Discovery of directional chromatin-associated regulatory motifs affecting human gene transcription

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 are essential in enhancer-promoter interactions (EPIs) and transcription regulation. CTCF and cohesin proteins located at chromatin interaction anchors and other DNA-binding proteins such as YY1, ZNF143, and SMARCA4 are involved in chromatin interactions. However, overall understanding about proteins associated with chromatin interactions and insulator functions remains unclear. Here, I describe a systematic and comprehensive approach for discovering DNA-binding motifs of transcription factors (TFs) that affect EPIs and gene expression. This analysis identified 96 biased orientations (64 forward-reverse (FR) and 52 reverse-forward (RF)) of motifs that significantly affected the expression level of putative transcriptional target genes in monocytes, T cells, HMEC, and NPC and included CTCF, cohesin (RAD21 and SMC3), YY1, and ZNF143 (some TFs had more than one motif in databases; hence, the total number is smaller than the sum of FR and RF). KLF4, ERG, RFX, RFX2, HIF1, SP1, STAT3, and AP1 were associated with chromatin interactions. Many other TFs were also known to have chromatin-associated functions. The predicted biased orientations of motifs were compared with chromatin interaction data. Correlations in expression level of nearby genes separated by the motif sites were then examined among 53 tissues. Most TFs showed weak directional biases at chromatin interaction anchors and were difficult to identify based on enrichment analysis of motifs. These findings contribute to the investigation of chromatin-associated motifs involved in transcriptional regulation, chromatin interactions/regulation, and histone modifications.

Keywords: gene expression, transcriptional regulation, transcription factors, enhancer-promoter interactions, chromatin interactions and regulation, CTCF, open chromatin regions, histone modifications

Background

Chromatin interactions play essential roles in enhancer-promoter interactions (EPIs) and transcriptional regulation. CTCF and cohesin proteins are located at the chromatin interaction anchors and facilitate the formation of the loop structures. CTCF functions as an insulator by limiting the activity of enhancers in the loops (Fig. 1a) [1, 2]. The DNA-binding sequence of CTCF indicates its orientation bias at chromatin interaction anchors as the forward-reverse (FR) orientation is frequently observed [1, 3, 4]. Other DNA-binding proteins such as ZNF143, YY1, and SMARCA4 (BRG1) are involved in chromatin interactions and EPIs [5-7]. ZNF143 is known to bind more to the promoters of genes than the distal regions, thereby establishing looping with the distal element bound by CTCF [5]. The DNA-binding motif of ZNF143 is significantly enriched at both chromatin interaction anchors (Z-score > 7), whereas the DNA-binding motif of YY1 is not enriched (Z-score < 2; [5] Figure 2a). CTCF, cohesin, ZNF143, YY1, and SMARCA4 are involved in various biological functions other than chromatin interaction. These transcription factors (TFs) are bound in open chromatin regions near transcriptional start sites (TSS) as well, which may reduce the enrichment of the motifs at chromatin interaction anchors.

55 Computational approaches to predict chromatin interactions have been developed
56 [8, 9]. However, they were not intended to find DNA motifs of TFs involved in
57 chromatin interactions. Xie et al. [10] have concluded that the next challenge will be to
58 develop systematic methods to discern the specific functions of these motifs in a
59 genome-wide fashion.

60 DNA-binding proteins involved in chromatin interactions are considered to affect
61 the transcription of genes within the loops formed by chromatin interactions. To
62 confirm this hypothesis, I previously found that the expression level of human putative
63 transcriptional target genes of a TF was significantly changed according to the
64 enhancer-promoter association (EPA) [11]. An EPA was shortened at the FR orientation
65 of CTCF-binding sites, and the transcriptional target genes of each TF bound to
66 enhancers and promoters were predicted based on the EPA (Fig. 1b) [12]. The
67 expression levels predicted based on EPA were significantly different compared to the
68 expression levels predicted based on only promoters (or closed chromatin regions). The
69 expression levels tended to increase in monocytes and $CD4^+$ T cells, implying that TFs
70 serve as activators of target genes. The expression levels tended to decrease in
71 embryonic stem (ES) and induced pluripotent stem (iPS) cells, implying that TFs serve
72 as repressors of target genes. The expression level of putative transcriptional target
73 genes of a TF bound in enhancers is changed significantly when EPIs are predicted
74 properly using the insulator sites of CTCF in different cell types. This conclusion
75 suggests that the difference in the expression level reflects the accuracy of prediction of
76 transcriptional target genes based on a given EPA. Using this approach, I planned to

77 search for other DNA-binding proteins involved in chromatin interactions, which affect
78 the expression level of putative transcriptional target genes through their insulator
79 function.

80 Experimental analyses of chromatin interactions in molecular biology have
81 technical limitations. Chromatin interaction data are altered according to experimental
82 techniques, the number of DNA sequencing reads, and even replication sets of the same
83 cell type. Additionally, chromatin interaction data may not be saturated enough to
84 contain all chromatin interactions and may include experimental noise such as ligation
85 errors. To avoid the technical limitations of experimental analyses, a computational
86 search was conducted for DNA motifs of TFs that affect EPI and the expression level of
87 putative transcriptional target genes in monocytes, T cells, human mammary epithelial
88 cells (HMEC), and neural progenitor cells (NPC) without using chromatin interaction
89 data. Then, I compared putative EPIs and transcription factor binding sites (TFBSs)
90 with chromatin interaction and gene expression data.

91

92 ## Results

93 ### Search for biased orientations of DNA motifs

94 TFBSs were predicted using open chromatin regions and DNA motifs of TFs
95 collected from various databases and journals (see Methods). Transcriptional target
96 genes were predicted based on TFBSs in promoters and enhancers. An EPA was
97 shortened at the TFBSs with insulator function, such as CTCF and cohesin (RAD21 and
98 SMC3). To find motifs of TFs with insulator function other than CTCF and cohesin, an

99 EPA was shortened at the TFBSs of a putative insulator TF. Transcriptional target genes
100 of each general TF bound in (i) distal open chromatin regions (enhancers) and (ii)
101 closed chromatin regions in promoters were predicted based on the EPA in order to
102 estimate the enhancer activity of each general TF. The ratio of expression level between
103 (i) and (ii) was calculated. The distribution of the ratios of all general TFs was
104 compared among forward-reverse (FR), reverse-forward (RF) orientations, and
105 non-orientation (i.e., without considering orientation) groups of TFBSs using a
106 two-sided Mann-Whitney test (p-value $< 10^{-7}$). Putative insulator TFs with significant
107 differences in the distributions of the ratios of expression level among FR, RF, and
108 non-orientation were enumerated (Fig. 2a).

109 To confirm the predictions of TFBSs, the distributions of TFBSs were examined
110 in the human genome. TFBSs involved in chromatin interactions were identified near
111 TSS (Fig. 2b). ZNF143 is known to bind to the promoter of genes, thereby establishing
112 looping with the distal element bound by CTCF [5]. YY1 is a multi-functional
113 transcriptional regulator and is involved in transcriptional activation, repression, and
114 initiation [13, 14]. The distributions of TFBSs seemed to be consistent with the features
115 of the TFs. Most TFBSs were identified in intergenic and intron regions of the human
116 genome (Fig. 2c).

117 Several hundreds of FR and RF orientations were found in monocytes, T cells,
118 HMEC, and NPC (Fig. 3). The number of the TFs seemed to be changed, according to
119 the number of DNase-seq reads, since the number of the reads in monocytes was larger
120 than those of the other three cell types. When DNase-seq reads increase, more TFBSs

121 may be predicted to some extent. Ninety-six biased (sixty-four FR and fifty-two RF)
122 orientations of DNA-binding motifs of TFs were discovered and found to be common
123 between the four cell types, whereas non-orientation of DNA-binding motifs of TFs
124 were not detected (Fig. 3; Table 1; Additional file 1: Table S1). The FR orientation of
125 DNA motifs included CTCF, cohesin (RAD21 and SMC3), YY1, and ZNF143, which
126 are involved in chromatin interactions and EPIs.

127 Without considering the difference of orientations among cell types, 175 biased
128 orientations were found in common between the four cell types. These numbers (96 and
129 175) are the numbers of unique TFs with biased orientations of DNA motifs and not the
130 numbers of unique DNA motifs. Possible causes of the discrepancy are as follows: (1)
131 A TF or an isoform of the same TF may bind to a different DNA-binding motif,
132 according to cell types and/or in the same cell type. About 50% or more alternative
133 splicing isoforms are differently expressed among tissues, thereby indicating that most
134 alternative splicing is subject to tissue-specific regulation [15-17]; (2) The same TF has
135 several DNA-binding motifs, and in some cases, one of the motifs is almost the same as
136 the reverse complement motif of another motif of the same TF in a database; (3) I
137 previously found that a complex of TFs would bind to a slightly different DNA-binding
138 motif by combining DNA-binding motifs from the complex in *C. elegans* [18]. In other
139 reports, the difference of binding specificity of TF pairs was investigated [19].
140 Cofactors also alter the binding specificity of TFs [20-22]. These increase the variety
141 and number of (biased orientations of) DNA-binding motifs, which are potentially
142 involved in different biological functions among cell types.

143 Among the 96 TFs, KLF4 [23], ERG [24, 25], RFX [26], RFX2 [27], HIF1 [28],
144 SP1 [29], STAT3 [30], and AP1 [31] were found to be associated with chromatin
145 interactions. Myc [32, 33], the ETS family, ELK3 [34], ELK4 [35], RUNX3 [36], USF1
146 [37, 38], AHR/ARNT [39], SP2 [40], EGR2 [41-43], ZBTB33 [44], NF-Y [45, 46],
147 MITF [47, 48], NONO [49], FLI1 [50], E2F [51-53], BCL [54], CXXC1 [55], E2F1
148 [56], EGR1 [57], REST [58, 59], RFX [60], SIN3A [61-63], AP1 [64], and STAT3 and
149 STAT5 [65] were associated with chromatin remodeling, chromatin accessibility, and/or
150 histone modification. NRF1 [6, 66] and TFAP2 [67] were pioneer factors, which were
151 associated with chromatin remodeling and accessibility. TFBSs of the TFs other than
152 those involved in chromatin interactions would affect gene expression by
153 chromatin-associated functions. Huang et al. [37] have mentioned that the insulator
154 element at the 5' end of the chicken beta-globin locus acts as a barrier, thus protecting
155 transgenes against silencing effects of adjacent heterochromatin. The transcription
156 factor USF1 binds within the insulator at a site important in adjacent nucleosomes for
157 generating histone modifications associated with active chromatin and, by inference,
158 with barrier function.
159
160 **Comparison with chromatin interaction data**
161 To examine whether the biased orientations of DNA motifs were associated with
162 chromatin interactions, EPIs predicted based on an EPA were compared with chromatin
163 interaction data. EPIs predicted based on an EPA were compared with three replications
164 (B2T1, B2T2, and B3T1) of HiChIP chromatin interaction data in $CD4^+$ T cells

165 separately [68]. The resolutions of HiChIP chromatin interaction data and EPIs were
166 adjusted to 5 kilobases (kb). The number of unique HiChIP cis (i.e., in the same
167 chromatin) chromatin interactions was 19,926,360 at 5-kb resolution in B2T1
168 replication of T cells; 666,149 at 5-kb resolution with ≥1,000 counts for each
169 interaction; and 78,209 at 5-kb resolution with ≥6,000 counts for each interaction
170 (Additional file 1: Table S2). EPIs were predicted based on three types of EPAs: (i)
171 EPA shortened at the FR or RF orientation of TFBSs of a TF with putative insulator
172 function such as CTCF, (ii) EPA shortened at the TFBSs of a TF without considering
173 their orientation, and (iii) EPA which were not shortened by TFBSs. For a total of 201
174 biased orientations [133 of 273 (49%) FR and 90 of 211 (43%) RF] of DNA motifs in T
175 cells, which included CTCF, cohesin (RAD21 and SMC3), ZNF143, and YY1 in three
176 replications (B2T1, B2T2, and B3T1), a significantly higher ratio of EPIs overlapped
177 with HiChIP chromatin interactions with ≥1,000 counts for each interaction in EPA (i)
178 than in EPA (ii) and (iii) in T cells (Table 2). For a total of 390 biased orientations [261
179 of 273 (96%) FR and 199 of 211 (94%) RF] of DNA motifs in T cells, a significantly
180 higher ratio of EPIs overlapped with the chromatin interactions in EPA (i) than EPA
181 (iii) (Table 2; Additional file 1: Table S3). The difference between EPIs predicted based
182 on EPA (i) and (ii) seemed to be difficult to distinguish using the chromatin interaction
183 data and a statistical test in some cases. However, for the difference between EPIs
184 predicted based on EPA (i) and (iii), a larger number of biased orientations of DNA
185 motifs were correlated with chromatin interaction data. Most of the biased orientations
186 of DNA motifs (95%) were correlated with chromatin interactions when comparing

187 EPIs predicted based on EPA (i) and (iii) with HiChIP chromatin interactions.
188 Chromatin interaction data were obtained from different samples of DNase-seq open
189 chromatin regions; thus, individual differences may exist. Individual differences of
190 chromatin interactions were larger than those of open chromatin regions [68].

191 Additionally, to confirm these tendencies of comparison of EPIs with HiChIP
192 chromatin interactions, the same analysis was conducted using HiChIP chromatin
193 interaction data in Th17, Treg, and GM12878 cells. For a total of 128 biased
194 orientations [76 of 200 (38%) FR and 64 of 183 (35%) RF] of DNA motifs in Th17
195 cells, which included CTCF, cohesin (RAD21 and SMC3), ZNF143, and YY1 in three
196 replications (B1T2, B2T1, and B3T1), a significantly higher ratio of EPIs overlapped
197 with HiChIP chromatin interactions in EPA (i) than in EPA (ii) and (iii) (Additional file
198 1: Table S4). When comparing EPIs predicted based only on EPA (i) and (iii) with the
199 chromatin interactions, for a total of 282 biased orientations [182 of 200 (91%) FR and
200 162 of 183 (89%) RF] of DNA motifs, a significantly higher ratio of EPIs overlapped
201 with the chromatin interactions in EPA (i) than EPA (iii) (Additional file 1: Table S4).

202 In Treg cells, chromatin interactions in one biological replication were overlapped
203 with EPIs, but because those in the other two biological replications were less
204 overlapped with EPIs, one replication (B3T1) was used for this analysis. For a total of
205 281 biased orientations [154 of 290 (53%) FR and 176 of 323 (54%) RF] of DNA motif
206 sequences, which included CTCF, cohesin (RAD21 and SMC3), ZNF143, and YY1, a
207 significantly higher ratio of EPIs overlapped with HiChIP chromatin interactions in
208 EPA (i) than the other types of EPA (ii) and (iii) (Additional file 1: Table S4). When

209 comparing EPIs predicted based on only EPA (i) and (iii) with the chromatin
210 interactions, for total 482 biased orientations [283 of 290 (98%) FR and 309 of 323
211 (96%) RF] of DNA motif sequences, a significantly higher ratio of EPIs overlapped
212 with the chromatin interactions in EPA (i) than EPA (iii) (Additional file 1: Table S4).

213 In GM12878 cells, for a total of 168 biased orientations [96 of 229 (38%) FR and
214 98 of 237 (35%) RF] of DNA motifs including CTCF, cohesin (RAD21 and SMC3),
215 ZNF143, and YY1 in two replications (B1 and B2), a significantly higher ratio of EPIs
216 overlapped with HiChIP chromatin interactions in EPA (i) than EPA (ii) and (iii)
217 (Additional file 1: Table S4). When comparing EPIs predicted based on only EPA (i)
218 and (iii) with the chromatin interactions, for a total of 366 biased orientations [218 of
219 229 (95%) FR and 227 of 237 (96%) RF] of DNA motifs, a significantly higher ratio of
220 EPIs overlapped with the chromatin interactions in EPA (i) than EPA (iii) (Additional
221 file 1: Table S4).

222 In situ Hi-C data in HMEC, which are available in a 4D nucleome data portal,
223 were compared with putative EPIs. The number of unique in situ Hi-C cis chromatin
224 interactions was 121,873,642 at 3-kb resolution. For a total of 222 biased orientations
225 [155 of 310 (50%) FR and 108 of 269 (40%) RF] of DNA motifs including CTCF,
226 cohesin (RAD21 and SMC3), ZNF143, and YY1, a significantly higher ratio of EPIs
227 overlapped with in situ Hi-C chromatin interactions (≥1 count for each interaction, i.e.,
228 all interactions) in EPA (i) than in EPA (ii) and (iii) (Additional file 1: Table S5). When
229 comparing EPIs predicted based on only EPA (i) and (iii) with the chromatin
230 interactions, for a total of 406 biased orientations [268 of 310 (86%) FR and 230 of 269

231 (86%) RF] of DNA motifs, a significantly higher ratio of EPIs overlapped with the
232 chromatin interactions in EPA (i) than EPA (iii) (Additional file 1: Table S5). There
233 were six replications of in situ Hi-C experimental data of HMEC, but other replications
234 contained much smaller number of chromatin interactions. The HiChIP chromatin
235 interaction data in T cells were expected to be enriched with EPIs, since they were
236 produced with H3K27ac antibody to capture chromatin interactions. In situ Hi-C data
237 consist of all types of chromatin interactions and thus contain a smaller number of EPIs
238 than HiChIP data. Chromatin interactions with a small number of counts may be
239 artifacts caused by ligation errors during experiments [69, 70]. Therefore, other
240 replications of in situ Hi-C data with a smaller number of chromatin interactions will
241 not be available for this analysis.

242 Promoter capture Hi-C data of NPC were downloaded and analyzed by the same
243 procedure as that for T cells and HMEC [71]. The number of unique promoter capture
244 Hi-C cis chromatin interactions was 9,436,689 at 5-kb resolution. For a total of 109
245 biased orientations [61 of 205 (28%) FR and 60 of 241 (25%) RF] of DNA motifs,
246 which included CTCF, cohesin (RAD21 and SMC3), ZNF143, and YY1, a significantly
247 higher ratio of EPIs overlapped with promoter capture Hi-C chromatin interactions (≥1
248 count for each interaction, i.e. all interactions) in EPA (i) than in EPA (ii) and (iii)
249 (Additional file 1: Table S6). When comparing EPIs predicted based on only EPA (i)
250 and (iii) with the chromatin interactions, for a total of 186 biased orientations [99 of 205
251 (48%) FR and 103 of 241 (43%) RF] of DNA motifs, a significantly higher ratio of
252 EPIs overlapped with the chromatin interactions in EPA (i) than EPA (iii) (Additional

253 file 1: Table S6).

254 All three types of chromatin interaction data (HiChIP, in situ Hi-C, and promoter
255 capture Hi-C) indicated that EPIs predicted based on EPA (i) that was shortened at
256 TFBSs of known TFs involved in chromatin interactions (CTCF, RAD21, SMC3,
257 ZNF143 and YY1) overlapped with more chromatin interactions than those based on
258 EPA (ii) and (iii) in six cell types (T cells, Th17, Treg, GM12878, HMEC, and NPC). In
259 monocytes, Hi-C data was available, but it was not analyzed due to the low resolution of
260 chromatin interaction data (50 kb).

261 The biased orientations of TFBSs of known TFs involved in chromatin
262 interactions such as CTCF, RAD21, SMC3, YY1, and ZNF143 might overlap with the
263 biased orientations of TFBSs of other TFs; thus, they showed a significantly higher ratio
264 of EPIs overlapped with chromatin interactions in EPA (i) than in EPA (ii) and (iii).
265 However, among 273 FR and 211 RF orientations of DNA motifs of TFs not including
266 the known TFs in T cells, the TFBSs of only 15 (6%) FR and 9 (4%) RF orientations of
267 DNA motifs of TFs overlapped with $\geq$ 60% of the TFBSs of biased orientations of
268 DNA motifs of either of the known TFs at 5-kb resolution.

269 Moreover, to investigate the enhancer activity of EPIs, the distribution of
270 expression level of putative target genes of EPIs was compared between EPIs
271 overlapped with HiChIP chromatin interactions and EPIs not overlapped with them.
272 Though the target genes of EPIs were selected from top 4,000 transcripts (genes) in
273 terms of the expression level of all transcripts (genes) excluding transcripts not
274 expressed in T, Th17, Treg, and GM12878 cells separately, target genes of EPIs

275 overlapped with chromatin interactions indicated a significantly higher expression level
276 than non-overlapped EPIs, suggesting that EPIs overlapped with chromatin interactions
277 activated the expression of target genes in the four cell types. For almost all
278 (99%-100%) FR and RF orientations of DNA motifs, a significantly higher expression
279 level of putative target genes was observed for EPIs overlapped with chromatin
280 interactions than for EPIs not overlapped in the four cell types. When a biased
281 orientation of DNA motifs showed a significantly higher expression level of putative
282 target genes of EPIs overlapped with chromatin interactions than EPIs not overlapped,
283 '1' was marked with in the tables of comparison between EPIs and HiChIP chromatin
284 interactions (Additional file 1: Table S3). DNA motifs showing a significantly lower
285 expression level or not showing a significant difference of expression level were not
286 observed in this analysis. HiChIP data were produced using H3K27ac antibody; thus,
287 chromatin interactions acting as repressors would not be identified. However, for in situ
288 Hi-C and promoter capture Hi-C data in HMEC and NPC, the significant difference of
289 expression level of putative target genes of EPIs was observed in a small percentage
290 (less than 1%) of biased orientations of DNA motifs.

291 Recently, high resolution chromatin interaction data became publicly available
292 [72]. The biased orientations of TFBSs of TFs were examined using Micro-C chromatin
293 interaction data in human fibroblasts. The resolution of the data was adjusted to 200 bp.
294 TFBSs of CTCF, RAD21, and SMC3 showed a significant directional bias at the pairs
295 of chromatin interaction anchors. SMRC2-binding sites showed a directional bias at the
296 anchors (p-value $< 10^{-50}$). SMARCA4 is known to be involved in higher-order

297 chromatin structures such as the topologically associating domain (TAD) [7]. DNA
298 motifs of SMARCA4 did not exist in the databases of DNA motifs of TFs used in this
299 analysis. SMRC2 was not identified in this analysis because TFBSs of TFs affecting
300 EPI and gene expression were searched, which are not high-order chromatin structures.
301 ATF3 and ZBTB7A in Table 1 were found among the top 32 biased orientations of
302 TFBSs of TFs at the pairs of chromatin interaction anchors with p-value $< 10^{-50}$
303 (Additional file 1: Table S8). Deletion of an ATF3-binding site affected the expression
304 of a target gene, which was far apart from the site [73]. ZBTB7A is involved in the
305 changes in chromatin accessibility and the regulation of gene expression [74]. TFBSs of
306 YY1 and ZNF143 also indicated a weak directional bias at the chromatin interaction
307 anchors (Additional file 1: Table S7). In this analysis, biased orientations of TFBSs of
308 TFs were investigated based on the difference of distribution of the expression level of
309 putative transcriptional target genes. YY1 and ZNF143 were found to have significant
310 directional biases, thereby implying that the analysis in this study could find TFs that
311 did not have a strong directional bias based on the number of TFBSs at chromatin
312 interaction anchors. The other TFs, except for CTCF, RAD21, SMC3, ATF3, and
313 ZBTB7A in Table 1 showed weak directional biases at the chromatin interaction
314 anchors (Additional file 1: Table S7).
315
316 **Correlation of expression level of gene pairs separated by TFBS**
317 To investigate whether the biased orientations of TFBSs affected gene expression,
318 the expression level of protein-coding genes in monocytes, that are most closely located

319 upstream and downstream of any biased orientations of TFBSs in the human genome,
320 was compared. The expression levels of closely located genes including divergent gene
321 pairs tend to be correlated among tissues [75, 76]. When a biased orientation of TFBSs
322 serves as an insulator (e.g. CTCF), the correlation of gene expression level would be
323 reduced. Among 96 biased (64 FR and 52 RF) orientations of TFBSs found in common
324 in monocytes, T cells, HMEC, and NPC, TFs with ≥50 genomic locations of the
325 TFBSs were selected after the elimination of ubiquitously expressed genes among 53
326 tissues and TFBSs near the genes with a coefficient of variance (CV) < 90. After this
327 filtering, 43 of 60 (72%) FR and 40 of 50 (80%) RF orientations showed significantly
328 lower correlations of expression levels of gene pairs separated by the TFBSs of a TF
329 than the correlations of all pairs of neighboring genes with intergene distance <1
330 megabase (Mb) (Mann-Whitney test, p-value < 0.05) (Table 3). The FR and RF
331 orientations of TFs included known TFs involved in chromatin interactions such as
332 CTCF, RAD21, SMC3, YY1, and ZNF143. The top 20 FR and RF orientations of DNA
333 motifs were selected in ascending order of the median correlation coefficient of
334 expression levels of gene pairs separated by the TFBSs of a TF. The 20 TFs included
335 known TFs involved in chromatin interactions such as CTCF, RAD21, and SMC3 (Fig.
336 4) (Mann-Whitney test, p-value < 0.01). Instead of using all TFBSs of a TF, when
337 TFBSs of a TF were limited to sites used in the analysis of biased orientations of DNA
338 motifs in an EPA, 33 of 55 (60%) FR and 36 of 47 (77%) RF orientations showed
339 significantly lower correlations of expression levels of gene pairs separated by the
340 TFBSs (Additional file 1: Table S8). With intergene distance < 20 kb, 21 of 62 (34%)

341 FR and 8 of 52 (15%) RF orientations showed significantly lower correlations of
342 expression levels of gene pairs separated by the TFBSs of a TF than the correlations of
343 all pairs of neighboring genes (Additional file 1: Table S8).

344 A large number of TFBSs of a TF tend to show significantly lower correlations of
345 gene expression (median numbers of TFBSs showing significantly lower correlations
346 were 656 and 479 for FR and RF respectively, and median numbers of TFBSs not
347 showing significantly lower correlations were 122 and 115 for FR and RF respectively.
348 The Mann-Whitney test *p*-value = 0 and < 10^{-10} for FR and RF respectively). This
349 suggests that if the criteria to examine the correlations of expression levels of gene pairs
350 separated by TFBSs are more stringent and if the number of TFBSs is reduced, it would
351 be more difficult to use a statistical test to compare the correlations of expression levels
352 of gene pairs separated by the reduced TFBSs, particularly for TFs with a small number
353 of TFBSs.

354 Among 91 biased orientations of DNA motifs of TFs found in common among
355 the four cell types, after excluding the five known TFs involved in chromatin
356 interactions such as CTCF, RAD21, SMC3, YY1, and ZNF143, TFBSs of 88 TFs were
357 located at $\geq$1 genomic regions between gene pairs not including TFBSs of the known
358 TFs and at $\geq$1 genomic regions between low correlation (< 0.13) of expression level of
359 gene pairs. Ubiquitously expressed genes among 53 tissues with CV < 90 were
360 eliminated. The median correlation coefficient of expression level of all pairs of nearby
361 genes was 0.17 (Table 3). TFBSs of 52 TFs were placed at $\geq$10 genomic regions.
362

363 **Co-location of biased orientations of motifs**

364 To examine the association of biased orientations of DNA-binding motifs of TFs,
365 co-location of motifs within 200 bp (approximately the length of the minimum open
366 chromatin region of 146 bp) of each other was analyzed in monocytes, T cells, HMEC,
367 and NPC. The same pairs of motifs upstream and downstream of genes in EPAs were
368 enumerated, and the pairs of motifs found in $\geq$100 genomic regions were listed (Table
369 4; Additional file 1: Table S9). A majority of the pairs of motifs overlapped with more
370 than 1 base, but the same tendency was found in the analysis of motifs of all TFs
371 (Additional file 1: Table S9). CTCF partially overlapped with cohesin RAD21 and
372 SMC3 (Table 4). The top 30 pairs of FR and RF orientations of motifs overlapped or
373 co-localized have been shown (Table 4). Unique pairs of overlapping or co-localized
374 motifs were 404 FR and 662 RF, consisting of 117 FR and 218 RF of unique DNA
375 motifs in monocytes. Only 10 overlapping or co-localized motifs were found to be
376 common among the four cell types, including pairs of CTCF, RAD21, SMC3, and YY1,
377 thereby implying that most overlapping and co-localized motifs serve in a cell-specific
378 manner (Additional file 1: Table S9). Paired DNA-binding motifs of TFs with biased
379 orientations tend to overlap, and a relatively small number of motifs are co-localized
380 upstream and downstream of genes. Overlapping motifs act as TFBSs for several TFs
381 either simultaneously or competitively [20-22, 77-81]. CTCF and cohesin function
382 cooperatively with another TF and a protein such as Klf4 and Brd2 [82, 83]. Klf4
383 recruits cohesin to the Oct4 enhancer region. BRD2 supports boundary (insulator)
384 activity of CTCF. The analysis of 457 ChIP-seq data sets from 119 human TFs showed

385 that secondary motifs were detected in addition to the canonical motifs of the TFs,
386 indicating tethered binding and co-binding between multiple TFs [84]. Wang et al. [84]
387 observed significant position and orientation preferences between many co-binding TFs.
388 Overlapping motifs found in this study may be biologically functional rather than
389 artifact. Though repeat sequences in the human genome were masked, more than a
390 hundred pairs of motifs overlapped with hundreds or thousands of upstream and
391 downstream gene regions.
392

393 **Discussion**

394 An initial comprehensive method to find DNA-binding motifs of TFs affecting
395 EPIs and gene expression without using chromatin interaction data is presented through
396 this study. DNA motifs, such as CTCF and cohesin (RAD21 and SMC3), were searched
397 from the open chromatin regions of monocytes, T cells, HMEC, and NPC. A total of 96
398 biased (64 FR and 52 RF) orientations of DNA motifs of TFs were found to be common
399 among the four cell types. Non-orientation (i.e., without considering orientation) of
400 motifs was not found to affect the expression level of putative transcriptional target
401 genes, suggesting that an EPA shortened at the FR or RF orientation of TFBSs was a
402 more accurate prediction of EPIs. This was supported by the comparisons with
403 chromatin interaction and gene expression data. FF and RR orientations of motifs were
404 not examined in this analysis since CTCF has FR or RF orientation of its TFBSs.
405 Among 96 biased orientations of DNA motifs of TFs, 14 TFs were reported to be
406 associated with chromatin interactions, and many other TFs had chromatin-associated

407 functions. The enrichment analysis of biased orientations of TFBSs of the 96 TFs, for
408 pairs of chromatin interaction anchors indicated a weak statistical significance, except
409 for CTCF, RAD21, SMC3, ATB3, and ZBTB7A, suggesting that the TFs were difficult
410 to find using DNA motifs of TFs [5].

411 To predict TFBSs in genomic DNA, several tools and methods were explored and
412 then protein interaction quantitation (PIQ) was used [66]. PIQ predicts TFBSs by
413 reducing noise in DNase-seq data, and the comparison of the prediction with ChIP-seq
414 data has revealed that PIQ shows a better performance than other tools and digital
415 genome footprinting (DNase-DGF). The prediction of TFBSs using PIQ in this study
416 increased the number of biased orientations of motifs, thereby reducing the number of
417 non-orientation of motifs. At first, the number of predicted motifs was expected to
418 increase relative to using consensus DNA-binding sequences of TFs and open
419 chromatin regions available in ENCODE database, which seemed to be difficult to find
420 all TFBSs in the genome. However, the number of TFBSs was almost the same or
421 slightly decreased compared with the analysis of consensus sequences, suggesting that
422 PIQ increased true-positive predictions of TFBSs and reduced false-positive
423 predictions.

424 When forming a homodimer or heterodimer with another TF, TFs may bind to
425 genomic DNA with a specific orientation of their TFBSs. The approach in this study
426 would be able to identify TFs forming a heterodimer. If the TFBS of only the mate to a
427 pair of TFs was found in an EPA, the EPA was shortened at one side of the TFBS, and
428 transcriptional target genes were predicted based on the EPA shortened at the side. In

429 this analysis, the mate to both heterodimers and homodimers of TFs can be tested to
430 examine the effect on the expression level of transcriptional target genes predicted
431 based on the EPA shortened at one side. Biased orientations of DNA-binding motifs
432 may also be found in forward-forward or reverse-reverse orientation.

433 TFs that are multi-functional transcriptional regulators such as YY1 may not
434 indicate a strong directional bias of their TFBSs. Though some proportion of TFBSs of
435 a TF involved in chromatin interactions and insulator functions is found in a biased
436 orientation in the genome, the other TFBSs of the TF may not form a dimer with
437 another TF and may not be present in a specific orientation in the genome. Thus, the
438 overall proportion of biased orientations of TFBSs of a TF may be reduced such that a
439 statistical test cannot detect it.

440 The approach presented in this study was tested to search for repeat sequences
441 that serve as insulator sites and affect the expression level of putative transcriptional
442 target genes of TFs. A small number of repeat sequences were identified and included
443 MIR, which act as insulators [85].

444 CTCF and cohesin-binding sites are frequently mutated in cancer [86]. Some
445 biased orientations of DNA motifs involved in chromatin interactions may be associated
446 with diseases including cancer.

447 The analyses conducted here suggest that biased orientations of DNA motifs of
448 discovered TFs affect EPIs and the expression level of putative transcriptional target
449 genes of general TFs bound in enhancers and promoters. This gives insight into the
450 nature of TFs and contributes to the research of functions and molecular mechanisms of

chromatin interactions, regulation of chromatin, histone modifications, transcriptional regulation, and transcriptional cascades.

Conclusion

CTCF and cohesin proteins are located at the anchors of chromatin interactions and form loop structures. Other DNA-binding proteins such as YY1, ZNF143, and SMARCA4 are also reported to be involved in chromatin interactions. However, the overall understanding of proteins associated with chromatin interactions and insulator function remains unclear. The analysis performed here discovered a hundred of FR and RF orientations of DNA motifs of TFs which affected EPIs and gene expression among four cell types. However, without considering the orientations, DNA motifs of TFs affecting EPIs and gene expression were not detected. Among the biased orientations of TFs, KLF4, ERG, RFX, RFX2, HIF1, SP1, STAT3, and AP1 were reported to be associated with chromatin interactions. Many other TFs were also known to have chromatin-associated functions. Most of the TFs showed weak directional biases at chromatin interaction anchors and were difficult to identify based on their enrichment analysis of TFBSs. These findings contribute to investigate chromatin-associated motifs involved in transcriptional regulation, chromatin interactions, regulation of chromatin, and histone modifications.

473 # Methods

474 **Search for biased orientations of DNA motifs**

475 To identify transcription factor binding sites (TFBSs) from open and closed
476 chromatin regions, TRANSFAC (2019.2), JASPAR (2018), UniPROBE (2018),
477 high-throughput SELEX, transcription factor binding sequences of ENCODE ChIP-seq
478 data, and HOCOMOCO version 9 and 11 were used to predict insulator sites [87-94].
479 TRANSFAC (2019.2) was used to analyze enhancer-promoter interactions because
480 these data were sufficient to identify biased orientations of DNA motifs of insulator TFs
481 with less computational time, thereby reducing the number of non-orientation of DNA
482 motifs of TFs. To reduce false-positive predictions of TFBSs in open chromatin regions,
483 the DNA motif search was performed by using the protein interaction quantitation (PIQ)
484 tool with 12,249 position frequency matrices converted from DNA motifs of vertebrate
485 TFs in the above-mentioned databases and DNase-seq data from GEO and ENCODE
486 database (GSM1024791 $CD14^+$ monocytes; GSM665812 $CD4^+$ T cell; GSM736634
487 Human mammary epithelial cell, HMEC; GSM878615 Neural progenitor cell, NPC;
488 ENCFF263GMV, ENCFF886XEV Th17; GSM1024741 Treg; GSM736620 GM12878)
489 [66]. To find open chromatin regions associated with enhancer and promoter activity,
490 narrow peaks of ChIP-seq experimental data of H3K27ac were predicted using MACS2
491 callpeak (GEO: GSM773004) [95]. To predict TFBSs of TFs from closed chromatin
492 regions, another method was employed. Position weight matrices of vertebrate
493 transcription factor binding sequences were converted into DNA consensus sequences,
494 which consist of an alphabet of 15 characters (the four bases A, C, G, T, the six

495 two-fold degenerate IUPAC codes R=[AG], Y=[CT], K=[GT], M=[AC], S=[GC],
496 W=[AT], the four three-fold degenerate IUPAC codes B=[CGT], D=[AGT], H=[ACT],
497 V=[ACG], and the four-fold degenerate character N=[ATGC]) using a convert matrix in
498 the RSAT website [96]. Transcription factor binding sequences were searched from
499 narrow peaks of DNase-seq data in repeat-masked hg19 genome sequences using the
500 Match tool in TRANSFAC with a similarity score cutoff of 1 after transforming the
501 DNA consensus sequences of TFs into position frequency matrixes for the Match search
502 [97]. Repeat-masked hg19 genome sequences were downloaded from the UCSC
503 genome browser (http://genome.ucsc.edu/,
504 http://hgdownload.soe.ucsc.edu/goldenPath/hg19/bigZips/hg19.fa.masked.gz). TFs
505 corresponding to transcription factor binding sequences were searched computationally
506 and manually by comparing their names and gene symbols of HUGO Gene
507 Nomenclature Committee-approved gene nomenclature and 31,848 UCSC canonical
508 transcripts
509 (http://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/knownCanonical.txt.gz)
510 because transcription factor binding sequences were not linked to transcript IDs such as
511 UCSC, RefSeq, and Ensembl transcripts.
512 Target genes of a TF were assigned when its TFBS was found in the DNase-seq
513 narrow peaks in the promoter or extended regions of a gene's EPA. Promoter and
514 extended regions were defined as follows: promoter regions were those that were within
515 a distance of ±5 kb from transcriptional start sites (TSS). Promoter and extended
516 regions were defined as per the following association rule, which is the same as that

517 defined in Figure 3A of a previous study [12]: the single nearest gene association rule,
518 which extends the regulatory domain to the midpoint between the TSS of the gene and
519 that of the nearest gene upstream and downstream without the limitation of extension
520 length. The limitation of extension length was ＜1 Mb in this study. Extended regions
521 for an EPA were shortened at the TFBSs of a putative insulator TF closest to a TSS, and
522 transcriptional target genes were predicted based on the shortened EPA and TFBSs of a
523 general TF in open chromatin regions. Promoter and extended regions of an EPA were
524 shortened at the forward–reverse (FR) orientation of TFBSs of a putative insulator TF.
525 When the forward or reverse orientation of TFBSs was located in the genome several
526 times, the most internal (i.e., closest to a TSS) forward–reverse orientation of the
527 TFBSs was selected. The genomic positions of genes were identified using the
528 'knownGene.txt.gz' file in UCSC bioinformatics sites [98]. The file
529 'knownCanonical.txt.gz' was also utilized to select representative transcripts among
530 various alternate forms for assigning promoter and extended regions of an EPA. From
531 the list of TFBSs and transcriptional target genes, redundant DNA motifs of TFs were
532 removed by comparing the putative target genes of a TFBS and its corresponding TF; if
533 identical, one of the DNA motifs of TF was chosen. When the number of transcriptional
534 target genes predicted based on a TFBS was <5, the DNA motif of TF was omitted.
535 For gene expression data, RNA-seq reads mapped onto the human hg19 genome
536 sequence were obtained from UCSF-UBC human reference epigenome mapping project
537 RNA-seq reads with poly-A of naive $CD4^+$ T cells (GEO: GSM669617). For monocytes,
538 Blueprint RNA-seq FPKM data

539 (‘C0010KB1.transcript_quantification.rsem_grape2_crg.GRCh38.20150622.results’)
540 were downloaded from Blueprint DCC portal
541 (http://dcc.blueprint-epigenome.eu/#/files). RNA-seq reads with the poly-A of human
542 mammary epithelial cells (HMEC) in the Encyclopedia of DNA Elements at UCSC
543 (http://genome.ucsc.edu/encode/,
544 ‘wgEncodeCshlLongRnaSeqHmecCellPapAlnRep1.bam’ file), H1-derived NPC (GEO:
545 GSM915326, ENCODE: ENCFF529SIO), Th17 (GSM2859479, NCBI SRA:
546 SRR6298326), Treg (GSM2859476, SRR6298323), and GM12878 (GSE78551,
547 ENCFF297QCE) were analyzed. RNA-seq reads were aligned in repeat-masked hg19
548 genome sequences using Burrows-Wheeler Aligner with default parameters [99].
549 FPKMs of the RNA-seq data were calculated by RSeQC [100]. Based on
550 log2-transformed FPKM, UCSC transcripts were arranged in a descending order of
551 expression level, and the top 4,000 transcripts were selected in each cell type.

552 The expression level of transcriptional target genes predicted based on EPAs
553 shortened at the TFBSs of a TF was compared with the expression level of
554 transcriptional target genes predicted based on closed chromatin regions in promoters.
555 For each DNA motif shortening EPAs, transcriptional target genes were predicted based
556 on DNA-binding motifs of vertebrate TFs in the TRANSFAC database, and the ratio of
557 expression level of putative transcriptional target genes of each TF was calculated
558 between an EPA and closed chromatin regions in the promoters. The distribution of the
559 ratios of all TFs was compared among forward-reverse (FR), reverse-forward (RF), and
560 non-orientation (i.e., without considering orientation) of TFBSs shortening an EPA

561 using a two-sided Mann-Whitney test, (p-value < 10^{-7}). Other parameters for the
562 analysis were set as follows: the number of transcriptional target genes of a TF was
563 ≥50, the number of general TFs (not insulator TFs) to predict transcriptional target
564 genes was ≥50, and the number of TFBSs of a TF (insulator TF) shortening an EPA
565 was ≥100.
566
567 **Comparison with chromatin interaction data**
568 For comparison of EPIs with chromatin interactions (HiChIP) in $CD4^+$ T, Th17,
569 Treg, and GM12878 cells,
570 'GSM2705049_Naive_HiChIP_H3K27ac_B2T1_allValidPairs.txt',
571 'GSM2705050_Naive_HiChIP_H3K27ac_B2T2_allValidPairs.txt',
572 'GSM2705051_Naive_HiChIP_H3K27ac_B3T1_allValidPairs.txt',
573 'GSM2705054_Th17_HiChIP_H3K27ac_B1T2_allValidPairs.txt',
574 'GSM2705055_Th17_HiChIP_H3K27ac_B2T1_allValidPairs.txt',
575 'GSM2705056_Th17_HiChIP_H3K27ac_B3T1_allValidPairs.txt',
576 'GSM2705059_Treg_HiChIP_H3K27ac_B3T1_allValidPairs.txt',
577 'GSM2705041_GM_HiChIP_H3K27ac_B1_allValidPairs.txt', and
578 'GSM2705042_GM_HiChIP_H3K27ac_B2_allValidPairs.txt' files were downloaded
579 from the Gene Expression Omnibus (GEO) database. The resolutions of chromatin
580 interactions and EPIs were adjusted to 5 kb before comparison. Chromatin interactions
581 with ≥ 6,000 and ≥ 1,000 counts for each interaction were analyzed in the comparison
582 in $CD4^+$ T cells. According to the unique numbers of chromatin interactions in

583 replications in Th17, Treg, and GM12878 cells, chromatin interactions with ≥1,000
584 (B1T2 in Th17), ≥2,000 (B2T1 and B3T1 in Th17), ≥ 6,000 (B3T1 in Treg), and
585 ≥4,000 (B1 and B2 in GM12878) counts for each interaction were analyzed in the
586 comparison.

587 EPIs were predicted on the basis of the following three types of EPAs in
588 monocytes: (i) EPA shortened at the FR or RF orientation of TFBSs of a TF, (ii) EPA
589 shortened at the non-orientation (i.e., without considering orientation) of TFBSs of a TF,
590 and (iii) EPA without being shortened by a TFBS. EPIs predicted based on the three
591 types of EPAs in common were removed. EPIs predicted based on EPA (i) were
592 removed from EPIs predicted based on EPA (ii). EPIs predicted based on EPA (i) and
593 (ii) were removed from EPIs predicted based on EPA (iii). The resolution of HiChIP
594 chromatin interaction data was 1-5 kb; thus, EPIs were adjusted to 5 kb before
595 comparison. The number and ratio of EPIs overlapped with chromatin interactions were
596 compared two times between EPI (i) and (iii), and EPI (i) and (ii) (binomial distribution,
597 p-value < 0.025 for each test, two-sided, 95% confidence interval).

598 For comparison of EPIs with chromatin interactions (in situ Hi-C) in HMEC, the
599 genomic positions of chromatin interactions in the 'HMEC_4DNFI97O9SAZ.pairs.gz'
600 file in 4D nucleome data portal (http://data.4dnucleome.org) were converted into the
601 hg19 version of human genome using the liftOver tool
602 (http://hgdownload.soe.ucsc.edu/downloads.html#liftover). The resolutions of
603 chromatin interaction data and EPIs were adjusted to 3 kb before comparison. All
604 chromatin interactions were used in the comparison. For comparison of EPIs with

605 chromatin interactions (promoter capture Hi-C) in NPC, the resolutions of chromatin
606 interaction data in the 'GSE86189_npc.po.all.txt.bz2' file (GEO: GSE86189) and EPIs
607 were adjusted to 5 kb before comparison. All chromatin interactions were used in the
608 comparison.

609 Putative target genes for the comparison of EPIs and chromatin interactions were
610 selected from top 4,000 transcripts in terms of the expression level. The expression level
611 of putative target genes of EPIs overlapped with chromatin interactions was compared
612 with EPIs not overlapped with them. When the same transcriptional target gene of a TF
613 in an enhancer was predicted based on both an EPI overlapping a chromatin interaction
614 and a non-overlapping EPI, the target gene was eliminated. The distribution of
615 expression level of putative target genes was compared using a two-sided
616 Mann-Whitney test (p-value < 0.05).

617 Micro-C chromatin interaction data in human fibroblasts (4D nucleome data
618 portal ID: 4DNFICOEXGPJ) was converted into the hg19 version of the human genome
619 using the liftOver tool and was adjusted to 200-bp resolution. The same genomic
620 locations of paired chromatin interaction anchors were merged.

621

622 **Correlation of expression level of gene pairs separated by TFBS**

623 To examine the correlation of expression level of genes pairs separated by
624 TFBSs of a TF, gene expression data among 53 tissues were obtained from the
625 'GTEx_Analysis_2016-01-15_v7_RNASeQCv1.1.8_gene_median_tpm.gct.gz' file in
626 the GTEx database (https://gtexportal.org/home/). The correlation of log2-transformed

627 gene expression levels was analyzed based on RefSeq protein-coding transcripts
628 (RefSeq ID with 'NM_'). The same genomic locations of RefSeq transcripts were
629 removed, and TFBSs overlapped with RefSeq transcripts were also eliminated. Genes
630 closest to a TFBS and another gene were analyzed using BEDOPS closest-features
631 [101]. The Mann-Whitney test was performed using R.
632

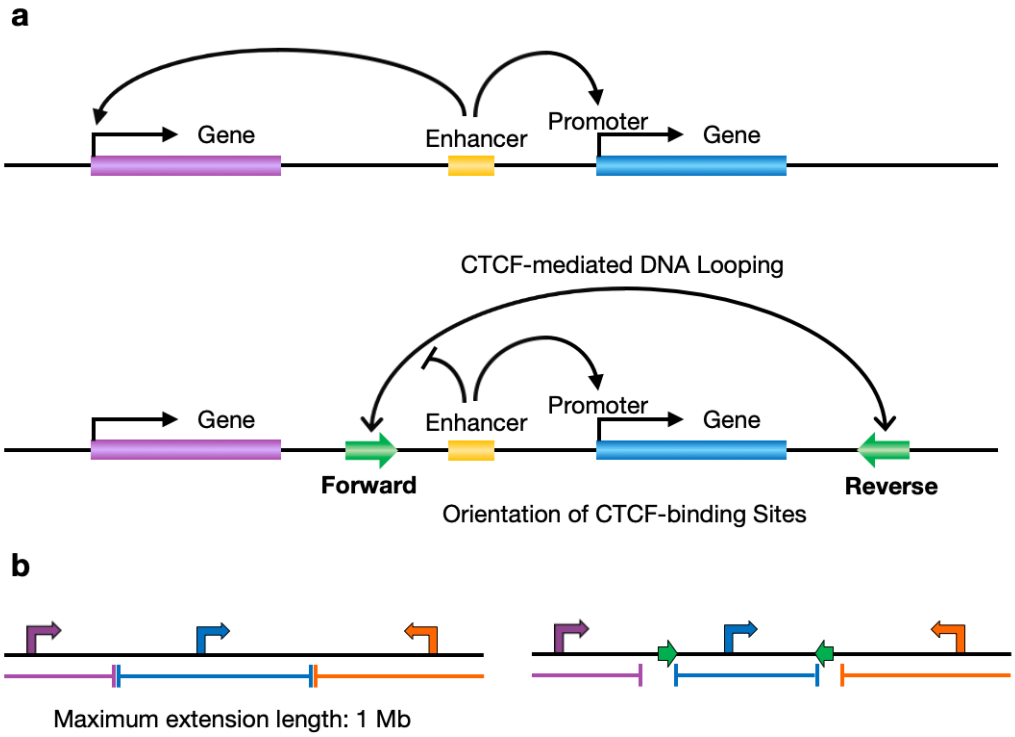
633 ## Abbreviations

634 TF: Transcription factors; TFBS: Transcription factor binding sites; TSS:
635 Transcriptional start sites; EPA: Promoter and extended regions for enhancer-promoter
636 association; EPI: Enhancer-promoter interaction; FF: Forward-forward; FR:
637 Forward-reverse; RF: Reverse-forward; RR: Reverse-reverse; ChIP-seq:
638 ChIP-sequencing, chromatin immunoprecipitation followed by massively parallel DNA
639 sequencing; DNase-seq: DNase I hypersensitive sites sequencing; RNA-seq: RNA
640 sequencing; ENCODE: Encyclopedia of DNA elements;
641

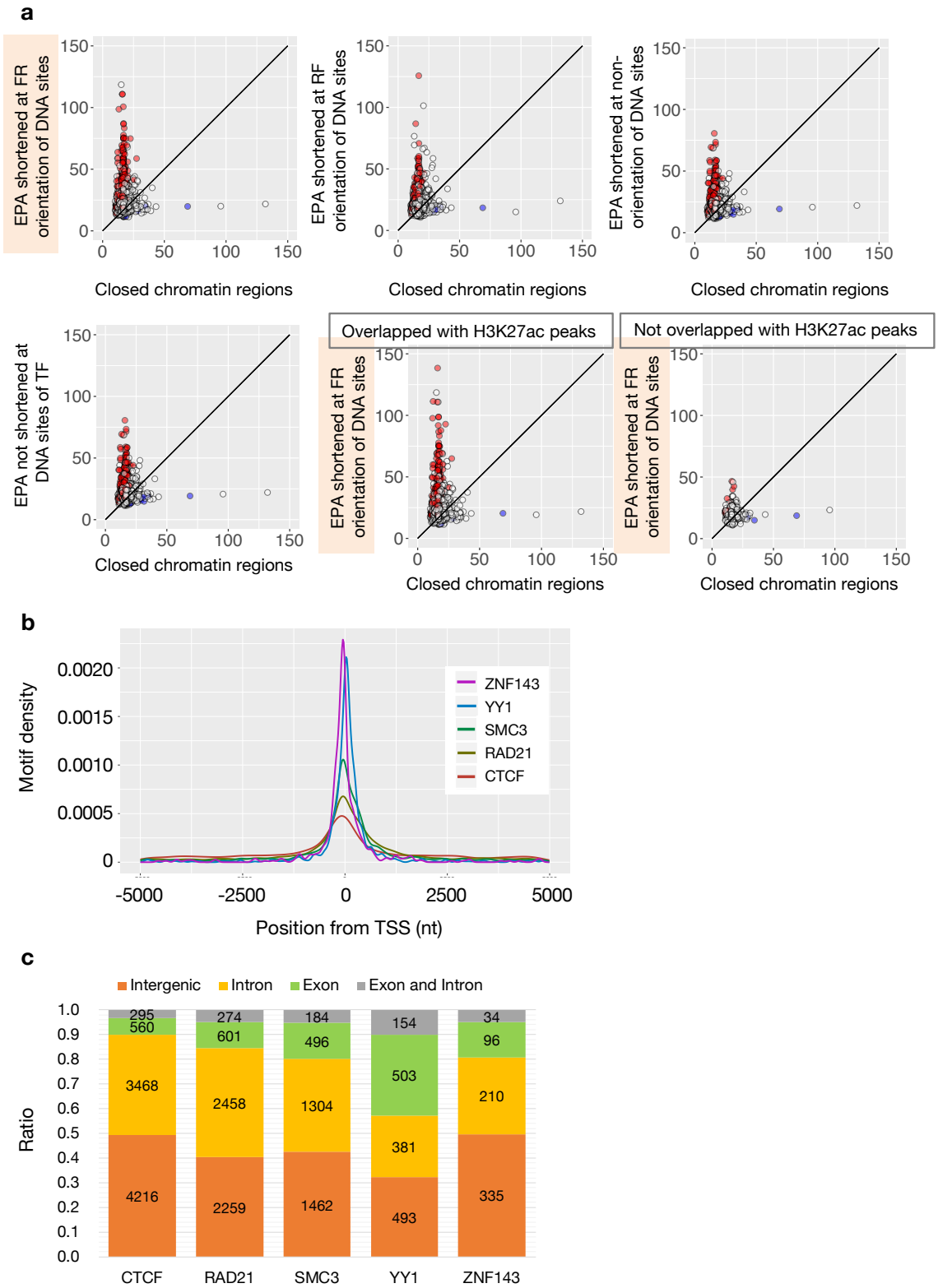
642 **Acknowledgements**

643 Supercomputing resources were provided by the 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

649 **Figures**



650 **Fig. 1** Chromatin interaction and enhancer-promoter association. **a** Forward–reverse
651 orientation of CTCF-binding sites is frequently found in chromatin interaction anchors.
652 CTCF can block the interaction between enhancers and promoters limiting the activity
653 of enhancers to certain functional domains [1-4]. **b** Computationally-defined regulatory
654 domains for enhancer-promoter association (EPA) [12]. The *single nearest gene*
655 association rule extends the regulatory domain to the midpoint between this gene's TSS
656 and the nearest gene's TSS both upstream and downstream. The EPA was shortened at
657 the forward-reverse orientation of TFBSs of a TF in this study (e.g. CTCF in the figure).



658 **Fig. 2** Identification of TFBSs in open chromatin regions and properties of TFBSs. **a**

659 Comparison of expression level of putative transcriptional target genes. Each dot

660 represents the median expression level (FPKM) of target genes predicted based on
661 TFBSs of a TF in closed chromatin regions in promoters (x-axis) and an EPA shortened
662 at the FR orientation of CTCF-binding sites (y-axis) in T cells (upper left graph). The
663 distribution of expression level was changed according to four criteria of EPAs (four
664 graphs) [11]. Red dots indicate the median expression level of target genes is
665 significantly higher in an EPA than that in closed chromatin regions, and blue dots
666 indicate the median expression level of target genes is significantly lower in an EPA
667 than that in closed chromatin regions (Mann-Whitney test). The expression level of
668 target genes predicted on the basis of an EPA tend to be higher than those predicted on
669 the basis of closed chromatin regions, thereby implying that TFs serve as activators of
670 target genes. TFBSs overlapping H3K27ac ChIP-seq peaks were associated with a
671 significant difference of expression level of target genes, and TFBSs not overlapping
672 H3K27ac peaks were not associated with the difference (lower right two graphs). This
673 suggests that the significant difference of expression level is associated with the
674 enhancer (and promoter) signals of H3K27ac. **b** Distribution of TFBSs of TFs near TSS.
675 ZNF143 and YY1 tended to be located near TSS, and CTCF tended to be observed
676 more in distant regions than those around TSS in monocytes. **c** Distribution of TFBSs of
677 TFs in the human genome in monocytes. Most of TFBSs were found in intergenic and
678 intron regions. YY1 is known to be at the TSS in the edge of exons of genes and is
679 involved in transcriptional activation, repression, and initiation.
680

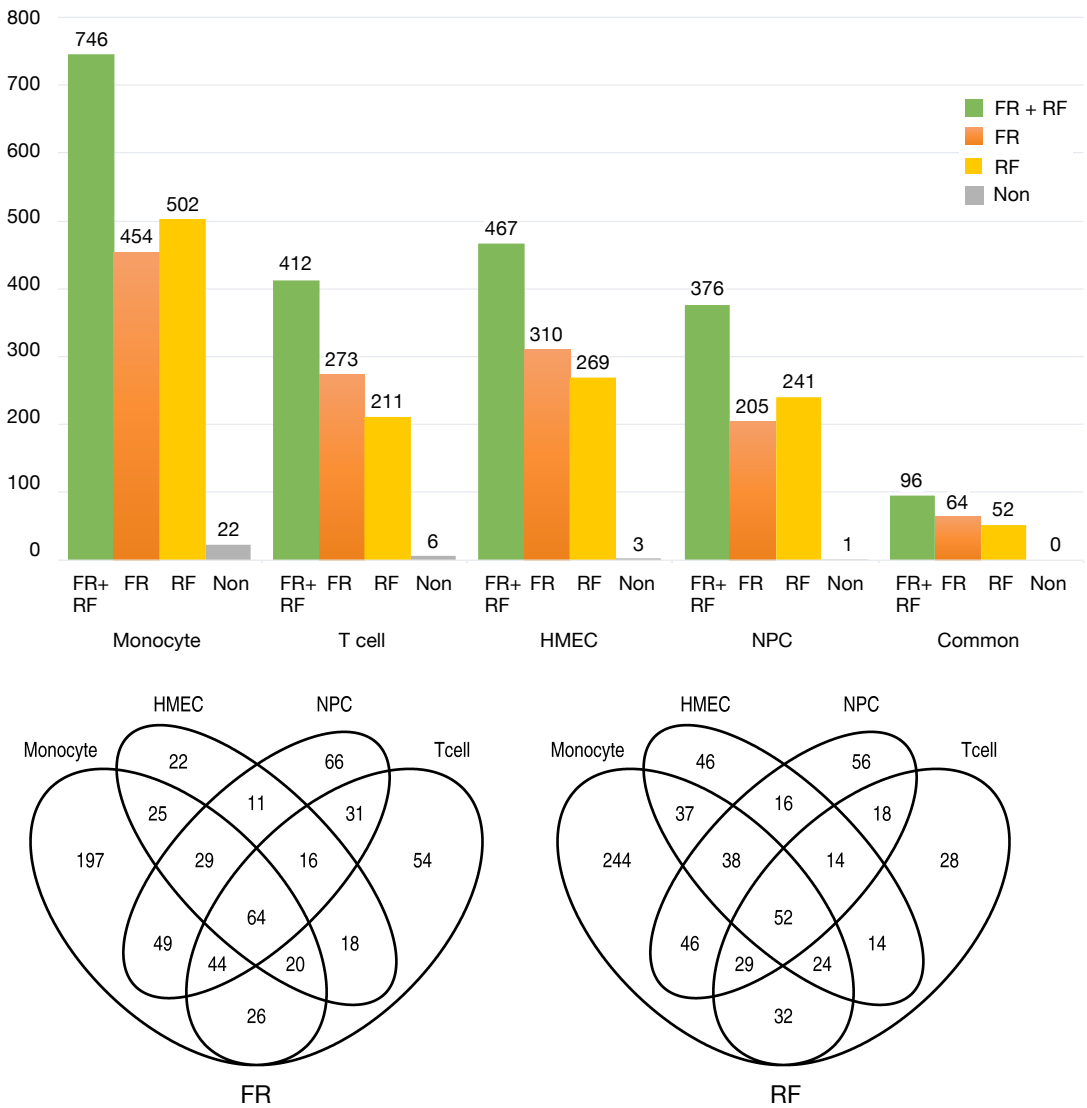


Fig. 3 Biased orientations of TFBSs of TFs affect the transcription of genes by the putative insulator function of the TFs. Ninety-six (sixty-four FR and fifty-two RF) biased orientations of TFBSs of TFs significantly affected the expression level of putative transcriptional target genes in monocytes, T cells, HMEC, and NPC, whereas non-orientation of TFBSs of a TF did not affect the expression level significantly.

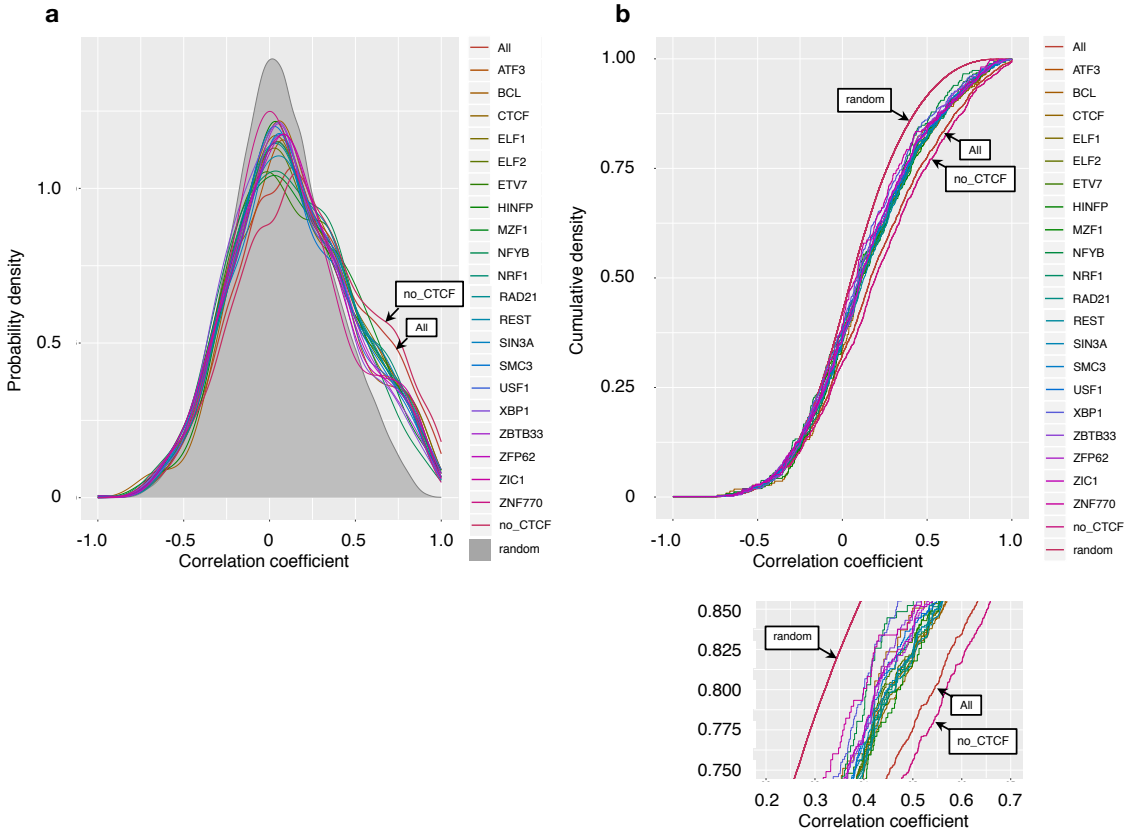


Fig. 4 Genes separated by predicted TFBSs of biased orientations of TFs are less correlated in gene expression. The top 20 TFs are selected according to ascending order of the median correlation coefficient. The correlation coefficient between neighboring gene pairs is shown in terms of probability density **a** and cumulative distribution **b**. The brown line (All) represents the correlation between all neighboring genes; gray shading in **a** and red line (random) in **b** represents correlation between randomly chosen gene pairs; and the purple line (no_CTCF) represents the correlation between gene pairs not separated by predicted CTCF-binding sites.

700 # Tables

701 **Table 1** Biased orientations of DNA-binding motifs of TFs in common between

702 monocytes, T cells, HMEC, and NPC. TF: DNA-binding motifs of transcription factors.

703

704 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

705

706

707 **Table 2** Comparison of enhancer-promoter interactions (EPIs) with chromatin
708 interactions in T cells. The upper tables show the number of biased orientations of DNA
709 motifs, where a significantly higher ratio of EPIs that were predicted on the basis of an
710 EPA (i) overlapped with HiChIP chromatin interactions than EPA (ii) and (iii) in T cells.
711 The middle tables show 133 FR and 90 RF orientations of DNA motifs found in
712 common among B2T1, B2T2, and B3T1 replications in T cells in the upper table. The
713 lower tables show that among 64 FR and 52 RF orientations of DNA motifs found in
714 common in monocytes, T cells, HMEC, and NPC, 44 (69%) FR and 28 (54%) RF were
715 matched with the analysis of HiChIP chromatin interactions according to three types of
716 EPAs. TF: DNA-binding motifs of a transcription factor. Score: $-\log_{10}$ (*p*-value). Score
717 1: Comparison of EPA (i) (an EPA shortened at the FR or RF orientation of TFBSs)
718 with EPA (iii) (an EPA not shortened). Score 2: Comparison of EPA (i) with EPA (ii)
719 (an EPA shortened at the non-orientation of TFBSs).

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

720

721

722 FR and RF orientations of DNA motifs found to be common among three replications in

723 T cells.

724 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

725

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 to be 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 3 Correlation of the expression level of gene pairs separated by TFBSs of a TF. Pairs of genes separated by 64 FR and 52 RF orientations of TFBSs of TFs found to be common between monocytes, T cells, HMEC, and NPC were analyzed to find significantly lower correlation of their expression level compared to the correlation of all pairs of nearby genes among 53 tissues. Score: $-\log_{10}$ (p-value) of the Mann-Whitney test. Median correlation: median correlation coefficient of expression level of nearby gene pairs separated by TFBSs. The median correlation coefficient of expression level of all pairs of nearby genes was 0.17. # TFBS: number of TFBSs of a

744 TF separating gene pairs with expression data. All TFBSs of a TF were used, except for

745 TFBSs separating pairs of ubiquitously expressed genes with coefficient of variance <

746 90.

747

748 Forward-reverse (FR)

TF	Score	Median correlation	# TFBS
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	# TFBS
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

749

750 **Table 4** Top 30 of the co-localized and overlapping biased orientations of DNA-binding
751 motifs of TFs in monocytes. Co-localized and overlapping DNA motifs of CTCF with
752 another biased orientation of DNA motifs are shown in separate tables. Motif 1 and
753 Motif 2: DNA-binding motifs of TFs. No.: the number of genomic locations with both
754 Motif 1 and Motif 2 upstream and downstream of genes in an EPA (< 1 Mb distance
755 from TSS of genes). The total number of genes was 30,523.

756

757 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

758

759

760 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-localized and overlapping DNA motifs of CTCF with other 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

770 ## References

771

772 1. Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT,
773 Sanborn AL, Machol I, Omer AD, Lander ES, Aiden EL: **A 3D map of the**
774 **human genome at kilobase resolution reveals principles of chromatin**
775 **looping.** *Cell* 2014, **159:**1665-1680.
776 2. Shlyueva D, Stampfel G, Stark A: **Transcriptional enhancers: from**
777 **properties to genome-wide predictions.** *Nat Rev Genet* 2014, **15:**272-286.
778 3. Guo Y, Xu Q, Canzio D, Shou J, Li J, Gorkin DU, Jung I, Wu H, Zhai Y, Tang
779 Y, et al: **CRISPR Inversion of CTCF Sites Alters Genome Topology and**
780 **Enhancer/Promoter Function.** *Cell* 2015, **162:**900-910.
781 4. de Wit E, Vos ES, Holwerda SJ, Valdes-Quezada C, Verstegen MJ, Teunissen H,
782 Splinter E, Wijchers PJ, Krijger PH, de Laat W: **CTCF Binding Polarity**
783 **Determines Chromatin Looping.** *Mol Cell* 2015, **60:**676-684.
784 5. Bailey SD, Zhang X, Desai K, Aid M, Corradin O, Cowper-Sal Lari R,
785 Akhtar-Zaidi B, Scacheri PC, Haibe-Kains B, Lupien M: **ZNF143 provides**
786 **sequence specificity to secure chromatin interactions at gene promoters.** *Nat*
787 *Commun* 2015, **2:**6186.
788 6. Weintraub AS, Li CH, Zamudio AV, Sigova AA, Hannett NM, Day DS,
789 Abraham BJ, Cohen MA, Nabet B, Buckley DL, et al: **YY1 Is a Structural**
790 **Regulator of Enhancer-Promoter Loops.** *Cell* 2017, **171:**1573-1588 e1528.
791 7. Barutcu AR, Lajoie BR, Fritz AJ, McCord RP, Nickerson JA, van Wijnen AJ,
792 Lian JB, Stein JL, Dekker J, Stein GS, Imbalzano AN: **SMARCA4 regulates**
793 **gene expression and higher-order chromatin structure in proliferating**
794 **mammary epithelial cells.** *Genome Res* 2016, **26:**1188-1201.
795 8. Schreiber J, Libbrecht M, Bilmes J, Noble W: **Nucleotide sequence and**
796 **DNaseI sensitivity are predictive of 3D chromatin architecture.** *bioRxiv*
797 2017.
798 9. Zhang H, Li F, Jia Y, Xu B, Zhang Y, Li X, Zhang Z: **Characteristic**
799 **arrangement of nucleosomes is predictive of chromatin interactions at**
800 **kilobase resolution.** *Nucleic Acids Res* 2017, **45:**12739-12751.
801 10. Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, Lander ES,

802 Kellis M: **Systematic discovery of regulatory motifs in human promoters**
803 **and 3' UTRs by comparison of several mammals.** *Nature* 2005, **434:**338-345.
804 11. Osato N: **Characteristics of functional enrichment and gene expression level**
805 **of human putative transcriptional target genes.** *BMC Genomics* 2018,
806 **19:**957.
807 12. McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger
808 AM, Bejerano G: **GREAT improves functional interpretation of**
809 **cis-regulatory regions.** *Nat Biotechnol* 2010, **28:**495-501.
810 13. Xi H, Yu Y, Fu Y, Foley J, Halees A, Weng Z: **Analysis of overrepresented**
811 **motifs in human core promoters reveals dual regulatory roles of YY1.**
812 *Genome Res* 2007, **17:**798-806.
813 14. Shrivastava A, Calame K: **An analysis of genes regulated by the**
814 **multi-functional transcriptional regulator Yin Yang-1.** *Nucleic Acids Res*
815 1994, **22:**5151-5155.
816 15. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF,
817 Schroth GP, Burge CB: **Alternative isoform regulation in human tissue**
818 **transcriptomes.** *Nature* 2008, **456:**470-476.
819 16. Chen M, Manley JL: **Mechanisms of alternative splicing regulation: insights**
820 **from molecular and genomics approaches.** *Nat Rev Mol Cell Biol* 2009,
821 **10:**741-754.
822 17. Das D, Clark TA, Schweitzer A, Yamamoto M, Marr H, Arribere J, Minovitsky
823 S, Poliakov A, Dubchak I, Blume JE, Conboy JG: **A correlation with exon**
824 **expression approach to identify cis-regulatory elements for tissue-specific**
825 **alternative splicing.** *Nucleic Acids Res* 2007, **35:**4845-4857.
826 18. Tabuchi TM, Deplancke B, Osato N, Zhu LJ, Barrasa MI, Harrison MM,
827 Horvitz HR, Walhout AJ, Hagstrom KA: **Chromosome-biased binding and**
828 **gene regulation by the Caenorhabditis elegans DRM complex.** *PLoS Genet*
829 2011, **7:**e1002074.
830 19. Jolma A, Yin Y, Nitta KR, Dave K, Popov A, Taipale M, Enge M, Kivioja T,
831 Morgunova E, Taipale J: **DNA-dependent formation of transcription factor**
832 **pairs alters their binding specificity.** *Nature* 2015, **527:**384-388.
833 20. Ansari AZ, Peterson-Kaufman KJ: **A partner evokes latent differences**
834 **between Hox proteins.** *Cell* 2011, **147:**1220-1221.

835 21. Slattery M, Riley T, Liu P, Abe N, Gomez-Alcala P, Dror I, Zhou T, Rohs R,
836 Honig B, Bussemaker HJ, Mann RS: **Cofactor binding evokes latent**
837 **differences in DNA binding specificity between Hox proteins.** *Cell* 2011,
838 **147:**1270-1282.
839 22. Merabet S, Mann RS: **To Be Specific or Not: The Critical Relationship**
840 **Between Hox And TALE Proteins.** *Trends Genet* 2016, **32:**334-347.
841 23. Di Giammartino DC, Kloetgen A, Polyzos A, Liu Y, Kim D, Murphy D,
842 Abuhashem A, Cavaliere P, Aronson B, Shah V, et al: **KLF4 is involved in the**
843 **organization and regulation of pluripotency-associated three-dimensional**
844 **enhancer networks.** *Nat Cell Biol* 2019, **21:**1179-1190.
845 24. Li F, Yuan Q, Di W, Xia X, Liu Z, Mao N, Li L, Li C, He J, Li Y, et al: **ERG**
846 **orchestrates chromatin interactions to drive prostate cell fate**
847 **reprogramming.** *J Clin Invest* 2020.
848 25. Zhang Z, Chng KR, Lingadahalli S, Chen Z, Liu MH, Do HH, Cai S, Rinaldi N,
849 Poh HM, Li G, et al: **An AR-ERG transcriptional signature defined by**
850 **long-range chromatin interactomes in prostate cancer cells.** *Genome Res*
851 2019, **29:**223-235.
852 26. Majumder P, Gomez JA, Chadwick BP, Boss JM: **The insulator factor CTCF**
853 **controls MHC class II gene expression and is required for the formation of**
854 **long-distance chromatin interactions.** *J Exp Med* 2008, **205:**785-798.
855 27. Quigley IK, Kintner C: **Rfx2 Stabilizes Foxj1 Binding at Chromatin Loops to**
856 **Enable Multiciliated Cell Gene Expression.** *PLoS Genet* 2017, **13:**e1006538.
857 28. Mimura I, Nangaku M, Kanki Y, Tsutsumi S, Inoue T, Kohro T, Yamamoto S,
858 Fujita T, Shimamura T, Suehiro J, et al: **Dynamic change of chromatin**
859 **conformation in response to hypoxia enhances the expression of GLUT3**
860 **(SLC2A3) by cooperative interaction of hypoxia-inducible factor 1 and**
861 **KDM3A.** *Mol Cell Biol* 2012, **32:**3018-3032.
862 29. Deshane J, Kim J, Bolisetty S, Hock TD, Hill-Kapturczak N, Agarwal A: **Sp1**
863 **regulates chromatin looping between an intronic enhancer and distal**
864 **promoter of the human heme oxygenase-1 gene in renal cells.** *J Biol Chem*
865 2010, **285:**16476-16486.
866 30. Park JH, Choi Y, Song MJ, Park K, Lee JJ, Kim HP: **Dynamic Long-Range**
867 **Chromatin Interaction Controls Expression of IL-21 in CD4+ T Cells.** *J*

868 *Immunol* 2016, **196:**4378-4389.
869 31. D'Ippolito AM, McDowell IC, Barrera A, Hong LK, Leichter SM, Bartelt LC,
870 Vockley CM, Majoros WH, Safi A, Song L, et al: **Pre-established Chromatin**
871 **Interactions Mediate the Genomic Response to Glucocorticoids.** *Cell Syst*
872 2018, **7:**146-160 e147.
873 32. Knoepfler PS, Zhang XY, Cheng PF, Gafken PR, McMahon SB, Eisenman RN:
874 **Myc influences global chromatin structure.** *EMBO J* 2006, **25:**2723-2734.
875 33. Sabo A, Amati B: **Genome recognition by MYC.** *Cold Spring Harb Perspect*
876 *Med* 2014, **4**.
877 34. Yang H, Schramek D, Adam RC, Keyes BE, Wang P, Zheng D, Fuchs E: **ETS**
878 **family transcriptional regulators drive chromatin dynamics and**
879 **malignancy in squamous cell carcinomas.** *Elife* 2015, **4:**e10870.
880 35. Li L, Bhatia R: **The controversial role of Sirtuins in tumorigenesis - SIRT7**
881 **joins the debate.** *Cell Res* 2013, **23:**10-12.
882 36. Lee JW, Kim DM, Jang JW, Park TG, Song SH, Lee YS, Chi XZ, Park IY,
883 Hyun JW, Ito Y, Bae SC: **RUNX3 regulates cell cycle-dependent chromatin**
884 **dynamics by functioning as a pioneer factor of the restriction-point.** *Nat*
885 *Commun* 2019, **10:**1897.
886 37. Huang S, Li X, Yusufzai TM, Qiu Y, Felsenfeld G: **USF1 recruits histone**
887 **modification complexes and is critical for maintenance of a chromatin**
888 **barrier.** *Mol Cell Biol* 2007, **27:**7991-8002.
889 38. Li Y, Schulz VP, Deng C, Li G, Shen Y, Tusi BK, Ma G, Stees J, Qiu Y, Steiner
890 LA, et al: **Setd1a and NURF mediate chromatin dynamics and gene**
891 **regulation during erythroid lineage commitment and differentiation.**
892 *Nucleic Acids Res* 2016, **44:**7173-7188.
893 39. Swanson HI: **DNA binding and protein interactions of the AHR/ARNT**
894 **heterodimer that facilitate gene activation.** *Chem Biol Interact* 2002,
895 **141:**63-76.
896 40. Volkel S, Stielow B, Finkernagel F, Stiewe T, Nist A, Suske G: **Zinc finger**
897 **independent genome-wide binding of Sp2 potentiates recruitment of**
898 **histone-fold protein Nf-y distinguishing it from Sp1 and Sp3.** *PLoS Genet*
899 2015, **11:**e1005102.
900 41. Frob F, Wegner M: **The role of chromatin remodeling complexes in Schwann**

901 **cell development.** *Glia* 2020, **68:**1596-1603.
902 42. Wust HM, Wegener A, Frob F, Hartwig AC, Wegwitz F, Kari V, Schimmel M,
903 Tamm ER, Johnsen SA, Wegner M, Sock E: **Egr2-guided histone H2B**
904 **monoubiquitination is required for peripheral nervous system myelination.**
905 *Nucleic Acids Res* 2020.
906 43. Mager GM, Ward RM, Srinivasan R, Jang SW, Wrabetz L, Svaren J: **Active**
907 **gene repression by the Egr2.NAB complex during peripheral nerve**
908 **myelination.** *J Biol Chem* 2008, **283:**18187-18197.
909 44. Blattler A, Yao L, Wang Y, Ye Z, Jin VX, Farnham PJ: **ZBTB33 binds**
910 **unmethylated regions of the genome associated with actively expressed**
911 **genes.** *Epigenetics Chromatin* 2013, **6:**13.
912 45. Oldfield AJ, Yang P, Conway AE, Cinghu S, Freudenberg JM, Yellaboina S,
913 Jothi R: **Histone-fold domain protein NF-Y promotes chromatin**
914 **accessibility for cell type-specific master transcription factors.** *Mol Cell* 2014,
915 **55:**708-722.
916 46. Nardini M, Gnesutta N, Donati G, Gatta R, Forni C, Fossati A, Vonrhein C,
917 Moras D, Romier C, Bolognesi M, Mantovani R: **Sequence-specific**
918 **transcription factor NF-Y displays histone-like DNA binding and H2B-like**
919 **ubiquitination.** *Cell* 2013, **152:**132-143.
920 47. Laurette P, Strub T, Koludrovic D, Keime C, Le Gras S, Seberg H, Van Otterloo
921 E, Imrichova H, Siddaway R, Aerts S, et al: **Transcription factor MITF and**
922 **remodeller BRG1 define chromatin organisation at regulatory elements in**
923 **melanoma cells.** *Elife* 2015, **4**.
924 48. Levy C, Fisher DE: **Dual roles of lineage restricted transcription factors: the**
925 **case of MITF in melanocytes.** *Transcription* 2011, **2:**19-22.
926 49. Ma C, Karwacki-Neisius V, Tang H, Li W, Shi Z, Hu H, Xu W, Wang Z, Kong
927 L, Lv R, et al: **Nono, a Bivalent Domain Factor, Regulates Erk Signaling and**
928 **Mouse Embryonic Stem Cell Pluripotency.** *Cell Rep* 2016, **17:**997-1007.
929 50. Riggi N, Knoechel B, Gillespie SM, Rheinbay E, Boulay G, Suva ML, Rossetti
930 NE, Boonseng WE, Oksuz O, Cook EB, et al: **EWS-FLI1 utilizes divergent**
931 **chromatin remodeling mechanisms to directly activate or repress enhancer**
932 **elements in Ewing sarcoma.** *Cancer Cell* 2014, **26:**668-681.
933 51. Blais A, Dynlacht BD: **E2F-associated chromatin modifiers and cell cycle**

934 **control.** *Curr Opin Cell Biol* 2007, **19:**658-662.
935 52. Taubert S, Gorrini C, Frank SR, Parisi T, Fuchs M, Chan HM, Livingston DM,
936 Amati B: **E2F-dependent histone acetylation and recruitment of the Tip60**
937 **acetyltransferase complex to chromatin in late G1.** *Mol Cell Biol* 2004,
938 **24:**4546-4556.
939 53. Subtil-Rodriguez A, Vazquez-Chavez E, Ceballos-Chavez M,
940 Rodriguez-Paredes M, Martin-Subero JI, Esteller M, Reyes JC: **The chromatin**
941 **remodeller CHD8 is required for E2F-dependent transcription activation of**
942 **S-phase genes.** *Nucleic Acids Res* 2014, **42:**2185-2196.
943 54. Choi J, Jeon S, Choi S, Park K, Seong RH: **The SWI/SNF chromatin**
944 **remodeling complex regulates germinal center formation by repressing**
945 **Blimp-1 expression.** *Proc Natl Acad Sci U S A* 2015, **112:**E718-727.
946 55. Smith E, Shilatifard A: **The chromatin signaling pathway: diverse**
947 **mechanisms of recruitment of histone-modifying enzymes and varied**
948 **biological outcomes.** *Mol Cell* 2010, **40:**689-701.
949 56. Magri L, Swiss VA, Jablonska B, Lei L, Pedre X, Walsh M, Zhang W, Gallo V,
950 Canoll P, Casaccia P: **E2F1 coregulates cell cycle genes and chromatin**
951 **components during the transition of oligodendrocyte progenitors from**
952 **proliferation to differentiation.** *J Neurosci* 2014, **34:**1481-1493.
953 57. Sun Z, Xu X, He J, Murray A, Sun MA, Wei X, Wang X, McCoig E, Xie E,
954 Jiang X, et al: **EGR1 recruits TET1 to shape the brain methylome during**
955 **development and upon neuronal activity.** *Nat Commun* 2019, **10:**3892.
956 58. Ooi L, Belyaev ND, Miyake K, Wood IC, Buckley NJ: **BRG1 chromatin**
957 **remodeling activity is required for efficient chromatin binding by repressor**
958 **element 1-silencing transcription factor (REST) and facilitates**
959 **REST-mediated repression.** *J Biol Chem* 2006, **281:**38974-38980.
960 59. Cunliffe VT: **Eloquent silence: developmental functions of Class I histone**
961 **deacetylases.** *Curr Opin Genet Dev* 2008, **18:**404-410.
962 60. Xu Y, Sengupta PK, Seto E, Smith BD: **Regulatory factor for X-box family**
963 **proteins differentially interact with histone deacetylases to repress collagen**
964 **alpha2(I) gene (COL1A2) expression.** *J Biol Chem* 2006, **281:**9260-9270.
965 61. Saunders A, Huang X, Fidalgo M, Reimer MH, Jr., Faiola F, Ding J,
966 Sanchez-Priego C, Guallar D, Saenz C, Li D, Wang J: **The SIN3A/HDAC**

967 **Corepressor Complex Functionally Cooperates with NANOG to Promote**
968 **Pluripotency.** *Cell Rep* 2017, **18:**1713-1726.
969 62. Salas-Armenteros I, Perez-Calero C, Bayona-Feliu A, Tumini E, Luna R,
970 Aguilera A: **Human THO-Sin3A interaction reveals new mechanisms to**
971 **prevent R-loops that cause genome instability.** *EMBO J* 2017, **36:**3532-3547.
972 63. Adams GE, Chandru A, Cowley SM: **Co-repressor, co-activator and general**
973 **transcription factor: the many faces of the Sin3 histone deacetylase (HDAC)**
974 **complex.** *Biochem J* 2018, **475:**3921-3932.
975 64. Biddie SC, John S, Sabo PJ, Thurman RE, Johnson TA, Schiltz RL, Miranda TB,
976 Sung MH, Trump S, Lightman SL, et al: **Transcription factor AP1 potentiates**
977 **chromatin accessibility and glucocorticoid receptor binding.** *Mol Cell* 2011,
978 **43:**145-155.
979 65. Orlova A, Wagner C, de Araujo ED, Bajusz D, Neubauer HA, Herling M,
980 Gunning PT, Keseru GM, Moriggl R: **Direct Targeting Options for STAT3**
981 **and STAT5 in Cancer.** *Cancers (Basel)* 2019, **11**.
982 66. Sherwood RI, Hashimoto T, O'Donnell CW, Lewis S, Barkal AA, van Hoff JP,
983 Karun V, Jaakkola T, Gifford DK: **Discovery of directional and**
984 **nondirectional pioneer transcription factors by modeling DNase profile**
985 **magnitude and shape.** *Nat Biotechnol* 2014, **32:**171-178.
986 67. Rothstein M, Simoes-Costa M: **Heterodimerization of TFAP2 pioneer factors**
987 **drives epigenomic remodeling during neural crest specification.** *Genome Res*
988 2020, **30:**35-48.
989 68. Mumbach MR, Satpathy AT, Boyle EA, Dai C, Gowen BG, Cho SW, Nguyen
990 ML, Rubin AJ, Granja JM, Kazane KR, et al: **Enhancer connectome in**
991 **primary human cells identifies target genes of disease-associated DNA**
992 **elements.** *Nat Genet* 2017, **49:**1602-1612.
993 69. Yardimci GG, Ozadam H, Sauria MEG, Ursu O, Yan KK, Yang T, Chakraborty
994 A, Kaul A, Lajoie BR, Song F, et al: **Measuring the reproducibility and**
995 **quality of Hi-C data.** *Genome Biol* 2019, **20:**57.
996 70. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T,
997 Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, et al: **Comprehensive**
998 **mapping of long-range interactions reveals folding principles of the human**
999 **genome.** *Science* 2009, **326:**289-293.

1000 71. Jung I, Schmitt A, Diao Y, Lee AJ, Liu T, Yang D, Tan C, Eom J, Chan M,
1001 Chee S, et al: **A compendium of promoter-centered long-range chromatin**
1002 **interactions in the human genome.** *Nat Genet* 2019, **51:**1442-1449.
1003 72. Krietenstein N, Abraham S, Venev SV, Abdennur N, Gibcus J, Hsieh TS, Parsi
1004 KM, Yang L, Maehr R, Mirny LA, et al: **Ultrastructural Details of**
1005 **Mammalian Chromosome Architecture.** *Mol Cell* 2020, **78:**554-565 e557.
1006 73. Kim DE, Procopio MG, Ghosh S, Jo SH, Goruppi S, Magliozzi F, Bordignon P,
1007 Neel V, Angelino P, Dotto GP: **Convergent roles of ATF3 and CSL in**
1008 **chromatin control of cancer-associated fibroblast activation.** *J Exp Med*
1009 2017, **214:**2349-2368.
1010 74. Ramos Pittol JM, Oruba A, Mittler G, Saccani S, van Essen D: **Zbtb7a is a**
1011 **transducer for the control of promoter accessibility by NF-kappa B and**
1012 **multiple other transcription factors.** *PLoS Biol* 2018, **16:**e2004526.
1013 75. Purmann A, Toedling J, Schueler M, Carninci P, Lehrach H, Hayashizaki Y,
1014 Huber W, Sperling S: **Genomic organization of transcriptomes in mammals:**
1015 **Coregulation and cofunctionality.** *Genomics* 2007, **89:**580-587.
1016 76. Xie X, Mikkelsen TS, Gnirke A, Lindblad-Toh K, Kellis M, Lander ES:
1017 **Systematic discovery of regulatory motifs in conserved regions of the**
1018 **human genome, including thousands of CTCF insulator sites.** *Proc Natl*
1019 *Acad Sci U S A* 2007, **104:**7145-7150.
1020 77. Yoon SO, Chikaraishi DM: **Tissue-specific transcription of the rat tyrosine**
1021 **hydroxylase gene requires synergy between an AP-1 motif and an**
1022 **overlapping E box-containing dyad.** *Neuron* 1992, **9:**55-67.
1023 78. Ackerman SL, Minden AG, Williams GT, Bobonis C, Yeung CY: **Functional**
1024 **significance of an overlapping consensus binding motif for Sp1 and Zif268**
1025 **in the murine adenosine deaminase gene promoter.** *Proc Natl Acad Sci U S A*
1026 1991, **88:**7523-7527.
1027 79. Chatterjee R, Zhao J, He X, Shlyakhtenko A, Mann I, Waterfall JJ, Meltzer P,
1028 Sathyanarayana BK, FitzGerald PC, Vinson C: **Overlapping ETS and CRE**
1029 **Motifs ((G/C)CGGAAGTGACGTCA) preferentially bound by GABPalpha**
1030 **and CREB proteins.** *G3 (Bethesda)* 2012, **2:**1243-1256.
1031 80. He X, Syed KS, Tillo D, Mann I, Weirauch MT, Vinson C: **GABPalpha**
1032 **Binding to Overlapping ETS and CRE DNA Motifs Is Enhanced by**

1033 **CREB1: Custom DNA Microarrays.** *G3 (Bethesda)* 2015, **5:**1909-1918.
1034 81. Kin K, Chen X, Gonzalez-Garay M, Fakhouri WD: **The effect of non-coding**
1035 **DNA variations on P53 and cMYC competitive inhibition at cis-overlapping**
1036 **motifs.** *Hum Mol Genet* 2016, **25:**1517-1527.
1037 82. Wei Z, Gao F, Kim S, Yang H, Lyu J, An W, Wang K, Lu W: **Klf4 organizes**
1038 **long-range chromosomal interactions with the oct4 locus in reprogramming**
1039 **and pluripotency.** *Cell Stem Cell* 2013, **13:**36-47.
1040 83. Hsu SC, Gilgenast TG, Bartman CR, Edwards CR, Stonestrom AJ, Huang P,
1041 Emerson DJ, Evans P, Werner MT, Keller CA, et al: **The BET Protein BRD2**
1042 **Cooperates with CTCF to Enforce Transcriptional and Architectural**
1043 **Boundaries.** *Mol Cell* 2017, **66:**102-116 e107.
1044 84. Wang J, Zhuang J, Iyer S, Lin X, Whitfield TW, Greven MC, Pierce BG, Dong
1045 X, Kundaje A, Cheng Y, et al: **Sequence features and chromatin structure**
1046 **around the genomic regions bound by 119 human transcription factors.**
1047 *Genome Res* 2012, **22:**1798-1812.
1048 85. Wang J, Vicente-Garcia C, Seruggia D, Molto E, Fernandez-Minan A, Neto A,
1049 Lee E, Gomez-Skarmeta JL, Montoliu L, Lunyak VV, Jordan IK: **MIR**
1050 **retrotransposon sequences provide insulators to the human genome.** *Proc*
1051 *Natl Acad Sci U S A* 2015, **112:**E4428-4437.
1052 86. Katainen R, Dave K, Pitkanen E, Palin K, Kivioja T, Valimaki N, Gylfe AE,
1053 Ristolainen H, Hanninen UA, Cajuso T, et al: **CTCF/cohesin-binding sites are**
1054 **frequently mutated in cancer.** *Nat Genet* 2015, **47:**818-821.
1055 87. Wingender E, Dietze P, Karas H, Knuppel R: **TRANSFAC: a database on**
1056 **transcription factors and their DNA binding sites.** *Nucleic Acids Res* 1996,
1057 **24:**238-241.
1058 88. Portales-Casamar E, Thongjuea S, Kwon AT, Arenillas D, Zhao X, Valen E,
1059 Yusuf D, Lenhard B, Wasserman WW, Sandelin A: **JASPAR 2010: the greatly**
1060 **expanded open-access database of transcription factor binding profiles.**
1061 *Nucleic Acids Res* 2010, **38:**D105-110.
1062 89. Newburger DE, Bulyk ML: **UniPROBE: an online database of protein**
1063 **binding microarray data on protein-DNA interactions.** *Nucleic Acids Res*
1064 2009, **37:**D77-82.
1065 90. Zhao Y, Stormo GD: **Quantitative analysis demonstrates most transcription**

1066 **factors require only simple models of specificity.** *Nat Biotechnol* 2011,
1067 **29:**480-483.
1068 91. Jolma A, Yan J, Whitington T, Toivonen J, Nitta KR, Rastas P, Morgunova E,
1069 Enge M, Taipale M, Wei G, et al: **DNA-binding specificities of human**
1070 **transcription factors.** *Cell* 2013, **152:**327-339.
1071 92. Xie Z, Hu S, Blackshaw S, Zhu H, Qian J: **hPDI: a database of experimental**
1072 **human protein-DNA interactions.** *Bioinformatics* 2010, **26:**287-289.
1073 93. Kheradpour P, Kellis M: **Systematic discovery and characterization of**
1074 **regulatory motifs in ENCODE TF binding experiments.** *Nucleic Acids Res*
1075 2014, **42:**2976-2987.
1076 94. Kulakovskiy IV, Vorontsov IE, Yevshin IS, Sharipov RN, Fedorova AD,
1077 Rumynskiy EI, Medvedeva YA, Magana-Mora A, Bajic VB, Papatsenko DA, et
1078 al: **HOCOMOCO: towards a complete collection of transcription factor**
1079 **binding models for human and mouse via large-scale ChIP-Seq analysis.**
1080 *Nucleic Acids Res* 2018, **46:**D252-D259.
1081 95. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum
1082 C, Myers RM, Brown M, Li W, Liu XS: **Model-based analysis of ChIP-Seq**
1083 **(MACS).** *Genome Biol* 2008, **9:**R137.
1084 96. van Helden J: **Regulatory sequence analysis tools.** *Nucleic Acids Res* 2003,
1085 **31:**3593-3596.
1086 97. Kel AE, Gossling E, Reuter I, Cheremushkin E, Kel-Margoulis OV, Wingender
1087 E: **MATCH: A tool for searching transcription factor binding sites in DNA**
1088 **sequences.** *Nucleic Acids Res* 2003, **31:**3576-3579.
1089 98. Karolchik D, Barber GP, Casper J, Clawson H, Cline MS, Diekhans M, Dreszer
1090 TR, Fujita PA, Guruvadoo L, Haeussler M, et al: **The UCSC Genome Browser**
1091 **database: 2014 update.** *Nucleic Acids Res* 2014, **42:**D764-770.
1092 99. Li H, Durbin R: **Fast and accurate short read alignment with**
1093 **Burrows-Wheeler transform.** *Bioinformatics* 2009, **25:**1754-1760.
1094 100. Wang L, Wang S, Li W: **RSeQC: quality control of RNA-seq experiments.**
1095 *Bioinformatics* 2012, **28:**2184-2185.
1096 101. Neph S, Kuehn MS, Reynolds AP, Haugen E, Thurman RE, Johnson AK, Rynes
1097 E, Maurano MT, Vierstra J, Thomas S, et al: **BEDOPS: high-performance**
1098 **genomic feature operations.** *Bioinformatics* 2012, **28:**1919-1920.

1099