD4^+$ T cells,

EPIs predicted based on an enhancer-promoter association (EPA) that was shortened at the biased orientations of DNA motif sites, overlapped with chromatin interactions significantly more than other types of EPAs. Second, using gene expression data among 53 tissues, 43 (72%) FR and 40 (80%) RF orientations of DNA motifs showed significantly reduced correlation in expression level of nearby genes separated by the motif sites. Some DNA motif sites of biased orientations of DNA motifs of TFs were found between nearby genes not including DNA motif sites of the known TFs involved in chromatin interactions and between low correlation of expression level of nearby genes. These analyses suggest that the DNA motifs are associated with insulator function.

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

Background

Chromatin interactions play essential roles in enhancer-promoter interactions (EPIs) and regulating the transcription of genes. CTCF and cohesin proteins are located at the anchors of chromatin interactions, forming their loop structures. CTCF has insulator function limiting the activity of enhancers into the loops (Fig. 1A). The DNA binding sequence of CTCF indicate its orientation bias at chromatin interaction anchors – forward-reverse (FR) orientation is frequently observed (Rao et al. 2014; de Wit et al.

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

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

72 DNA binding proteins involved in chromatin interactions are considered to affect
73 the transcription of genes in the loops formed by chromatin interactions. To confirm this
74 notion, I previously found that the expression level of human putative transcriptional
75 target genes of a TF was significantly changed, according to the criteria of
76 enhancer-promoter association (EPA) (Osato 2018). An EPA was shortened at the FR

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

91 Experimental analyses of chromatin interactions in molecular biology have
92 technical limitations: chromatin interaction data are altered, according to experimental
93 techniques, the number of reads of DNA sequencing and even replication sets of the
94 same cell type. Chromatin interaction data may not be saturated enough to contain all
95 chromatin interactions and may include experimental noises such as ligation error.
96 Avoiding the technical limitations of experimental analyses, here I searched
97 computationally for DNA motif sequences of TFs, affecting EPI and the expression
98 level of putative transcriptional target genes in monocytes, T cells, HMEC and NPC

99 without using chromatin interaction data. Then, putative EPIs and DNA motif sites of
100 TFs were compared with chromatin interaction and gene expression data.
101

102 ## Results

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

104 Transcription factor binding sites (TFBSs) were predicted using open chromatin
105 regions and DNA motif sequences of transcription factors (TF) collected from various
106 databases and journal papers (see Methods). Transcriptional target genes were predicted
107 based on TFBS in promoters and enhancers. An enhancer-promoter association (EPA)
108 was shortened at the DNA binding motif sites of a TF acting as insulator such as CTCF
109 and cohesin (RAD21 and SMC3) (Fig. 1B). To find DNA motif sequences of TFs with
110 insulator function, other than CTCF and cohesin, affecting the expression level of genes,
111 An EPA was shortened at the DNA motif sites of a putative insulator TF, and
112 transcriptional target genes of each general TF bound in (i) distal open chromatin
113 regions (enhancers) and (ii) closed chromatin regions in promoters were predicted based
114 on the EPA, in order to estimate the enhancer activity of each general TF. The ratio of
115 expression level of putative transcriptional target genes of each TF between (i) and (ii)
116 was calculated. The distribution of the ratios of all general TFs was compared among
117 forward-reverse (FR), reverse-forward (RF) orientations and non-orientation (i.e.
118 without considering orientation) of DNA motif sites of an insulator TF shortening an
119 EPA using Mann-Whitney test, two-sided (p-value $< 10^{-7}$), and putative insulator TFs
120 with significant differences of the distributions of the ratios of expression level among

121 FR, RF orientations and non-orientation were enumerated (Fig. 2A).

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

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

140 Without considering the difference of orientations of DNA binding motif
141 sequences among cell types, 175 biased orientations of DNA binding motif sequences
142 of TFs were found in the four cell types in common. These numbers (96 and 175) are

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

161 There may be a possibility that DNA binding sites of biased orientations of TFs
162 were located near those of CTCF, so the biased orientations of TFs seem to have
163 insulator function. To address the issue, I shortened an EPA using the DNA binding sites
164 of biased orientations of CTCF and those of another biased orientation of a TF in 96

165 TFs found in common among the four cell types. Transcriptional target genes of TFs
166 were predicted based on the EPA, and the ratio of expression level of the transcriptional
167 target genes of a TF in open chromatin regions in the EPA and closed chromatin regions
168 in promoter was examined. The ratio was compared between an EPA shortened at the
169 DNA binding sites of CTCF and the merger of those of CTCF and another biased
170 orientation of a TF. If the DNA binding sites of biased orientation of a TF include
171 insulator sites not associated with those of CTCF, the ratio would be higher than an EPA
172 shortened at the DNA binding sites of only CTCF. However, there are many DNA
173 binding sites of CTCF in the human genome, the difference between the ratios may not
174 be large. The biased orientations of CTCF consist of 19 DNA binding motif sequences
175 in monocyte. When comparing the DNA sites of each DNA binding motif sequence of
176 CTCF with those of the merger of the same CTCF and another biased orientation of a
177 TF, among 96 biased orientations of TFs, 79, 70, 95 and 62 TFs showed a higher ratio in
178 monocyte, $CD4^+$ T cell, HMEC and NPC respectively (p-value < 0.05) (Supplemental
179 material). YY1 was in a higher rank (the top in HMEC and NPC, second in T cell, and
180 fifth in monocyte). The rank of each TF was normalized by the lowest rank in each cell
181 type and was summed in the four cell types, where 45 TFs were found in common
182 (Table 2). The result is consistent that the DNA binding sites of YY1 are not frequently
183 located close to those of CTCF and are associated with chromatin interactions inside
184 chromatin interactions formed by CTCF (Weintraub et al. 2017) (Figure 2C, D, E and
185 Figure 7E in the reference). About 45 – 65% of DNA binding sites of ZNF143 are
186 closely located to those of CTCF in three cell types (Wen et al. 2018) (Figure 1a in the

187 reference). DNA binding sites of ZNF143 tend to be close to transcriptional start sites in
188 promoters. Other biased orientations of TFs potentially have DNA sites with insulator
189 function other than DNA sites of CTCF. When the DNA sites of DNA motif sequences
190 of all biased orientations of CTCF were merged, and an EPA was shortened at the
191 merged DNA sites, among 96 biased orientations of TFs, 1, 0, 34 and 14 TFs showed a
192 higher ratio in monocyte, $CD4^+$ T cell, HMEC and NPC respectively (p-value < 0.05),
193 and included YY1 in HMEC and NPC, but not included YY1 and ZNF143 in monocyte
194 and T cell (Supplemental material). The criterion of the comparison seemed to be strict.
195 DNA binding sites of CTCF involved in non-insulator function and false-positive
196 predictions of DNA binding sites of CTCF may affect the result of the analysis.
197
198 **Comparison with chromatin interaction data**
199 To ask whether the biased orientations of DNA motif sequences are associated
200 with chromatin interaction, enhancer-promoter interactions (EPIs) predicted based on an
201 enhancer-promoter association (EPA) were compared with chromatin interaction data.
202 EPIs predicted based on an EPA were compared with three replications (B2T1, B2T2
203 and B3T1) of HiChIP chromatin interaction data in $CD4^+$ T cells separately (Mumbach
204 et al. 2017). The resolutions of HiChIP chromatin interaction data and EPIs were
205 adjusted to 5 kilobases (kb). The number of unique HiChIP cis (i.e. in the same
206 chromatin) chromatin interactions was 19,926,360 at 5 kb resolution in B2T1
207 replication of T cells, 666,149 at 5 kb resolution with $\geq$1,000 counts for each
208 interaction and 78,209 at 5 kb resolution with $\geq$6,000 counts for each interaction

209 (Supplemental material). EPI were predicted based on three types of EPAs: (i) EPA
210 shortened at the FR or RF orientation of DNA motif sites of a TF with putative insulator
211 function such as CTCF, (ii) EPA shortened at the DNA motif sites of a TF without
212 considering their orientation, and (iii) EPA without being shortened by DNA motif sites.
213 For total 201 biased orientations [133 (49%) FR and 90 (43%) RF of 273 FR and 211
214 RF] of DNA motif sequences in T cells, which included CTCF, cohesin (RAD21 and
215 SMC3), ZNF143 and YY1 in three replications (B2T1, B2T2, and B3T1), a
216 significantly higher ratio of EPIs overlapped with HiChIP chromatin interactions with
217 ≥1,000 counts for each interaction in EPA (i) than the other types of EPA (ii) and (iii) in
218 T cells (Table 3). When comparing EPIs predicted based on only EPA (i) and (iii) with
219 the chromatin interactions, for total 390 biased orientations [261 (96%) FR and 199
220 (94%) RF of 273 FR and 211 RF] of DNA motif sequences in T cells, a significantly
221 higher ratio of EPIs overlapped with the chromatin interactions in EPA (i) than EPA (iii)
222 (Table 3; Supplemental material). The difference between EPIs predicted based on EPA
223 (i) and (ii) seemed to be difficult to distinguish using the chromatin interaction data and
224 a statistical test in some cases. However, as for the difference between EPIs predicted
225 based on EPA (i) and (iii), a larger number of biased orientations of DNA motif
226 sequences were correlated with chromatin interaction data. Most of biased orientations
227 of DNA motif sequences (95%) were correlated with chromatin interactions, when
228 comparing EPIs predicted based on EPA (i) and (iii) with HiChIP chromatin interactions.
229 Chromatin interaction data were obtained from different samples from DNase-seq open
230 chromatin regions, so individual differences may exist. Individual differences of

231 chromatin interactions were larger than those of open chromatin regions (Mumbach et
232 al. 2017).

233 Additionally, to confirm these tendencies of comparison of EPIs with HiChIP
234 chromatin interactions, the same analysis was conducted using HiChIP chromatin
235 interaction data in Th17, Treg and GM12878 cells. For total 128 biased orientations [76
236 (38%) FR and 64 (35%) RF of 200 FR and 183 RF] of DNA motif sequences in Th17
237 cells, which included CTCF, cohesin (RAD21 and SMC3), ZNF143 and YY1 in three
238 replications (B1T2, B2T1, and B3T1), a significantly higher ratio of EPIs overlapped
239 with HiChIP chromatin interactions in EPA (i) than the other types of EPA (ii) and (iii)
240 (Supplemental material). When comparing EPIs predicted based on only EPA (i) and
241 (iii) with the chromatin interactions, for total 282 biased orientations [182 (91%) FR
242 and 162 (89%) RF of 200 FR and 183 RF] of DNA motif sequences, a significantly
243 higher ratio of EPIs overlapped with the chromatin interactions in EPA (i) than EPA (iii)
244 (Supplemental material).

245 In Treg cells, chromatin interactions in one biological replication were overlapped
246 with EPIs, but those in the other two biological replications were less overlapped with
247 EPIs, so the one replication (B3T1) was used for this analysis. For total 281 biased
248 orientations [154 (53%) FR and 176 (54%) RF of 290 FR and 323 RF] of DNA motif
249 sequences, which included CTCF, cohesin (RAD21 and SMC3), ZNF143 and YY1, a
250 significantly higher ratio of EPIs overlapped with HiChIP chromatin interactions in EPA
251 (i) than the other types of EPA (ii) and (iii) (Supplemental material). When comparing
252 EPIs predicted based on only EPA (i) and (iii) with the chromatin interactions, for total

253 482 biased orientations [283 (98%) FR and 309 (96%) RF of 290 FR and 323 RF] of
254 DNA motif sequences, a significantly higher ratio of EPIs overlapped with the
255 chromatin interactions in EPA (i) than EPA (iii) (Supplemental material).

256 In GM12878 cells, for total 168 biased orientations [96 (38%) FR and 98 (35%)
257 RF of 229 FR and 237 RF] of DNA motif sequences, which included CTCF, cohesin
258 (RAD21 and SMC3), ZNF143 and YY1 in two replications (B1 and B2), a significantly
259 higher ratio of EPIs overlapped with HiChIP chromatin interactions in EPA (i) than the
260 other types of EPA (ii) and (iii) (Supplemental material). When comparing EPIs
261 predicted based on only EPA (i) and (iii) with the chromatin interactions, for total 366
262 biased orientations [218 (95%) FR and 227 (96%) RF of 229 FR and 237 RF] of DNA
263 motif sequences, a significantly higher ratio of EPIs overlapped with the chromatin
264 interactions in EPA (i) than EPA (iii) (Supplemental material).

265 In situ Hi-C data in HMEC, which are available in 4D nucleome data portal, were
266 compared with putative EPIs. The number of unique in situ Hi-C cis chromatin
267 interactions was 121,873,642 at 3 kb resolution. For total 222 biased orientations [155
268 (50%) FR and 108 (40%) RF of 310 FR and 269 RF] of DNA motif sequences, which
269 included CTCF, cohesin (RAD21 and SMC3), ZNF143 and YY1, a significantly higher
270 ratio of EPIs overlapped with in situ Hi-C chromatin interactions (≥1 count for each
271 interaction, i.e. all interactions) in EPA (i) than the other types of EPA (ii) and (iii)
272 (Supplemental material). When comparing EPIs predicted based on only EPA (i) and
273 (iii) with the chromatin interactions, for total 406 biased orientations [268 (86%) FR
274 and 230 (86%) RF of 310 FR and 269 RF] of DNA motif sequences, a significantly

275 higher ratio of EPIs overlapped with the chromatin interactions in EPA (i) than EPA (iii)
276 (Supplemental material). There were six replications of in situ Hi-C experimental data
277 of HMEC, but other replications contained much smaller number of chromatin
278 interactions. The HiChIP chromatin interaction data in T cells were expected to be
279 enriched with enhancer-promoter interactions, since they were produced with H3K27ac
280 antibody to capture chromatin interactions. In situ Hi-C data consist of all types of
281 chromatin interactions and thus, contain a smaller number of enhancer-promoter
282 interactions than HiChIP data. Chromatin interactions with a small number of counts
283 may be artifacts occurred by ligation errors during experiments (Lieberman-Aiden et al.
284 2009; Yardimci et al. 2019). Therefore, other replications of in situ Hi-C data with a
285 smaller number of chromatin interactions will not be available for this analysis.

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

297 interactions in EPA (i) than EPA (iii) (Supplemental material).

298 All three types of chromatin interaction data (HiChIP, in situ Hi-C and promoter
299 capture Hi-C) indicated that EPIs predicted based on EPA (i) that was shortened at DNA
300 motif sites of known TFs involved in chromatin interactions (CTCF, RAD21, SMC3,
301 ZNF143 and YY1), overlapped with more chromatin interactions than the other EPA (ii)
302 and (iii) in six cell types (T cells, Th17, Treg, GM12878, HMEC and NPC). In
303 monocyte, Hi-C data was available, but it was not analyzed, due to the low resolution of
304 chromatin interaction data (50 kb).

305 The biased orientations of DNA motif sites of known TFs involved in chromatin
306 interactions such as CTCF, RAD21, SMC3, YY1 and ZNF143 may be overlapped with
307 the biased orientations of DNA motif sites of other TFs, and thus, they showed a
308 significantly higher ratio of EPIs overlapped with chromatin interactions in EPA (i) than
309 the other types of EPA (ii) and (iii). However, among 273 FR and 211 RF orientations of
310 DNA motifs of TFs not including the known TFs in T cells, the DNA motif sites of only
311 15 (6%) FR and 9 (4%) RF orientations of DNA motifs of TFs overlapped with $\geq 60\%$
312 of DNA motif sites of biased orientations of DNA motifs of either of the known TFs at 5
313 kb resolution.

314 Moreover, to investigate the enhancer activity of EPIs, the distribution of
315 expression level of putative target genes of EPIs was compared between EPIs
316 overlapped with HiChIP chromatin interactions and EPIs not overlapped with them.
317 Though the target genes of EPIs were selected from top 4,000 transcripts (genes) in
318 terms of the expression level of all transcripts (genes) excluding transcripts not

319 expressed in T, Th17, Treg and GM12878 cells separately, target genes of EPIs
320 overlapped with chromatin interactions indicated a significantly higher expression level
321 than EPIs not overlapped with them, suggesting that EPIs overlapped with chromatin
322 interactions activated the expression of target genes in the four cell types. For almost all
323 (99% - 100%) FR and RF orientations of DNA motifs, a significantly higher expression
324 level of putative target genes of EPIs overlapped with chromatin interactions than EPIs
325 not overlapped in the four cell types. When a biased orientation of DNA motif showed a
326 significantly higher expression level of putative target genes of EPIs overlapped with
327 chromatin interactions than EPIs not overlapped, '1' was marked with in the tables of
328 comparison between EPIs and HiChIP chromatin interactions in Supplemental material.
329 DNA motifs showing a significantly lower expression level or not showing a significant
330 difference of expression level were not observed in this analysis. HiChIP data were
331 produced using H3K27ac antibody, so chromatin interactions acting as repressor would
332 not be identified. However, for in situ Hi-C and promoter capture Hi-C data in HMEC
333 and NPC, the significant difference of expression level of putative target genes of EPIs
334 was observed in a small percentage (less than 1%) of biased orientations of DNA
335 motifs.
336
337 **Correlation of expression level of gene pairs separated by DNA motif sites of a TF**
338 To ask whether the biased orientations of DNA motif sites affect gene expression,
339 I compared the expression level in monocytes of protein-coding genes the most closely
340 located to upstream and downstream of biased orientations of DNA motif sites in the

341 human genome. The expression levels of closely located genes including divergent gene
342 pairs tend to be correlated among tissues (Purmann et al. 2007; Xie et al. 2007). When a
343 biased orientation of DNA motif sites serve as insulator (e.g. CTCF), the correlation of
344 gene expression level would be reduced. Among 96 biased (64 FR and 52 RF)
345 orientations of DNA motifs of TFs found in common in monocytes, T cells, HMEC and
346 NPC, TFs with ≥50 genomic locations of the DNA motif sites were selected, after the
347 elimination of ubiquitously expressed genes among 53 tissues and DNA motif sites near
348 the genes with coefficient of variance (CV) <90. After this filtering, 43 (72%) FR and
349 40 (80%) RF of 60 FR and 50 RF showed significantly lower correlations of expression
350 levels of gene pairs separated by the DNA motif sites of a TF than the correlations of all
351 pairs of neighbor genes with intergene distance <1 megabase (Mb) (Mann-Whitney test,
352 p-value < 0.05) (Table 4). The FR and RF orientations of TFs included known TFs
353 involved in chromatin interactions such as CTCF, RAD21, SMC3, YY1 and ZNF143.
354 Top 20 of the FR and RF orientations of DNA motifs were selected in ascending order
355 of the median correlation coefficient of expression levels of gene pairs separated by the
356 DNA motif sites of a TF. The 20 TFs included known TFs involved in chromatin
357 interactions such as CTCF, RAD21 and SMC3 (Fig. 4) (Mann-Whitney test, p-value <
358 0.01). Instead of using all DNA motif sites of a TF, when DNA motif sites of a TF were
359 limited to sites examined for biased orientations of DNA motifs in an EPA, 33 (60%)
360 FR and 36 (77%) RF of 55 FR and 47 RF showed significantly lower correlations of
361 expression levels of gene pairs separated by the DNA motif sites (Supplemental
362 material). With intergene distance <20 kb, 21 (34%) FR and 8 (15%) RF of 62 FR and

363 52 RF showed significantly lower correlations of expression levels of gene pairs
364 separated by the DNA motif sites of a TF than the correlations of all pairs of neighbor
365 genes (Supplemental material).

366 A large number of DNA motif sites of a TF tend to show significantly lower
367 correlations of gene expression (Median numbers of DNA motif sites showing
368 significantly lower correlations were 656 and 479 for FR and RF respectively, and
369 median numbers of DNA motif sites not showing lower correlations significantly were
370 122 and 115 for FR and RF respectively. Mann-Whitney test p-value = 0 and < 10^{-10} for
371 FR and RF respectively). This suggests that if the criteria to examine the correlations of
372 expression levels of gene pairs separated by DNA motif sites are more stringent, and
373 thus the number of DNA motif sites is reduced, it would be more difficult to identify
374 using a statistical test the difference of correlations of expression levels of gene pairs
375 separated by the reduced DNA motif sites of TFs, particularly for TFs with a small
376 number of DNA motif sites.

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

385 regions.

386

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

388 To examine the association of biased orientations of DNA binding motif
389 sequences of TFs, co-location of the DNA binding motif sequences within 200bp to
390 each other, which is close to the length of core DNA around octamer of core histone
391 (146bp), that is, the length of the minimum open chromatin region, was analyzed in
392 monocytes, T cells, HMEC and NPC. The same pairs of DNA binding motif sequences
393 in upstream and downstream of genes in EPAs were enumerated, and the pairs of DNA
394 binding motif sequences found in ≥100 genomic regions were listed (Table 5;
395 Supplemental material). Majority of pairs of DNA motif sites overlapped with more
396 than 1 base, but the same tendency was found in the analysis of DNA motif sites of all
397 TFs (Supplemental material). CTCF was partially overlapped with cohesin RAD21 and
398 SMC3 (Table 5). Top 30 pairs of FR and RF orientations of DNA motifs overlapped or
399 co-localized were shown (Table 5). Total numbers of unique pairs of DNA motif sites of
400 TFs overlapped or co-localized were 404 (FR) and 662 (RF) in monocytes, consisting of
401 177 (39% of all FR) and 218 (43% of all RF) unique DNA motifs. Only ten overlapping
402 or co-location of DNA motifs of TFs were identified in the four cell types in common,
403 including pairs of CTCF, RAD21, SMC3 and YY1. This implies that most of
404 overlapping and co-location of DNA motifs of TFs serve in cell-type-specific manner
405 (Supplemental material). Pairs of biased orientations of DNA binding motifs of TFs
406 tend to overlap, and relatively a small number of motifs are co-localized in upstream

407 and downstream of genes. Overlapping DNA motifs act as DNA binding sites of several
408 TFs simultaneously or competitively (Ackerman et al. 1991; Yoon and Chikaraishi
409 1992; Ansari and Peterson-Kaufman 2011; Slattery et al. 2011; Chatterjee et al. 2012;
410 He et al. 2015; Kin et al. 2016; Merabet and Mann 2016). CTCF and cohesin function
411 coordinately with another TF and a protein such as Klf4 and Brd2 (Wei et al. 2013; Hsu
412 et al. 2017). Klf4 recruits cohesin to the Oct4 enhancer region. BRD2 supports
413 boundary (insulator) activity of CTCF. The analysis of 457 ChIP-seq data sets on 119
414 human TFs showed that secondary motifs were detected in addition to the canonical
415 motifs of the TFs, indicating tethered binding and co-binding between multiple TFs
416 (Wang et al. 2012a). The authors observed significant position and orientation
417 preferences between many co-binding TFs. Overlapping DNA motif sites of TFs found
418 in this study may not be artifact, but biologically functional. Though repeat sequences in
419 human genome were masked, more than a hundred of pairs of DNA motif sites
420 overlapped at hundreds or thousands upstream and downstream regions of genes.
421

422 **Discussion**

423 The approach in this study is an initial comprehensive method to find DNA motif
424 sequences of transcription factors (TFs) associated with chromatin interaction and
425 insulator function. DNA motif sequences were searched from open chromatin regions of
426 monocytes, T cells, HMEC and NPC. Total 96 biased [64 forward-reverse (FR) and 52
427 reverse-forward (RF)] orientations of DNA motif sequences of TFs were found in the
428 four cell types in common, whereas non-orientation (i.e. without considering

429 orientation) of DNA motif sequences of TFs were not found to affect the expression
430 level of putative transcriptional target genes, suggesting that an enhancer-promoter
431 association (EPA) shortened at the FR or RF orientation of DNA motif sites of a TF
432 includes more accurate prediction of enhancer-promoter interactions (EPIs), which were
433 supported by the comparisons with chromatin interaction and gene expression data.

434 To predict DNA motif sites of TFs, I tried several tools and methods, and then
435 used protein interaction quantitation (PIQ) tool (Sherwood et al. 2014). PIQ predicts
436 DNA motif sites of TFs by reducing noises of DNase-seq data, and the comparison of
437 the prediction with ChIP-seq data has revealed that PIQ shows a better performance
438 than other tools and digital genome footprinting (DNase-DGF). The prediction of DNA
439 motif sites of TFs using PIQ in this study increased the number of biased orientations of
440 DNA motif sequences of TFs, reducing the number of non-orientation of DNA motif
441 sequences of TFs. At first, the number of predicted DNA motif sites of TFs was
442 expected to increase, relative to using consensus DNA sequences of TFs, which seemed
443 to be stringent to find all DNA motif sites of a TF in the genome. However, the number
444 of DNA motif sites was almost the same or slightly decreased, compared with the
445 analysis of consensus DNA sequences, suggesting that PIQ increases true positive
446 predictions of DNA motif sites, reducing false positive predictions.

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). The
449 approach in this study would be able to identify TFs forming heterodimer. If the DNA
450 motif site of only the mate to a pair of TFs was found in an EPA, the EPA was shortened

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

457 TFs that are multi-functional transcriptional regulator such as YY1 may not
458 indicate a strong biased orientation of their DNA motif sites. Though some proportion
459 of DNA motif sites of a TF involved in chromatin interaction and insulator function lie
460 in biased orientation in the genome, the other DNA motif sites of the TF may not form a
461 dimer with another TF, and not be present in a specific orientation in the genome. Thus,
462 overall proportion of biased orientations of DNA motif sites of a TF may be reduced, so
463 that a statistical test cannot detect it.

464 The approach in this study was tested to search for repeat sequences serving as
465 insulator sites and affecting the expression level of putative transcriptional target genes
466 of TFs. A small number of repeat sequences were identified and included MIR, which
467 acts as insulators (Wang et al. 2015).

468 CTCF and cohesin-binding sites are frequently mutated in cancer (Katainen et al.
469 2015). Some biased orientations of DNA motif sequences would be involved in
470 chromatin interactions and might be associated with diseases including cancer.

471 The analyses taken here suggest that biased orientations of DNA motifs of
472 discovered insulator TFs affect enhancer-promoter interactions and the expression level

473 of putative transcriptional target genes of general TFs bound in enhancers and
474 promoters through insulator function. This gives insight into the nature of insulators and
475 contributes to the researches of functions and molecular mechanisms of chromatin
476 interactions, TFs, transcriptional regulation and transcriptional cascades.
477

478 # Methods

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

480 To identify transcription factor binding sites (TFBSs) from open and closed
481 chromatin regions, TRANSFAC (2019.2), JASPAR (2018), UniPROBE (2018),
482 high-throughput SELEX, transcription factor binding sequences of ENCODE ChIP-seq
483 data, and HOCOMOCO version 9 and 11 were used to predict insulator sites
484 (Wingender et al. 1996; Newburger and Bulyk 2009; Portales-Casamar et al. 2010; Xie
485 et al. 2010; Zhao and Stormo 2011; Jolma et al. 2013; Kheradpour and Kellis 2014;
486 Kulakovskiy et al. 2018). TRANSFAC (2019.2) was used to analyze enhancer-promoter
487 interactions, because these data were sufficient to identify biased orientations of DNA
488 motif sequences of insulator TFs with less computational time, reducing the number of
489 non-orientation of DNA motif sequences of TFs. To reduce false positive prediction of
490 DNA motif sites of TFs in open chromatin regions, DNA motif search was performed
491 by protein interaction quantitation (PIQ) tool with 12,249 position frequency matrices
492 converted from DNA motifs of vertebrate TFs in the above databases and DNase-seq
493 data from GEO and ENCODE database (GSM1024791 $CD14^+$ monocytes; GSM665812
494 $CD4^+$ T cell; GSM736634 Human mammary epithelial cell, HMEC; GSM878615

495 Neural progenitor cell, NPC; ENCFF263GMV, ENCFF886XEV Th17; GSM1024741
496 Treg; GSM736620 GM12878) (Sherwood et al. 2014). To find open chromatin regions
497 associated with enhancer (and promoter) activity, narrow peaks of ChIP-seq
498 experimental data of H3K27ac were predicted using macs2 callpeak (GEO:
499 GSM773004) (Zhang et al. 2008). To predict DNA motif sites of TFs from closed
500 chromatin regions, another method was employed. Position weight matrices of
501 vertebrate transcription factor binding sequences were converted into DNA consensus
502 sequences, which consist of alphabet of 15 characters (the four bases A, C, G, T, the six
503 two-fold degenerate IUPAC codes R=[AG], Y=[CT], K=[GT], M=[AC], S=[GC],
504 W=[AT], the four three-fold degenerate IUPAC codes B=[CGT], D=[AGT], H=[ACT],
505 V=[ACG], and the four-fold degenerate character N=[ATGC]) using convert matrix in
506 RSAT website (van Helden 2003). Transcription factor binding sequences were
507 searched from narrow peaks of DNase-seq data in repeat-masked hg19 genome
508 sequences using Match tool in TRANSFAC with similarity score cutoff of 1 after
509 transforming the DNA consensus sequences of TFs into position frequency matrixes for
510 Match search (Kel et al. 2003). Repeat-masked hg19 genome sequences were
511 downloaded from UCSC genome browser (http://genome.ucsc.edu/,
512 http://hgdownload.soe.ucsc.edu/goldenPath/hg19/bigZips/hg19.fa.masked.gz). TFs
513 corresponding to transcription factor binding sequences were searched computationally
514 and manually by comparing their names and gene symbols of HGNC (HUGO Gene
515 Nomenclature Committee) -approved gene nomenclature and 31,848 UCSC known
516 canonical transcripts

517 (http://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/knownCanonical.txt.gz),
518 because transcription factor binding sequences were not linked to transcript IDs such as
519 UCSC, RefSeq and Ensembl transcripts.

520 Target genes of a TF were assigned when its TFBS was found in DNase-seq
521 narrow peaks in promoter or extended regions of enhancer-promoter association of
522 genes (EPA). Promoter and extended regions were defined as follows: promoter regions
523 were those that were within distance of ±5 kb from transcriptional start sites (TSS).
524 Promoter and extended regions were defined as per the following association rule,
525 which is the same as that defined in Figure 3A of a previous study (McLean et al. 2010):
526 the single nearest gene association rule, which extends the regulatory domain to the
527 midpoint between the TSS of the gene and that of the nearest gene upstream and
528 downstream without the limitation of extension length. The limitation of extension
529 length was ＜1 Mb in this study. Extended regions for an EPA were shortened at the
530 DNA motif sites of a putative insulator TF that were the closest to a TSS, and
531 transcriptional target genes were predicted based on the shortened EPA and DNA motif
532 sites of a general TF in open chromatin regions. Promoter and extended regions of an
533 EPA were shortened at the forward–reverse (FR) orientation of DNA binding sites of a
534 putative insulator TF. When forward or reverse orientation of DNA binding sites were
535 continuously located in the genome several times, the most internal (i.e. closest to a
536 TSS) forward–reverse orientation of DNA binding sites were selected. The genomic
537 positions of genes were identified using 'knownGene.txt.gz' file in UCSC
538 bioinformatics sites (Karolchik et al. 2014). The file 'knownCanonical.txt.gz' was also

539 utilized to select representative transcripts among various alternate forms for assigning
540 promoter and extended regions of an EPA. From the list of transcription factor binding
541 sequences and transcriptional target genes, redundant transcription factor binding
542 sequences were removed by comparing the putative target genes of a transcription
543 factor binding sequence and its corresponding TF; if identical, one of the transcription
544 factor binding sequences was chosen. When the number of transcriptional target genes
545 predicted based on a transcription factor binding sequence was less than five, the
546 transcription factor binding sequence was omitted.

547 For gene expression data, RNA-seq reads mapped onto the human hg19 genome
548 sequence were obtained from UCSF-UBC human reference epigenome mapping project
549 RNA-seq reads with poly-A of naive $CD4^+$ T cells (GEO: GSM669617). For monocytes,
550 Blueprint RNA-seq FPKM data
551 ('C0010KB1.transcript_quantification.rsem_grape2_crg.GRCh38.20150622.results')
552 were downloaded from Blueprint DCC portal
553 (http://dcc.blueprint-epigenome.eu/#/files). RNA-seq reads with poly-A of human
554 mammary epithelial cells (HMEC) in Encyclopedia of DNA Elements at UCSC
555 (http://genome.ucsc.edu/encode/,
556 'wgEncodeCshlLongRnaSeqHmecCellPapAlnRep1.bam' file), H1 derived neural
557 progenitor cell (NPC) (GEO: GSM915326, ENCODE: ENCFF529SIO), Th17
558 (GSM2859479, NCBI SRA: SRR6298326), Treg (GSM2859476, SRR6298323) and
559 GM12878 (GSE78551, ENCFF297QCE) were analyzed. RNA-seq reads were aligned
560 in repeat-masked hg19 genome sequences using BWA with default parameters (Li and

561 Durbin 2009). FPKMs of the RNA-seq data were calculated by RSeQC (Wang et al.
562 2012b). Based on log2-transformed FPKM, UCSC transcripts were arranged in
563 descending order of expression level, and top 4,000 transcripts were selected in each
564 cell type.

565 The expression level of transcriptional target genes predicted based on EPAs
566 shortened at the DNA motif sites of a TF was compared with the expression level of
567 transcriptional target genes predicted based on closed chromatin regions in promoters.
568 For each DNA motif sequence shortening EPAs, transcriptional target genes were
569 predicted based on DNA binding motif sequences of vertebrate TFs in TRANSFAC
570 database, and the ratio of expression level of putative transcriptional target genes of
571 each TF was calculated between an EPA and closed chromatin regions in promoters.
572 The distribution of the ratios of all TFs was compared among forward-reverse (FR),
573 reverse-forward (RF) and non-orientation (i.e. without considering orientation) of a
574 DNA motif sequence shortening an EPA using Mann-Whitney test, two-sided (p-value <
575 10^{-7}). Other parameters for the analysis were set as follows: the number of
576 transcriptional target genes of a TF was $\geq$50, the number of general TFs (not insulator
577 TFs) to predict transcriptional target genes was $\geq$50, and the number of DNA motif
578 sites of a TF (insulator TF) shortening an EPA was $\geq$100.

579

580 **Comparison with chromatin interaction data**

581 For comparison of EPIs with chromatin interactions (HiChIP) in $CD4^+$ T, Th17,
582 Treg and GM12878 cells,

583 ‘GSM2705049_Naive_HiChIP_H3K27ac_B2T1_allValidPairs.txt’,

584 ‘GSM2705050_Naive_HiChIP_H3K27ac_B2T2_allValidPairs.txt’,

585 ‘GSM2705051_Naive_HiChIP_H3K27ac_B3T1_allValidPairs.txt’,

586 ‘GSM2705054_Th17_HiChIP_H3K27ac_B1T2_allValidPairs.txt’,

587 ‘GSM2705055_Th17_HiChIP_H3K27ac_B2T1_allValidPairs.txt’,

588 ‘GSM2705056_Th17_HiChIP_H3K27ac_B3T1_allValidPairs.txt’,

589 ‘GSM2705059_Treg_HiChIP_H3K27ac_B3T1_allValidPairs.txt’,

590 ‘GSM2705041_GM_HiChIP_H3K27ac_B1_allValidPairs.txt’,

591 ‘GSM2705042_GM_HiChIP_H3K27ac_B2_allValidPairs.txt’ files were downloaded

592 from Gene Expression Omnibus (GEO) database. The resolutions of chromatin

593 interactions and EPIs were adjusted to 5 kb before their comparison. Chromatin

594 interactions with ≥6,000 and ≥1,000 counts for each interaction were analyzed in the

595 comparison in $CD4^+$ T cells. According to the unique numbers of chromatin interactions

596 in replications in Th17, Treg and GM12878 cells, chromatin interactions with ≥1,000

597 (B1T2 in Th17), ≥2,000 (B2T1 and B3T1 in Th17), ≥6,000 (B3T1 in Treg), ≥4,000

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

599 Enhancer-promoter interactions (EPIs) were predicted based on three types of

600 EPAs in monocytes: (i) EPA shortened at the FR or RF orientation of DNA motif sites

601 of a TF, (ii) EPA shortened at the non-orientation (i.e. without considering orientation)

602 of DNA motif sites of a TF, and (iii) EPA without being shortened by a DNA motif site.

603 EPIs predicted based on the three types of EPAs in common were removed. EPIs

604 predicted based on EPA (i) were removed from EPIs predicted based on EPA (ii). EPIs

605 predicted based on EPA (i) and (ii) were removed from EPIs predicted based on EPA
606 (iii). The resolution of HiChIP chromatin interaction data was 1-5 kb, so EPIs were
607 adjusted to 5 kb before their comparison. The number and ratio of EPIs overlapped with
608 chromatin interactions were compared two times between EPI (i) and (iii), and EPI (i)
609 and (ii) (binomial distribution, p-value < 0.025 for each test, two-sided, 95% confidence
610 interval).

611 For comparison of EPIs with chromatin interactions (in situ Hi-C) in HMEC, the
612 genomic positions of chromatin interactions in 'HMEC_4DNFI97O9SAZ.pairs.gz' file
613 in 4D nucleome data portal (http://data.4dnucleome.org) were converted into hg19
614 version of human genome using liftOver tool
615 (http://hgdownload.soe.ucsc.edu/downloads.html#liftover). The resolutions of
616 chromatin interaction data and EPIs were adjusted to 3 kb before their comparison. All
617 chromatin interactions were used in the comparison. For comparison of EPIs with
618 chromatin interactions (promoter capture Hi-C) in NPC, the resolutions of chromatin
619 interaction data in 'GSE86189_npc.po.all.txt.bz2' file (GEO: GSE86189) and EPIs were
620 adjusted to 5 kb before their comparison. All chromatin interactions were used in the
621 comparison.

622 Putative target genes for the comparison of EPIs and chromatin interactions were
623 selected from top 4,000 transcripts in term of the expression level. The expression level
624 of putative target genes of EPIs overlapped with chromatin interactions was compared
625 with EPIs not overlapped with them. When the same transcriptional target gene of a TF
626 in an enhancer was predicted based on both an EPI overlapped with a chromatin

627 interaction and an EPI not overlapped with, the target gene was eliminated. The
628 distribution of expression level of putative target genes was compared using
629 Mann-Whitney test, two-sided (p-value < 0.05).
630
631 **Correlation of expression level of gene pairs separated by DNA motif sites of a TF**
632 To examine the correlation of expression level of genes pairs separated by DNA
633 motif sites of a TF, gene expression data among 53 tissues were obtained from
634 'GTEx_Analysis_2016-01-15_v7_RNASeQCv1.1.8_gene_median_tpm.gct.gz' file in
635 GTEx database (https://gtexportal.org/home/). The correlation of log2-transformed gene
636 expression level was analyzed based on protein-coding transcripts of RefSeq (RefSeq
637 ID with 'NM_'). The same genomic location of RefSeq transcripts were removed. DNA
638 motif sites overlapped with RefSeq transcripts were eliminated. Genes closest to a DNA
639 motif sites and another gene were analyzed using BEDOPS closest-features (Neph et al.
640 2012). Mann-Whitney test was performed using R.
641
642 ## Acknowledgements
643 The supercomputing resource was provided by Human Genome Center of the Institute
644 of Medical Science at the University of Tokyo. Publication charges for this article were
645 funded by JSPS KAKENHI Grant Number 16K00387. This work was partially
646 supported by JST CREST Grant Number JPMJCR15G1, Japan.
647

648 **Figures**

649 Web links to original figures.

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

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

652 76_2014_BFnrg3682_MOESM18_ESM.ppt

653 Graphical Abstract (de Wit et al. 2015)

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

655 92c5eba1d/fx1.jpg

656

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

658 2010)

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

660 87_2010_BFnbt1630_MOESM11_ESM.pdf

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

662 DNA Looping (Guo et al. 2015)

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

664 2af081251/gr4.jpg

665

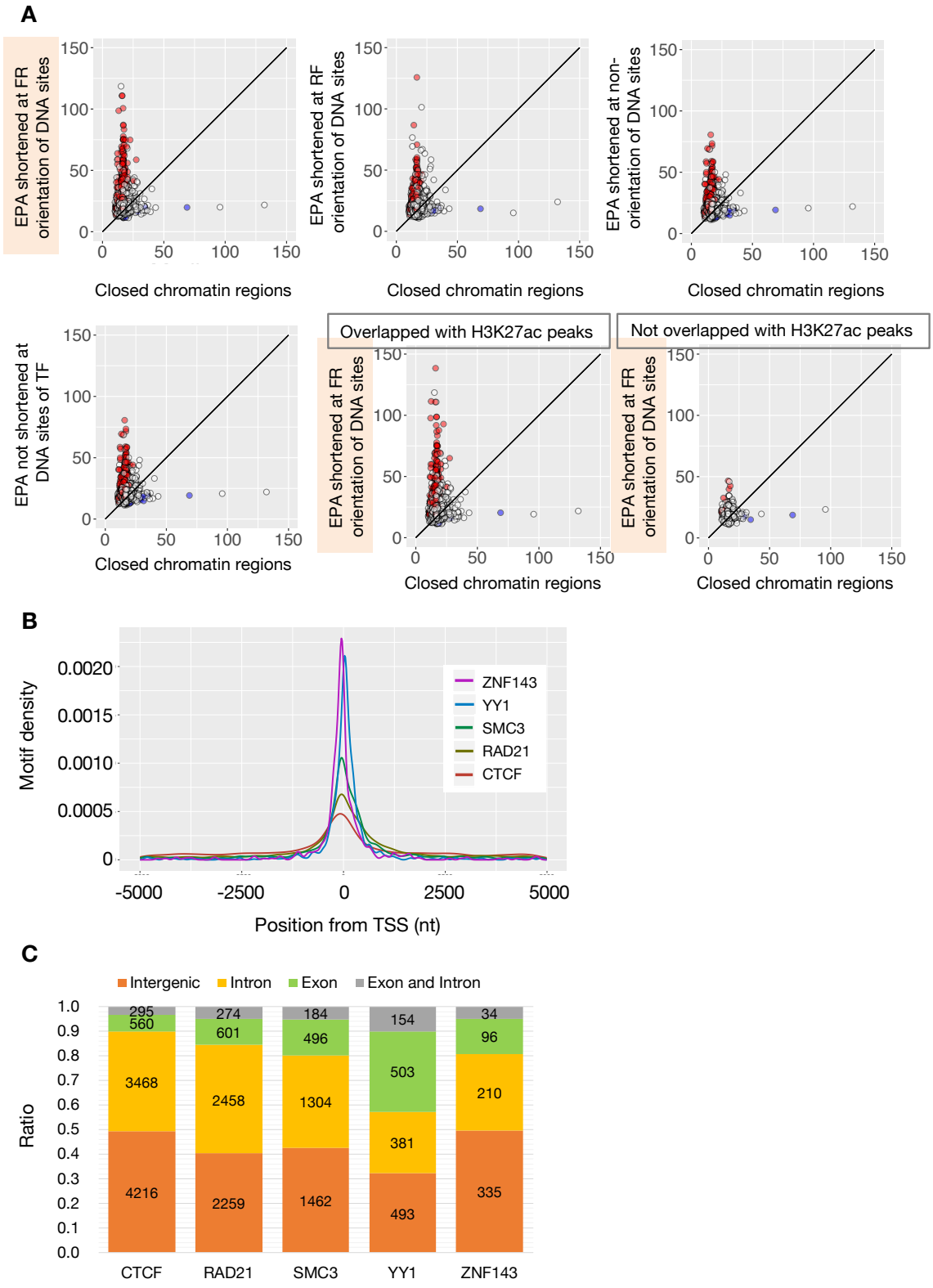
666 **Figure 1.** Chromatin interaction and enhancer-promoter association. (A) Forward–

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

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

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

670 2014; de Wit et al. 2015; Guo et al. 2015). (B) Computationally-defined regulatory
671 domains for enhancer-promoter association (McLean et al. 2010). The *single nearest*
672 *gene* association rule extends the regulatory domain to the midpoint between this gene's
673 TSS and the nearest gene's TSS both upstream and downstream. The enhancer-promoter
674 association was shortened at the forward-reverse orientation of DNA binding motif sites
675 of a TF in this study (e.g. CTCF in the figure).
676



677

678 **Figure 2.** Identification of DNA motif sites of TFs in open chromatin regions and

679 properties of DNA motif sites of TFs. (A) Comparison of expression level of putative

680 transcriptional target genes. Each dot represents the median expression level (FPKM) of
681 target genes predicted based on DNA motif sites of a transcription factor in closed
682 chromatin regions in promoters (X-axis) and an enhancer-promoter association (EPA)
683 shortened at the forward-reverse orientation of DNA motif sites of CTCF (Y-axis) in T
684 cell (upper left graph). The distribution of expression level was changed according to
685 four criteria of EPAs (four graphs) (Osato 2018). Red dots indicate the median
686 expression level of target genes was significantly higher in an EPA than closed
687 chromatin regions, and blue dots indicate the median expression level of target genes
688 was significantly lower in an EPA than closed chromatin regions (Mann-Whitney test).
689 The expression level of target genes predicted based on an EPA tend to be higher than
690 closed chromatin regions, implying that TFs serve as activators of target genes. DNA
691 motif sites of TFs overlapped with H3K27ac ChIP-seq peaks were associated with a
692 significant difference of expression level of target genes, and DNA motif sites not
693 overlapped with H3K27ac peaks were not associated with the difference (lower right
694 two graphs). This suggests that the significant difference of expression level is
695 associated with enhancer (and promoter) signals of H3K27ac. (B) Distribution of DNA
696 motif sites of TFs near TSS. ZNF143 and YY1 tend to be located near TSS, and CTCF
697 tend to be observed more in distant regions than around TSS in monocyte. (C)
698 Distribution of DNA motif sites of TFs in the human genome in monocytes. Most of
699 DNA motif sites were found in intergenic and intron regions. YY1 is known to be at
700 TSS in the edge of exons of genes, and is involved in transcriptional activation,
701 repression and initiation.

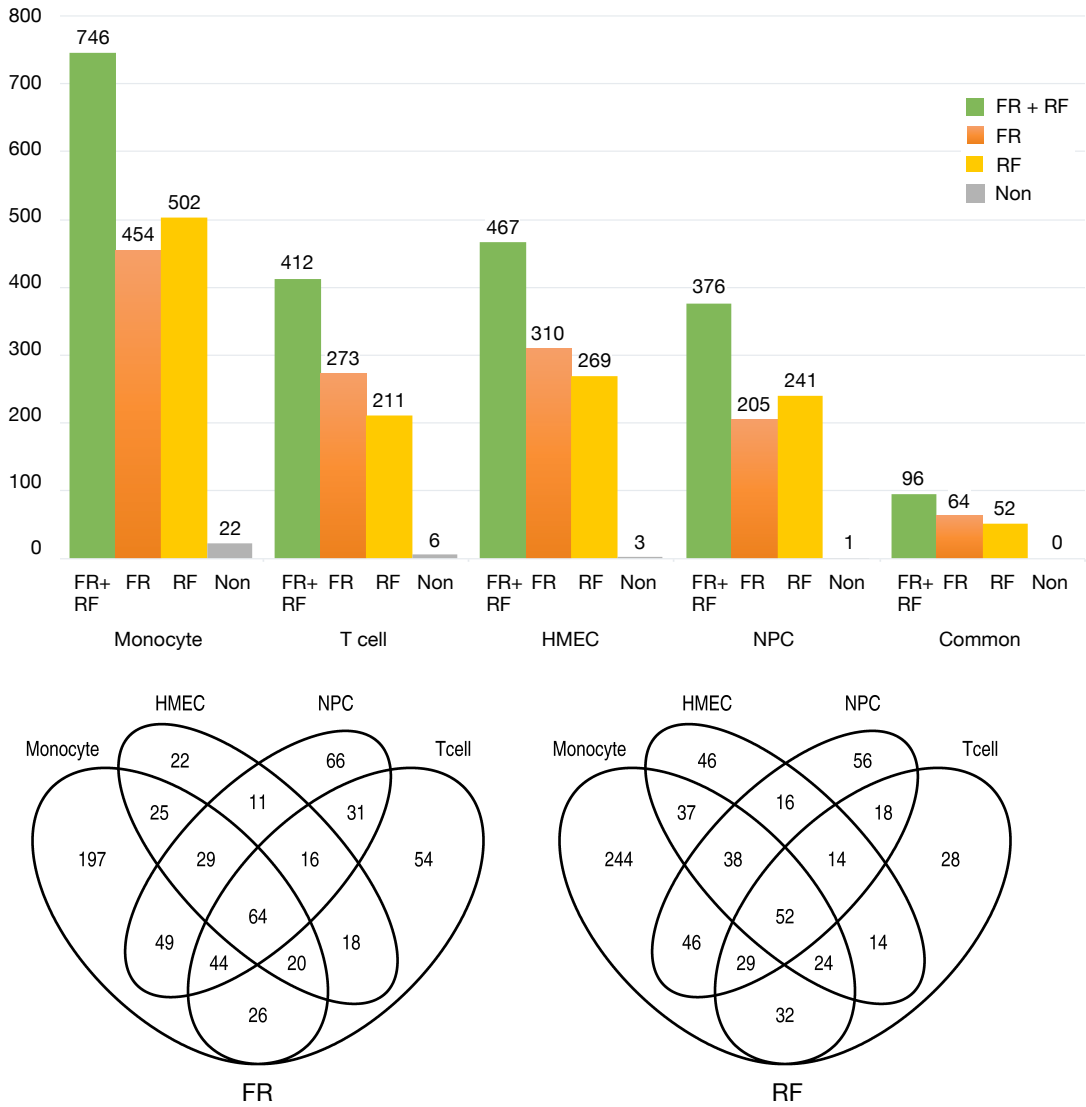
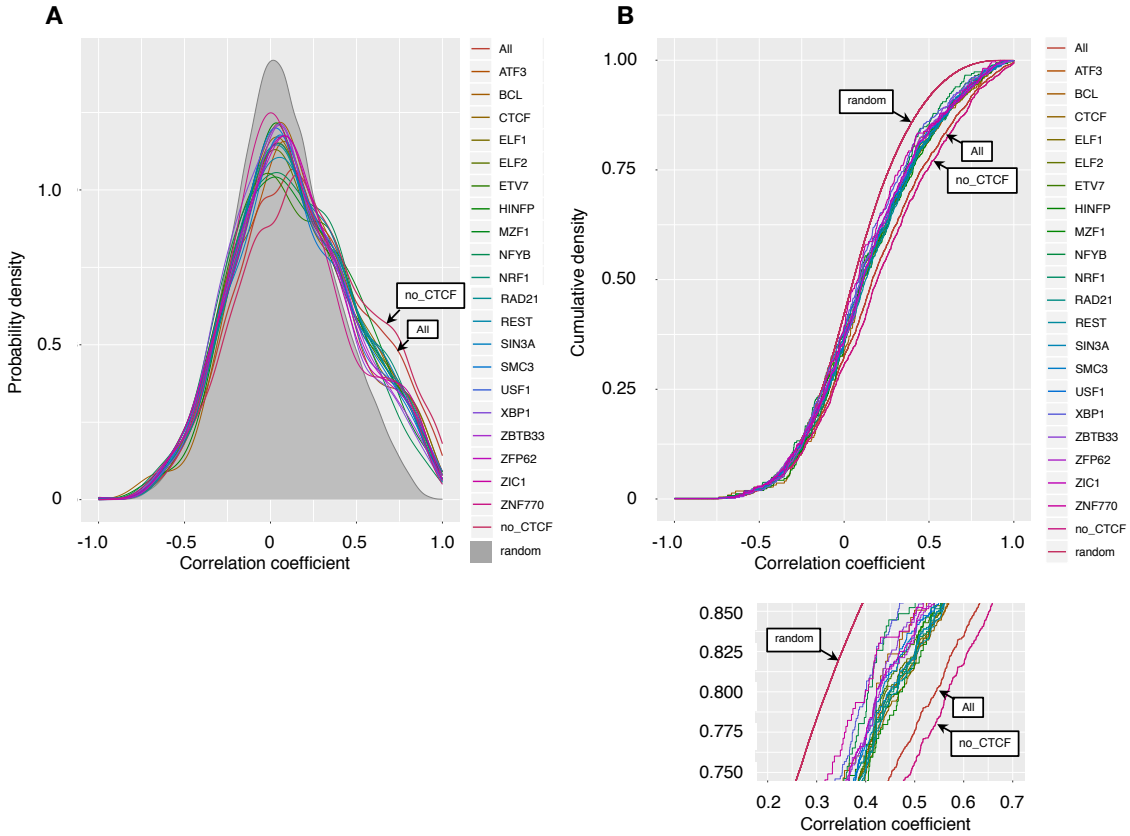


Figure 3. Biased orientations of DNA motif sequences of TFs affect the transcription of genes by the putative insulator function of the DNA motif sites. Total 96 (64 FR and 52 RF) of biased orientations of DNA binding motif sequences of TFs affected significantly the expression level of putative transcriptional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

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

710 significantly.

711



712

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

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

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

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

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

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

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

720 by predicted CTCF sites.

721 Web links to original figures.

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

723 (Bailey et al. 2015).

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

725

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

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

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

729 2015).

730

731

Tables

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

Forward-reverse orientation

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

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

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

Reverse-forward orientation

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

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

740 **Table 2.** Comparison of biased orientation of a TF and CTCF. As insulator and
741 chromatin-interaction sites, the merger of DNA motif sites of a TF and CTCF showed a
742 higher expression level of putative transcriptional target genes than DNA motif sites of
743 only CTCF. Each rank was divided by the lowest rank in each cell type for
744 normalization. Known TFs associated with chromatin interactions are indicated in bold.

Rank total	TF	monocyte	T cell	HMEC	NPC
0.30	**YY1**	0.28	0.11	0.11	0.08
0.41	TATA	0.78	0.22	0.11	0.08
0.52	YY2	0.56	0.33	0.11	0.08
0.54	NR3C1	0.06	0.06	0.33	0.15
0.88	MYC	0.06	0.39	0.33	0.15
1.21	USF1	0.94	0.83	0.22	0.15
1.23	MZF1	0.33	0.39	0.22	0.62
1.25	SREBF2	0.39	0.33	0.22	0.69
1.27	MAX	0.06	0.28	0.22	0.77
1.28	ETS1	0.50	0.33	0.33	0.62
1.31	EGR2	0.94	0.17	0.22	0.92
1.42	SP2	0.83	0.89	0.22	0.31
1.43	XBP1	0.94	0.56	0.33	0.54
1.46	CREB3L1	0.28	0.89	0.11	0.46
1.56	ETS	0.28	0.83	0.11	0.62
1.62	FOS	1.00	0.89	0.11	0.62
1.66	ATF3	0.06	0.67	0.22	0.77
1.66	AHR	1.00	0.94	0.33	0.38
1.68	TP73	0.44	1.00	0.44	0.23
1.71	RUNX3	0.28	0.89	0.67	0.15
1.76	NFY	0.72	0.78	0.44	0.54
1.77	YBX1	0.56	0.78	0.22	0.77
1.77	ARNT	0.39	0.94	0.44	0.38
1.83	PLAG1	0.17	0.72	0.11	1.00
1.84	**CTCF**	0.33	0.78	0.44	0.62
1.90	PBX3	0.50	0.89	0.78	0.23
1.97	MYB	0.17	0.83	0.44	0.69
1.97	KLF4	0.72	0.94	0.33	0.69
2.00	NONO	0.89	0.78	0.22	1.00
2.00	ZBTB7A	0.56	0.67	0.33	1.00
2.05	**ZNF143**	0.39	0.94	0.33	0.77
2.06	FLI1	0.94	0.83	0.22	1.00
2.12	NFYB	0.33	0.72	0.78	0.62
2.15	E2F	1.00	0.89	0.33	0.92
2.16	**RAD21**	0.50	0.61	0.78	0.77
2.17	MITF	0.67	0.72	0.44	1.00
2.17	ERG	0.78	0.83	0.33	1.00
2.17	ELK3	0.72	0.83	0.33	1.00
2.20	TFAP2C	0.28	0.94	0.33	0.92
2.21	**SMC3**	0.56	0.89	0.56	0.77
2.22	ZBTB33	0.06	1.00	0.22	1.00
2.28	HIC1	0.89	0.94	0.33	1.00
2.35	ZNF770	0.72	0.83	0.67	0.85
2.39	BHLHE41	0.94	0.94	0.44	1.00
2.78	PLAGL1	0.56	1.00	0.78	1.00

745

746

747

748 **Table 3.** Comparison of enhancer-promoter interactions (EPIs) with chromatin
749 interactions in T cells. The upper tables show the number of biased orientations of DNA
750 motifs, where a significantly higher ratio of EPIs that were predicted based on an
751 enhancer-promoter association (EPA) (i) overlapped with HiChIP chromatin
752 interactions than the other types of EPA (ii) and (iii) in T cells. The middle tables show
753 133 FR and 90 RF orientations of DNA motifs found in common among B2T1, B2T2
754 and B3T1 replications in T cells in the upper table. The lower tables show that among

755 64 FR and 52 RF orientations of DNA motifs found in common in monocytes, T cell,
756 HMEC and NPC, 44 (69%) FR and 28 (54%) RF were matched with the analysis of
757 HiChIP chromatin interactions according to three types of EPAs. TF: DNA binding
758 motif sequence of a transcription factor. Score: $-\log_{10}$ (*p*-value). Score 1: Comparison of
759 an EPA shortened at the FR or RF orientation of DNA motif sites [EPA (i)] with an EPA
760 not shortened [EPA (iii)]. Score 2: Comparison of an EPA shortened at the FR or RF
761 orientation of DNA motif sites [EPA (i)] with an EPA shortened at the non-orientation of
762 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

763
764
765
766
767
768
769

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

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

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

Forward-reverse (FR)

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

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

Reverse-forward (RF)

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

Table 4. Correlation of expression level of gene pairs separated by DNA motif sites of a TF. Pairs of genes separated by 64 FR and 52 RF orientations of DNA motif sites of TFs found in common in monocytes, T cells, HMEC and NPC were analyzed to find significantly lower correlation of their expression level than the correlation of all pairs of nearby genes among 53 tissues. Score: $-\log_{10}$ (p-value) of Mann-Whitney test. Median correlation: median correlation coefficient of expression level of nearby gene pairs separated by DNA motif sites. The median correlation coefficient of expression level of all pairs of nearby genes is 0.17. # DNA site: number of DNA motif sites of a

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

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

794 coefficient of variance <90.

795

796 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

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

804

805 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

806

807

808

809

Reverse-forward (RF)

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

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

FR orientation of CTCF

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

RF orientation of CTCF

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

819

820 **References**

821

822 Ackerman SL, Minden AG, Williams GT, Bobonis C, Yeung CY. 1991. Functional
823 significance of an overlapping consensus binding motif for Sp1 and Zif268 in
824 the murine adenosine deaminase gene promoter. *Proc Natl Acad Sci U S A* **88**:
825 7523-7527.
826 Ansari AZ, Peterson-Kaufman KJ. 2011. A partner evokes latent differences between
827 Hox proteins. *Cell* **147**: 1220-1221.
828 Bailey SD, Zhang X, Desai K, Aid M, Corradin O, Cowper-Sal Lari R, Akhtar-Zaidi B,
829 Scacheri PC, Haibe-Kains B, Lupien M. 2015. ZNF143 provides sequence
830 specificity to secure chromatin interactions at gene promoters. *Nat Commun* **2**:
831 6186.
832 Barutcu AR, Lajoie BR, Fritz AJ, McCord RP, Nickerson JA, van Wijnen AJ, Lian JB,
833 Stein JL, Dekker J, Stein GS et al. 2016. SMARCA4 regulates gene expression
834 and higher-order chromatin structure in proliferating mammary epithelial cells.
835 *Genome research* **26**: 1188-1201.
836 Chatterjee R, Zhao J, He X, Shlyakhtenko A, Mann I, Waterfall JJ, Meltzer P,
837 Sathyanarayana BK, FitzGerald PC, Vinson C. 2012. Overlapping ETS and CRE
838 Motifs ((G/C)CGGAAGTGACGTCA) preferentially bound by GABPalpha and
839 CREB proteins. *G3 (Bethesda)* **2**: 1243-1256.
840 Chen M, Manley JL. 2009. Mechanisms of alternative splicing regulation: insights from
841 molecular and genomics approaches. *Nat Rev Mol Cell Biol* **10**: 741-754.
842 Das D, Clark TA, Schweitzer A, Yamamoto M, Marr H, Arribere J, Minovitsky S,
843 Poliakov A, Dubchak I, Blume JE et al. 2007. A correlation with exon
844 expression approach to identify cis-regulatory elements for tissue-specific
845 alternative splicing. *Nucleic acids research* **35**: 4845-4857.
846 de Wit E, Vos ES, Holwerda SJ, Valdes-Quezada C, Verstegen MJ, Teunissen H,
847 Splinter E, Wijchers PJ, Krijger PH, de Laat W. 2015. CTCF Binding Polarity
848 Determines Chromatin Looping. *Molecular cell* **60**: 676-684.
849 Guo Y, Xu Q, Canzio D, Shou J, Li J, Gorkin DU, Jung I, Wu H, Zhai Y, Tang Y et al.
850 2015. CRISPR Inversion of CTCF Sites Alters Genome Topology and

851 Enhancer/Promoter Function. *Cell* **162**: 900-910.
852 He X, Syed KS, Tillo D, Mann I, Weirauch MT, Vinson C. 2015. GABPalpha Binding
853 to Overlapping ETS and CRE DNA Motifs Is Enhanced by CREB1: Custom
854 DNA Microarrays. *G3 (Bethesda)* **5**: 1909-1918.
855 Hsu SC, Gilgenast TG, Bartman CR, Edwards CR, Stonestrom AJ, Huang P, Emerson
856 DJ, Evans P, Werner MT, Keller CA et al. 2017. The BET Protein BRD2
857 Cooperates with CTCF to Enforce Transcriptional and Architectural Boundaries.
858 *Molecular cell* **66**: 102-116 e107.
859 Jolma A, Yan J, Whitington T, Toivonen J, Nitta KR, Rastas P, Morgunova E, Enge M,
860 Taipale M, Wei G et al. 2013. DNA-binding specificities of human transcription
861 factors. *Cell* **152**: 327-339.
862 Jolma A, Yin Y, Nitta KR, Dave K, Popov A, Taipale M, Enge M, Kivioja T, Morgunova
863 E, Taipale J. 2015. DNA-dependent formation of transcription factor pairs alters
864 their binding specificity. *Nature* **527**: 384-388.
865 Jung I, Schmitt A, Diao Y, Lee AJ, Liu T, Yang D, Tan C, Eom J, Chan M, Chee S et al.
866 2019. A compendium of promoter-centered long-range chromatin interactions in
867 the human genome. *Nature genetics* **51**: 1442-1449.
868 Karolchik D, Barber GP, Casper J, Clawson H, Cline MS, Diekhans M, Dreszer TR,
869 Fujita PA, Guruvadoo L, Haeussler M et al. 2014. The UCSC Genome Browser
870 database: 2014 update. *Nucleic acids research* **42**: D764-770.
871 Katainen R, Dave K, Pitkanen E, Palin K, Kivioja T, Valimaki N, Gylfe AE, Ristolainen
872 H, Hanninen UA, Cajuso T et al. 2015. CTCF/cohesin-binding sites are
873 frequently mutated in cancer. *Nature genetics* **47**: 818-821.
874 Kel AE, Gossling E, Reuter I, Cheremushkin E, Kel-Margoulis OV, Wingender E. 2003.
875 MATCH: A tool for searching transcription factor binding sites in DNA
876 sequences. *Nucleic acids research* **31**: 3576-3579.
877 Kheradpour P, Kellis M. 2014. Systematic discovery and characterization of regulatory
878 motifs in ENCODE TF binding experiments. *Nucleic acids research* **42**:
879 2976-2987.
880 Kin K, Chen X, Gonzalez-Garay M, Fakhouri WD. 2016. The effect of non-coding
881 DNA variations on P53 and cMYC competitive inhibition at cis-overlapping
882 motifs. *Hum Mol Genet* **25**: 1517-1527.
883 Kulakovskiy IV, Vorontsov IE, Yevshin IS, Sharipov RN, Fedorova AD, Rumynskiy EI,

884 Medvedeva YA, Magana-Mora A, Bajic VB, Papatsenko DA et al. 2018.
885 HOCOMOCO: towards a complete collection of transcription factor binding
886 models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic acids*
887 *research* **46**: D252-D259.
888 Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler
889 transform. *Bioinformatics (Oxford, England)* **25**: 1754-1760.
890 Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A,
891 Amit I, Lajoie BR, Sabo PJ, Dorschner MO et al. 2009. Comprehensive
892 mapping of long-range interactions reveals folding principles of the human
893 genome. *Science (New York, NY)* **326**: 289-293.
894 McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM,
895 Bejerano G. 2010. GREAT improves functional interpretation of cis-regulatory
896 regions. *Nature biotechnology* **28**: 495-501.
897 Merabet S, Mann RS. 2016. To Be Specific or Not: The Critical Relationship Between
898 Hox And TALE Proteins. *Trends Genet* **32**: 334-347.
899 Mumbach MR, Satpathy AT, Boyle EA, Dai C, Gowen BG, Cho SW, Nguyen ML,
900 Rubin AJ, Granja JM, Kazane KR et al. 2017. Enhancer connectome in primary
901 human cells identifies target genes of disease-associated DNA elements. *Nature*
902 *genetics* **49**: 1602-1612.
903 Neph S, Kuehn MS, Reynolds AP, Haugen E, Thurman RE, Johnson AK, Rynes E,
904 Maurano MT, Vierstra J, Thomas S et al. 2012. BEDOPS: high-performance
905 genomic feature operations. *Bioinformatics (Oxford, England)* **28**: 1919-1920.
906 Newburger DE, Bulyk ML. 2009. UniPROBE: an online database of protein binding
907 microarray data on protein-DNA interactions. *Nucleic acids research* **37**:
908 D77-82.
909 Osato N. 2018. Characteristics of functional enrichment and gene expression level of
910 human putative transcriptional target genes. *BMC Genomics* **19**: 957.
911 Portales-Casamar E, Thongjuea S, Kwon AT, Arenillas D, Zhao X, Valen E, Yusuf D,
912 Lenhard B, Wasserman WW, Sandelin A. 2010. JASPAR 2010: the greatly
913 expanded open-access database of transcription factor binding profiles. *Nucleic*
914 *acids research* **38**: D105-110.
915 Purmann A, Toedling J, Schueler M, Carninci P, Lehrach H, Hayashizaki Y, Huber W,
916 Sperling S. 2007. Genomic organization of transcriptomes in mammals:

917 Coregulation and cofunctionality. *Genomics* **89**: 580-587.
918 Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn
919 AL, Machol I, Omer AD, Lander ES et al. 2014. A 3D map of the human
920 genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**:
921 1665-1680.
922 Schreiber J, Libbrecht M, Bilmes J, Noble W. 2017. Nucleotide sequence and DNaseI
923 sensitivity are predictive of 3D chromatin architecture. *bioRxiv*.
924 Sherwood RI, Hashimoto T, O'Donnell CW, Lewis S, Barkal AA, van Hoff JP, Karun V,
925 Jaakkola T, Gifford DK. 2014. Discovery of directional and nondirectional
926 pioneer transcription factors by modeling DNase profile magnitude and shape.
927 *Nature biotechnology* **32**: 171-178.
928 Shlyueva D, Stampfel G, Stark A. 2014. Transcriptional enhancers: from properties to
929 genome-wide predictions. *Nature reviews Genetics* **15**: 272-286.
930 Shrivastava A, Calame K. 1994. An analysis of genes regulated by the multi-functional
931 transcriptional regulator Yin Yang-1. *Nucleic acids research* **22**: 5151-5155.
932 Slattery M, Riley T, Liu P, Abe N, Gomez-Alcala P, Dror I, Zhou T, Rohs R, Honig B,
933 Bussemaker HJ et al. 2011. Cofactor binding evokes latent differences in DNA
934 binding specificity between Hox proteins. *Cell* **147**: 1270-1282.
935 Tabuchi TM, Deplancke B, Osato N, Zhu LJ, Barrasa MI, Harrison MM, Horvitz HR,
936 Walhout AJ, Hagstrom KA. 2011. Chromosome-biased binding and gene
937 regulation by the Caenorhabditis elegans DRM complex. *PLoS Genet* **7**:
938 e1002074.
939 van Helden J. 2003. Regulatory sequence analysis tools. *Nucleic acids research* **31**:
940 3593-3596.
941 Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF,
942 Schroth GP, Burge CB. 2008. Alternative isoform regulation in human tissue
943 transcriptomes. *Nature* **456**: 470-476.
944 Wang J, Vicente-Garcia C, Seruggia D, Molto E, Fernandez-Minan A, Neto A, Lee E,
945 Gomez-Skarmeta JL, Montoliu L, Lunyak VV et al. 2015. MIR retrotransposon
946 sequences provide insulators to the human genome. *Proc Natl Acad Sci U S A*
947 **112**: E4428-4437.
948 Wang J, Zhuang J, Iyer S, Lin X, Whitfield TW, Greven MC, Pierce BG, Dong X,
949 Kundaje A, Cheng Y et al. 2012a. Sequence features and chromatin structure

950 around the genomic regions bound by 119 human transcription factors. *Genome*
951 *research* **22**: 1798-1812.
952 Wang L, Wang S, Li W. 2012b. RSeQC: quality control of RNA-seq experiments.
953 *Bioinformatics (Oxford, England)* **28**: 2184-2185.
954 Wei Z, Gao F, Kim S, Yang H, Lyu J, An W, Wang K, Lu W. 2013. Klf4 organizes
955 long-range chromosomal interactions with the oct4 locus in reprogramming and
956 pluripotency. *Cell Stem Cell* **13**: 36-47.
957 Weintraub AS, Li CH, Zamudio AV, Sigova AA, Hannett NM, Day DS, Abraham BJ,
958 Cohen MA, Nabet B, Buckley DL et al. 2017. YY1 Is a Structural Regulator of
959 Enhancer-Promoter Loops. *Cell* **171**: 1573-1588 e1528.
960 Wen Z, Huang ZT, Zhang R, Peng C. 2018. ZNF143 is a regulator of chromatin loop.
961 *Cell Biol Toxicol* **34**: 471-478.
962 Wingender E, Dietze P, Karas H, Knuppel R. 1996. TRANSFAC: a database on
963 transcription factors and their DNA binding sites. *Nucleic acids research* **24**:
964 238-241.
965 Xi H, Yu Y, Fu Y, Foley J, Halees A, Weng Z. 2007. Analysis of overrepresented motifs
966 in human core promoters reveals dual regulatory roles of YY1. *Genome research*
967 **17**: 798-806.
968 Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, Lander ES, Kellis M.
969 2005. Systematic discovery of regulatory motifs in human promoters and 3'
970 UTRs by comparison of several mammals. *Nature* **434**: 338-345.
971 Xie X, Mikkelsen TS, Gnirke A, Lindblad-Toh K, Kellis M, Lander ES. 2007.
972 Systematic discovery of regulatory motifs in conserved regions of the human
973 genome, including thousands of CTCF insulator sites. *Proc Natl Acad Sci U S A*
974 **104**: 7145-7150.
975 Xie Z, Hu S, Blackshaw S, Zhu H, Qian J. 2010. hPDI: a database of experimental
976 human protein-DNA interactions. *Bioinformatics (Oxford, England)* **26**:
977 287-289.
978 Yardimci GG, Ozadam H, Sauria MEG, Ursu O, Yan KK, Yang T, Chakraborty A, Kaul
979 A, Lajoie BR, Song F et al. 2019. Measuring the reproducibility and quality of
980 Hi-C data. *Genome Biol* **20**: 57.
981 Yoon SO, Chikaraishi DM. 1992. Tissue-specific transcription of the rat tyrosine
982 hydroxylase gene requires synergy between an AP-1 motif and an overlapping E

box-containing dyad. *Neuron* **9**: 55-67.

Zhang H, Li F, Jia Y, Xu B, Zhang Y, Li X, Zhang Z. 2017. Characteristic arrangement of nucleosomes is predictive of chromatin interactions at kilobase resolution. *Nucleic acids research* **45**: 12739-12751.

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W et al. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**: R137.

Zhao Y, Stormo GD. 2011. Quantitative analysis demonstrates most transcription factors require only simple models of specificity. *Nature biotechnology* **29**: 480-483.