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

Naoki Osato[1*]

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

[*]Corresponding author

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

Abstract

Chromatin interactions have important roles for enhancer-promoter interactions (EPI) and regulating the transcription of genes. CTCF and cohesin proteins are located at the anchors of chromatin interactions, forming their loop structures. CTCF has insulator function limiting the activity of enhancers into the loops. DNA binding sequences of CTCF indicate their orientation bias at chromatin interaction anchors – forward-reverse (FR) orientation is frequently observed. DNA binding sequences of CTCF were found in open chromatin regions at about 40% - 80% of chromatin interaction anchors in Hi-C and in situ Hi-C experimental data. It has been reported that long range of chromatin interactions tends to include less CTCF at their anchors. It is still unclear what proteins are associated with chromatin interactions.

To find DNA binding motif sequences of transcription factors (TF) such as CTCF affecting the interaction between enhancers and promoters of transcriptional target genes and their expression, here, I predicted human transcriptional target genes of TF bound in open chromatin regions in enhancers and promoters in monocytes and other cell types using experimental data in public database. Transcriptional target genes were predicted based on enhancer-promoter association (EPA). EPA was shortened at the genomic locations of FR orientation of DNA binding motifs of TF, which were supposed to be at chromatin interaction anchors.

The expression level of the target genes predicted based on EPA was compared with target genes predicted from only promoters. Total 351 biased orientation of DNA motifs (192 forward-reverse (FR) and 179 reverse-forward (RF) orientation, the reverse

complement sequences of some DNA motifs are also registered in databases, so the total number is smaller than the number of FR and RF) affected the expression level of putative transcriptional target genes significantly in monocytes of four people in common. Moreover, EPI predicted using FR or RF orientation of some DNA motifs were overlapped with chromatin interaction data (Hi-C) more than the other EPA (Total 62 biased orientation of DNA motifs, 41 FR and 24 RF showed this result).

Keywords: transcriptional target genes, gene expression, transcription factors, enhancer, enhancer-promoter interactions, chromatin interactions, CTCF, cohesion, open chromatin regions

Background

Chromatin interactions have important roles for enhancer-promoter interactions (EPI) and regulating the transcription of genes. CTCF and cohesin proteins are located at the anchors of chromatin interactions, forming their loop structures. CTCF has insulator function limiting the activity of enhancers into the loops (Fig. 1A). DNA binding sequences of CTCF indicate their orientation bias at chromatin interaction anchors – forward-reverse (FR) orientation is frequently observed (de Wit et al. 2015; Guo et al. 2015). About 40% - 80% of chromatin interaction anchors of Hi-C and in situ Hi-C experiments include DNA binding motif sequences of CTCF. However, it has been reported that long range of chromatin interactions tends to include less CTCF at their anchors (Jeong et al. 2017). Other DNA binding proteins such as ZNF143, YY1, and

55 SMARCA4 (BRG1) are found to be associated with chromatin interactions and EPI
56 (Bailey et al. 2015; Barutcu et al. 2016; Weintraub et al. 2017). CTCF, cohesin, ZNF143,
57 YY1 and SMARCA4 have other biological functions as well as chromatin interactions
58 and enhancer-promoter interactions. The DNA binding motif sequences of the
59 transcription factors are found in open chromatin regions near transcriptional start sites
60 (TSS) as well as chromatin interaction anchors.

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

68 DNA binding motif sequence of YY1 does not seem to be enriched at both
69 chromatin interaction anchors (Z-score < 2), whereas DNA binding motif sequence of
70 ZNF143 is significantly enriched (Z-score > 7; Bailey et al. 2015 Figure 2a). In the
71 analysis of YY1, to identify a protein factor that might contribute to EPI, (Ji et al. 2015)
72 performed chromatin immune precipitation with mass spectrometry (ChIP-MS), using
73 antibodies directed toward histones with modifications characteristic of enhancer and
74 promoter chromatin (H3K27ac and H3K4me3, respectively). Of 26 transcription factors
75 that occupy both enhancers and promoters, four are essential based on a CRISPR
76 cell-essentiality screen and two (CTCF, YY1) are expressed in >90% of tissues

77 examined (Weintraub et al. 2017). These analyses started from the analysis of histone
78 modifications of an enhancer mark rather than chromatin interactions. Other protein
79 factors associated with chromatin interactions may be found from other researches.

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

85 DNA binding proteins involved in chromatin interactions are supposed to affect
86 the transcription of genes in the loop formed by chromatin interactions. In my previous
87 analyses, the expression level of human putative transcriptional target genes was
88 affected, according to the criteria of enhancer-promoter association (EPA) (Fig. 1B;
89 (Osato 2018)). EPI predicted based on EPA shortened at the genomic locations of
90 forward-reverse orientation of CTCF binding sites affected the expression level of
91 putative transcriptional target genes the most, compared with the expression level of
92 transcriptional target genes predicted from only promoters (Fig. 1B). The expression
93 level tended to be increased in monocytes and $CD4^+$ T cells, implying that enhancers
94 activated the transcription of genes, and decreased in ES and iPS cells, implying that
95 enhancers repressed the transcription of genes. These analyses suggested that enhancers
96 affected the transcription of genes significantly, when EPI were predicted properly.
97 Other DNA binding proteins involved in chromatin interactions, as well as CTCF,
98 would locate at chromatin interaction anchors with a pair of biased orientation of DNA

99 binding motif sequences, affecting the expression level of putative transcriptional target
100 genes in the loop formed by the chromatin interactions. As experimental issues of the
101 analyses of chromatin interactions, chromatin interaction data are changed, according to
102 experimental techniques, depth of DNA sequencing, and even replication sets of the
103 same cell type. Chromatin interaction data may not be saturated enough to cover all
104 chromatin interactions. Supposing these characters of DNA binding proteins associated
105 with chromatin interactions and avoiding experimental issues of the analyses of
106 chromatin interactions, here I searched for DNA motif sequences of transcription factors
107 and DNA repeat sequences, affecting EPI and the expression level of putative
108 transcriptional target genes in monocytes of four people and other cell types without
109 using chromatin interaction data. Then, putative EPI were compared with chromatin
110 interaction data.
111

112 ## Results

113 **Search for biased orientation of DNA motif sequences**

114 Transcription factor binding sites (TFBS) were predicted using open chromatin
115 regions and DNA motif sequences of transcription factors collected from various
116 databases and journal papers (see Methods). Transcriptional target genes were predicted
117 using TFBS in promoters and enhancer-promoter association (EPA) shortened at the
118 genomic locations of DNA binding motif sequences of transcription factors such as
119 CTCF and cohesins (RAD21 and SMC3). To find DNA motif sequences of transcription
120 factors, other than CTCF and cohesin, and DNA repeat sequences affecting the

121 expression level of putative transcriptional target genes, transcriptional target genes of
122 each transcription factor were predicted based on EPA shortened at the genomic
123 locations of DNA motif sequences, and the expression level of the putative
124 transcriptional target genes was compared with those predicted from promoters using
125 Mann Whiteney U test (p-value < 0.05). The number of transcription factors showing a
126 significant difference of expression level of putative transcriptional target genes was
127 counted. To examine whether the orientation of DNA motif sequences shortening EPA
128 affected the number of transcription factors showing a significant difference of
129 expression level of putative transcriptional target genes, the number of the transcription
130 factors was compared among forward-reverse, reverse-forward, and any orientation of
131 DNA motif sequences shortening EPA, using chi-square test (p-value < 0.05). To avoid
132 missing DNA motif sequences showing a relatively weak statistical significance by
133 multiple testing collection, the above analyses were conducted using monocytes of four
134 people independently, and DNA motif sequences found in monocytes of four people in
135 common were selected. Total 351 of biased (192 forward-reverse and 179
136 reverse-forward) orientation of DNA binding motif sequences of transcription factors
137 have been found in monocytes of four people in common, whereas only one any
138 orientation (i.e. without considering orientation) of DNA binding motif sequences have
139 been found in monocytes of four people in common (Fig. 2; Table 1; Supplemental
140 Table S1). Forward-reverse orientation of DNA motif sequences included CTCF,
141 cohesin (RAD21 and SMC3), ZNF143 and YY1, which are associated with chromatin
142 interactions and enhancer-promoter interactions (EPI). SMARCA4 (BRG1) is

143 associated with topologically associated domain (TAD), which is higher-order
144 chromatin organization. The DNA binding motif sequences of SMARCA4 was not
145 registered in the databases used in this study. Forward-reverse orientation of DNA motif
146 sequences included SMRC2, a member of the same SWI/SNF family of proteins as
147 SMARCA4. These results suggested that the expression level of putative transcription
148 target genes of some transcription factors would be different, depending on the genomic
149 locations (enhancers or promoters) of DNA binding motif sequences of the transcription
150 factors in monocytes of four people.

151 To examine whether the biased orientation of DNA binding motif sequences of
152 transcription factors was observed in other cell types, the same analyses were conducted
153 in other cell types. However, for other cell types, experimental data of one sample were
154 available in ENCODE database, so the analyses of DNA motif sequences were
155 performed by comparing with the result in monocytes of four people. Among the biased
156 orientation of DNA binding motif sequences found in monocytes, 79, 76, 91, 61, and 59
157 DNA binding motif sequences were also observed in $CD4^+$ T cells, H1-hESC, iPS,
158 Huvec and MCF-7 respectively, including CTCF and cohesin (RAD21 and SMC3)
159 (Table 2; Supplemental Table S2). The scores of DNA binding motif sequences were the
160 highest in monocytes, and the other cell types showed lower scores. The results of the
161 analysis of DNA motif sequences in $CD20^+$ B cells and macrophages did not include
162 CTCF and cohesin, because these analyses can be utilized in cells where the expression
163 level of putative transcriptional target genes of each transcription factor show a
164 significant difference between promoters and EPA shortened at the genomic locations of

165 DNA motif sequences. Some experimental data of a cell did not show a significant
166 difference between promoters and EPA (Osato 2018).

167 Instead of DNA binding motif sequences of transcription factors, DNA repeat
168 sequences were also examined. The expression level of transcriptional target genes
169 predicted based on EPA shortened at the genomic locations of DNA repeat sequences
170 was compared with the expression level of transcriptional target genes predicted from
171 promoters. Three reverse-forward orientation of DNA repeat sequences showed a
172 significant difference of expression level of putative transcriptional target genes in
173 monocytes of four people in common (Table 3). Among them, LTR16C DNA repeat
174 sequence was observed in iPS and H1-hESC.

175

176 **Co-location of biased orientation of DNA motif sequences**

177 To examine association of the 351 biased (forward-reverse and reverse-forward)
178 orientation of DNA binding motif sequences of transcription factors. Co-location of the
179 DNA binding motif sequences in open chromatin regions in monocytes was analyzed.
180 The number of open chromatin regions with the same pairs of DNA binding motif
181 sequences was counted, and when the pairs of DNA binding motif sequences were
182 enriched with statistical significance (chi-square test, p-value $< 1.0 \times 10^{-10}$), they were
183 listed (Table 4; Supplemental Table S3). Open chromatin regions overlapped with
184 histone modification of an enhancer mark (H3K27ac) (total 26,095 regions) showed a
185 larger number of enriched pairs of DNA motifs than all open chromatin regions (Table
186 4; Supplemental Table S3). H3K27ac is known to be enriched at chromatin interaction

187 anchors (Phanstiel et al. 2017). As already known, CTCF was found with cohesin such
188 as RAD21 and SMC3 (Table 4). Top 30 pairs of FR and RF orientations of DNA motifs
189 co-occupied in the same open chromatin regions were shown (Table 4). Total number of
190 pairs of DNA motifs was 601, consisting of 155 kinds of DNA motifs, when the pair of
191 DNA motifs were observed in more than 80% of the number of open chromatin regions
192 with the DNA motifs.
193
194 **Comparison with chromatin interaction data**
195 To examine whether the biased orientation of DNA motif sequences is associated
196 with chromatin interactions, enhancer-promoter interactions (EPI) predicted based on
197 enhancer-promoter associations (EPA) were compared with chromatin interaction data
198 (Hi-C). Due to the resolution of Hi-C experimental data used in this study (50kb), EPI
199 were adjusted to 50kb resolution. EPI were predicted based on three types of EPA: (i)
200 EPA without being shortened by DNA motif sequences, (ii) EPA shortened at the
201 genomic locations of DNA motif sequences of transcription factors such as CTCF and
202 cohesin (RAD21, SMC3) without considering their orientation, and (iii) EPA shortened
203 at the genomic locations of forward-reverse or reverse-forward orientation of DNA
204 motif sequences of transcription factors. EPA (iii) showed the highest ratio of EPI
205 overlapped with chromatin interactions (Hi-C) using DNA binding motif sequences of
206 CTCF and cohesin (RAD21 and SMC3), compared with the other two types of EPA
207 (Figure 3). Total 62 biased orientation (41 FR and 24 RF) of DNA motif sequences
208 including CTCF and cohesin showed a higher ratio of EPI overlapped with Hi-C than

209 the other types of EPA in monocytes (Table 5). Chromatin interaction data were
210 obtained using different samples from DNase-seq, open chromatin regions, so
211 individual differences seemed to be large from the results of this analysis. Since, for
212 some DNA motif sequences of transcription factors, the number of EPI overlapped with
213 chromatin interactions was small, if higher resolution of chromatin interaction data
214 (such as Hi-ChIP, in situ DNase Hi-C, and in situ Hi-C data, or a tool to improve the
215 resolution such as HiCPlus) is available, the number of EPI overlapped with chromatin
216 interactions would be increased and the difference of the numbers among three types of
217 EPA might be larger and more significant than the current results (Rao et al. 2014;
218 Ramani et al. 2016; Mumbach et al. 2017; Zhang et al. 2018).
219

220 **Discussion**

221 To find DNA motif sequences of transcription factors and DNA repeat sequences
222 affecting human putative transcriptional target genes, DNA motif sequences were
223 searched from open chromatin regions of monocytes of four people. Total 351 biased
224 (192 forward-reverse and 179 reverse-forward) orientation of DNA motif sequences
225 were found in monocytes of four people in common, whereas only one any orientation
226 (i.e. without considering orientation) of DNA motif sequence was found to affect
227 putative transcriptional target genes, suggesting that EPA shortened at the genomic
228 locations of forward-reverse or reverse-forward orientation of DNA motif sequences of
229 transcription factors is an important character for EPI and transcription of genes.

230 When DNA motif sequences were searched from monocytes of one person, a

231 larger number of biased orientation of DNA motif sequences affecting human putative
232 transcriptional target genes has been found. When the number of people, from which
233 experimental data were obtained, was increased, the number of DNA motif sequences
234 found in all people in common decreased and in some cases, known transcription
235 factors involved in chromatin interactions such as CTCF and cohesin (RAD21 and
236 SMC3) were not identified by statistical tests. This would be caused by individual
237 difference of the same cell type, low quality of experimental data, and experimental
238 errors. Moreover, though forward-reverse orientation of DNA binding motif sequences
239 of CTCF and cohesin is frequently observed at chromatin interaction anchors, the
240 percentage of forward-reverse orientation is not 100, and other orientations of the DNA
241 binding motif sequences are also observed. Though DNA binding motif sequences of
242 CTCF and cohesins are found in various open chromatin regions, DNA binding motif
243 sequences of some transcription factors would be observed less frequently in open
244 chromatin regions. The analyses of cells of a number of people would avoid missing
245 relatively weak statistical significance of DNA motif sequences in cells of each person
246 by multiple testing correction of thousands of statistical tests. A DNA motif sequence
247 was found with p-value < 0.05 in cells of one person and the DNA motif sequence
248 found in the same cell type of four people in common would have p-value $< (0.05)^4 =$
249 0.00000625. Actually, DNA motif sequences with p-value slightly less than 0.05 in
250 monocytes of one person were observed in monocytes of four people in common.

251 EPI were compared with chromatin interactions (Hi-C) in monocytes. EPA
252 shortened at the genomic locations of DNA binding motif sequences of CTCF and

253 cohesin (RAD21 and SMC3) showed a significant difference of the ratios of EPI
254 overlapped with chromatin interactions, according to three types of EPA (FR, any, and
255 others). Using open chromatin regions overlapped with ChIP-seq experimental data of
256 histone modification of an enhancer mark (H3K27ac), the ratio of EPI not overlapped
257 with Hi-C was reduced and the difference of the ratio of EPI overlapped with chromatin
258 interactions among the three types of EPA became larger (data not shown). (Phanstiel et
259 al. 2017) also reported that there was an especially strong enrichment for loops with
260 H3K27 acetylation peaks at both ends (Fisher's Exact Test, $p = 1.4 \times 10^{-27}$). However,
261 the total number of EPI overlapped with chromatin interactions was also reduced using
262 H3K27ac peaks, so more chromatin interaction data would be needed to obtain reliable
263 results in this study. As an issue of experimental data, chromatin interaction data were
264 obtained using different samples from DNase-seq, open chromatin regions, so
265 individual differences were not reflected to chromatin interaction data. Moreover, the
266 resolution of chromatin interaction data used in this study is about 50kb, thus the
267 number of chromatin interactions was relatively small (72,284 at 50kb resolution with
268 cutoff score of CHiCAGO tool > 1). EPA shortened at the genomic locations of DNA
269 binding motif sequences of transcription factors found in various open chromatin
270 regions such as CTCF and cohesin (RAD21 and SMC3) tend to be overlapped with a
271 larger number of chromatin interactions than DNA binding motif sequences of
272 transcription factors less frequently observed in open chromatin regions. Therefore, to
273 examine the difference of the numbers of EPI overlapped with chromatin interactions,
274 according to three types of EPA (FR, any, and others), the number of chromatin

275 interactions should be large enough. If the resolution of chromatin interaction data
276 would be increased by using other experimental data such as Hi-ChIP, in situ DNase
277 Hi-C, and in situ Hi-C data, or a tool to improve the resolution of chromatin interaction
278 data such as HiCPlus, more DNA binding motif sequences of transcription factors less
279 frequently observed in open chromatin regions would be examined (Rao et al. 2014;
280 Ramani et al. 2016; Mumbach et al. 2017; Zhang et al. 2018).
281

282 ## Methods

283 **Search for biased orientation of DNA motif sequences**

284 To examine transcriptional regulatory target genes, bed files of hg38 of Blueprint
285 DNase-seq data for Monocytes $CD14^+$ (EGAD00001002286) were obtained from
286 Blueprint web site (http://dcc.blueprint-epigenome.eu/#/home), and the bed files of
287 hg38 were converted into those of hg19 using Batch Coordinate Conversion (liftOver)
288 web site (https://genome.ucsc.edu/util.html). Bed files of hg19 of ENCODE
289 $CD4^+$_Naive_Wb11970640 (GSM1014537; UCSC Accession: wgEncodeEH003156),
290 H1-hESC (GSM816632; UCSC Accession: wgEncodeEH000556), iPSC (GSM816642;
291 UCSC Accession: wgEncodeEH001110), HUVEC (GSM1014528; UCSC Accession:
292 wgEncodeEH002460), and MCF-7 (GSM816627; UCSC Accession:
293 wgEncodeEH000579) were obtained from the ENCODE websites
294 (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeUwDgf/;
295 http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeUwDnase/).
296 To identify transcription factor binding sites (TFBS) from the DNase-seq data,

297 TRANSFAC (2013.2), JASPAR (2010), UniPROBE, BEEML-PBM, high-throughput
298 SELEX, Human Protein-DNA Interactome, and transcription factor binding sequences
299 of ENCODE ChIP-seq data were used (Wingender et al. 1996; Newburger and Bulyk
300 2009; Portales-Casamar et al. 2010; Xie et al. 2010; Zhao and Stormo 2011; Jolma et al.
301 2013; Kheradpour and Kellis 2014). Position weight matrices of transcription factor
302 binding sequences were transformed into TRANSFAC matrices and then into MEME
303 matrices using in-house scripts and transfac2meme in MEME suite (Bailey et al. 2009).
304 Transcription factor binding sequences of transcription factors derived from vertebrates
305 were used for further analyses. Searches were conducted for transcription factor binding
306 sequences from each narrow peak using FIMO with *p*-value threshold of 10^{-5} (Grant et
307 al. 2011). Transcription factors corresponding to transcription factor binding sequences
308 were searched computationally by comparing their names and gene symbols of HGNC
309 (HUGO Gene Nomenclature Committee) -approved gene nomenclature and 31,848
310 UCSC known canonical transcripts
311 (http://hgdownload.cse.ucsc.edu/goldenpath/hg19/database/knownCanonical.txt.gz), as
312 transcription factor binding sequences were not linked to transcript IDs such as UCSC,
313 RefSeq, and Ensembl transcripts.

314 Target genes of a transcription factor were assigned when its TFBS was found in
315 DNase-seq narrow peaks in promoter or extended regions for enhancer-promoter
316 association of genes (EPA). Promoter and extended regions were defined as follows:
317 promoter regions were those that were within distances of ±5 kb from transcriptional
318 start sites (TSS). Promoter and extended regions were defined as per the following

319 association rule, which is the same as that defined in Figure 3A of a previous study
320 (McLean et al. 2010): the single nearest gene association rule, which extends the
321 regulatory domain to the midpoint between the TSS of the gene and that of the nearest
322 gene upstream and downstream without the limitation of extension length. Extended
323 regions for EPA were shortened at the genomic locations of DNA binding sites of
324 transcription factors that were the closest to a transcriptional start site, and
325 transcriptional target genes were predicted from the shortened enhancer regions using
326 TFBS. Furthermore, promoter and extended regions for EPA were shortened at the
327 genomic locations of forward–reverse orientation of DNA binding sites of transcription
328 factors. When forward or reverse orientation of DNA binding sites were continuously
329 located in genome sequences several times, the most external forward–reverse
330 orientation of DNA binding sites were selected. The genomic positions of genes were
331 identified using 'knownGene.txt.gz' file in UCSC bioinformatics sites (Karolchik et al.
332 2014). The file 'knownCanonical.txt.gz' was also utilized for choosing representative
333 transcripts among various alternate forms for assigning promoter and extended regions
334 for EPA. From the list of transcription factor binding sequences and transcriptional
335 target genes, redundant transcription factor binding sequences were removed by
336 comparing the target genes of a transcription factor binding sequence and its
337 corresponding transcription factor; if identical, one of the transcription factor binding
338 sequences was used. When the number of transcriptional target genes predicted from a
339 transcription factor binding sequence was less than five, the transcription factor binding
340 sequence was omitted.

341 For gene expression data, RNA-seq reads mapped onto human hg19 genome
342 sequences were obtained, including ENCODE long RNA-seq reads with poly-A of
343 H1-hESC, iPSC, HUVEC, IMR90, and MCF-7 (GSE26284, GSM958733,
344 GSM2344099, GSM2344100, GSM958734, and GSM765388), and UCSF-UBC human
345 reference epigenome mapping project RNA-seq reads with poly-A of naive $CD4^+$ T
346 cells (GSM669617). Two replicates were present for H1-hESC, iPSC, HUVEC, and
347 MCF-7, and a single one for $CD4^+$ T cells. RPKMs of the RNA-seq data were
348 calculated using RSeQC (Wang et al. 2012). For monocytes, Blueprint RNA-seq RPKM
349 data (GSE58310, GSE58310_GeneExpression.csv.gz, Monocytes_Day0_RPMI) was
350 used (Saeed et al. 2014). Based on RPKM, UCSC transcripts with expression levels
351 among top 30% of all the transcripts were selected in each cell type.
352 The expression level of transcriptional target genes predicted based on EPA
353 shortened at the genomic locations of DNA motif sequences of transcription factors or
354 DNA repeat sequences was compared with the expression level of transcriptional target
355 genes predicted from promoter. For each DNA motif sequence shortening EPA,
356 transcriptional target genes were predicted using about 3,000 – 5,000 DNA binding
357 motif sequences of transcription factors, and the expression level of putative
358 transcriptional target genes of each transcription factor was compared between EPA and
359 only promoter using Mann-Whitney test (p-value < 0.05). The number of transcription
360 factors showing a significant difference of expression level of putative transcriptional
361 target genes between EPA and promoter was compared among forward-reverse,
362 reverse-forward, and any orientation of DNA motif sequences shortening EPA using

363 chi-square test (p-value < 0.05). When DNA motif sequences of transcription factors or
364 DNA repeat sequences shortening EPA showed a significant difference of expression
365 level of putative transcriptional target genes in forward-reverse, reverse-forward, or any
366 orientation in monocytes of four people in common, the DNA motif sequences were
367 listed.

368 Though forward-reverse orientation of DNA binding motif sequences of CTCF
369 and cohesin is frequently observed at chromatin interaction anchors, the percentage of
370 forward-reverse orientation is not 100, and other orientations of the DNA binding motif
371 sequences are also observed. Though DNA binding motif sequences of CTCF and
372 cohesin are found in various open chromatin regions, DNA binding motif sequences of
373 some transcription factors would be observed less frequently in open chromatin regions.
374 The analyses of cells of a number of people would avoid missing relatively weak
375 statistical significance of DNA motif sequences in cells of each person by multiple
376 testing correction of thousands of statistical tests. A DNA motif sequence was found
377 with p-value < 0.05 in cells of one person and the DNA motif sequence found in the
378 same cell type of four people in common would have p-value < $0.05^4 = 0.00000625$.

379

380 **Co-location of biased orientation of DNA motif sequences**

381 Co-location of biased orientation of DNA binding motif sequences of
382 transcription factors was examined. The number of open chromatin regions with the
383 same pairs of DNA binding motif sequences was counted, and when the pairs of DNA
384 binding motif sequences were enriched with statistical significance (chi-square test,

385 p-value < 1.0×10^{-10}), they were listed.

386

387

388 **Comparison with chromatin interaction data**

389 For chromatin interactions, 'PCHiC_peak_matrix_cutoff0.txt.gz' file was
390 downloaded from 'Promoter Capture Hi-C in 17 human primary blood cell types' web
391 site (https://osf.io/u8tzp/files/), and chromatin interactions for Monocytes with scores of
392 CHiCAGO tool > 1 were extracted from the file (Javierre et al. 2016).

393 EPI were predicted using three types of EPA in monocytes: (i) EPA shortened at
394 the genomic locations of forward-reverse or reverse-forward orientation of DNA motif
395 sequences of transcription factors, (ii) EPA shortened at the genomic locations of any
396 orientation (i.e. without considering orientation) of DNA motif sequences of
397 transcription factors, and (iii) EPA without being shortened by DNA motif sequences.
398 EPI predicted using the three types of EPA in common were removed. First, EPI
399 predicted based on EPA (i) were compared with chromatin interactions (Hi-C). The
400 resolution of chromatin interaction data used in this study was 50kb, so EPI were
401 adjusted to 50kb, before their comparison. The number and ratio of EPI overlapped with
402 chromatin interactions were counted. Second, EPI were predicted based on EPA (ii), and
403 EPI predicted based on EPA (i) were removed from the EPI. The number and ratio of
404 EPI overlapped with chromatin interactions were counted. Third, EPI were predicted
405 based on EPA (iii), and EPI predicted based on EPA (i) and (ii) were removed from the
406 EPI. The number and ratio of EPI overlapped with chromatin interactions were counted.

407 The number and ratio of the EPI compared two times between EPA (i) and (iii), and
408 EPA (i) and (ii) (binomial distribution, p-value < 0.025 for each test).
409
410

411 **Acknowledgements**

412 The supercomputing resource was provided by Human Genome Center of the Institute
413 of Medical Science at the University of Tokyo. Computations were partially performed
414 on the NIG supercomputer at ROIS National Institute of Genetics. Publication charges
415 for this article were funded by JSPS KAKENHI Grant Number 16K00387. This
416 research was partially supported by the Platform Project for Supporting in Drug
417 Discovery and Life Science Research（Platform for Dynamic Approaches to Living
418 System）from Japan Agency for Medical Research and Development (AMED). This
419 research was partially supported by Development of Fundamental Technologies for
420 Diagnosis and Therapy Based upon Epigenome Analysis from Japan Agency for
421 Medical Research and Development (AMED). This work was partially supported by
422 JST CREST Grant Number JPMJCR15G1, Japan.
423
424

425 **Figures**

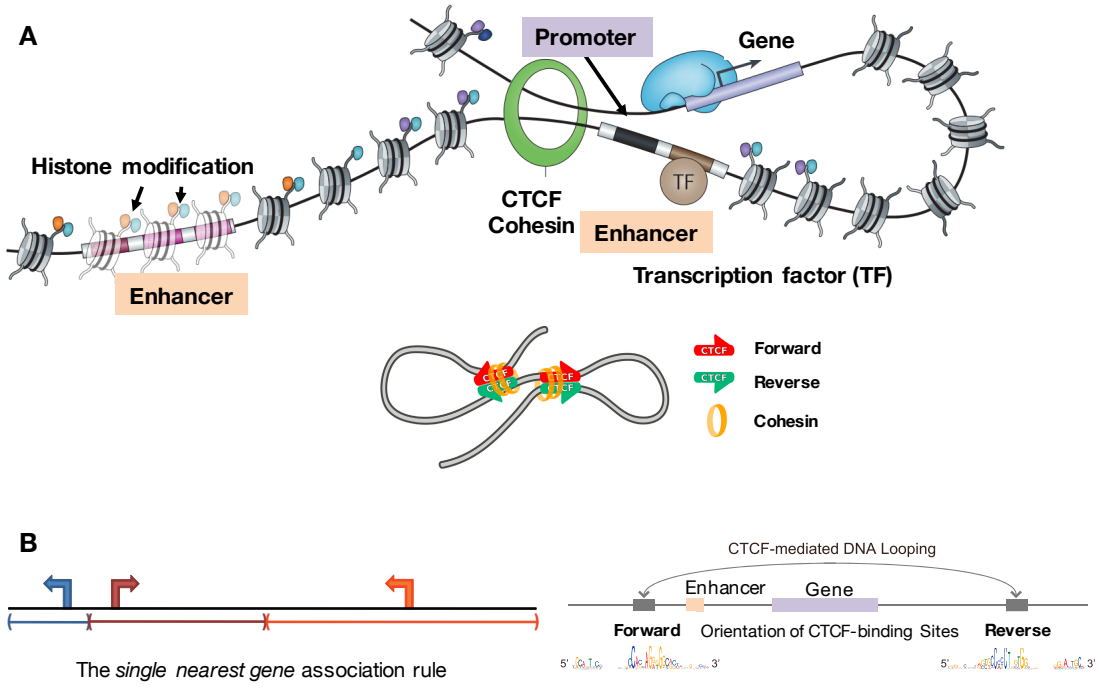

Figure 1. Chromatin interactions and enhancer-promoter association. (A) Forward–reverse orientation of CTCF-binding sites is frequently found in chromatin interaction anchors. CTCF can block the interaction between enhancers and promoters limiting the activity of enhancers to certain functional domains (de Wit et al. 2015; Guo et al. 2015). (B) Computationally-defined regulatory domains for enhancer-promoter association (McLean et al. 2010). The *single nearest gene* association rule extends the regulatory domain to the midpoint between this gene's TSS and the nearest gene's TSS both upstream and downstream. Enhancer-promoter association was shortened at the genomic locations of forward-reverse orientation of DNA binding motif sequences of transcription factors (e.g. CTCF in the figure).

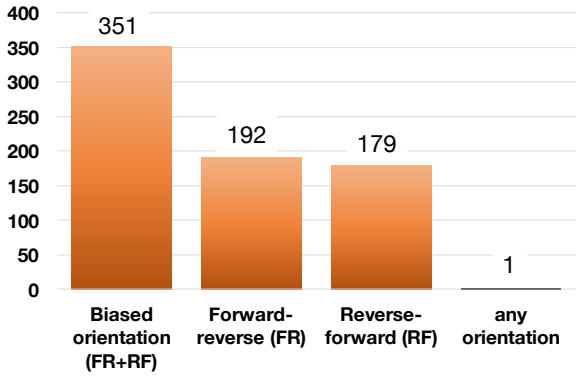

Figure 2. Biased-orientation of DNA motif sequences of transcription factors. Total 351 of biased orientation of DNA binding motif sequences of transcription factors have been found to affect the expression level of putative transcriptional target genes in monocytes of four people in common, whereas only one any orientation (i.e. without considering orientation) of DNA binding motif sequences have been found to affect the expression level.

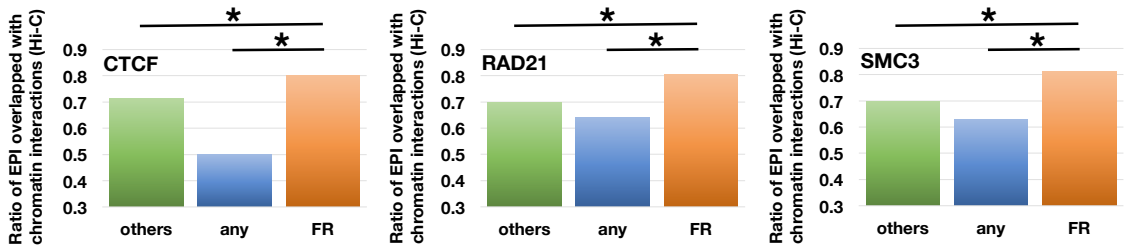

Figure 3. Comparison of DNA binding motif sequences of transcription factors with chromatin interactions. Total 62 biased orientation (41 FR and 24 RF) of DNA motif sequences including CTCF and cohesin showed a higher ratio of EPI overlapped with Hi-C than the other types of EPA in monocytes. FR: forward-reverse orientation of DNA motifs. any: any orientation of DNA motifs. others: enhancer-promoter association not shortened at the genomic locations of DNA motifs. * p-value < 10^{-5}.

Tables

457 **Table 1.** Top 90 biased orientation of DNA binding motif sequences of transcription

458 factors in monocytes. TF: DNA binding motif sequences of transcription factors. Score:

459 $-\log_{10}$ (*p*-value).

460 Forward-reverse orientation

TF	Score	TF	Score	TF	Score
SRF	75.75	YLL054C	34.01	NR2F2	25.16
IRF7	71.35	ZIC1	33.85	PLAGL1	24.95
MINI20	61.98	SMC3	33.67	REST	24.83
CTCF	61.90	ESR2	32.37	GAL4P	24.48
E2F	59.63	TP63	32.28	ZNF143	23.97
STAT1	59.24	ZNF836	31.96	HLH2_HLH3	23.85
ZNF317	56.27	STA5A	31.54	BM9	23.63
CEBPG	55.76	MYB	31.23	YB1	23.29
ZNF623	55.01	AT3G16280	31.22	PRDM15	22.93
ZNF93	54.53	E2F8	30.92	ZNF92	22.52
HMBOX1	53.29	P53	30.86	CPRF1	22.42
SMAD2_SMAD3_SMAD4	50.96	STA5B	30.24	ZNF214	22.41
FOXN2	50.18	STAT3	30.15	PTF1A	22.39
CLAMP	49.82	AP1	30.05	SMRC2	22.16
RDR1	46.44	AT3G04100	29.77	ZNF695	22.08
ZNF646	44.25	CEBPZ	29.73	ZNF28	22.03
TFE3	42.76	RAD21	28.98	ZNF316	21.96
CMYB	42.28	GATA2	28.82	TATA	21.90
KR	41.45	SMAD2	28.81	NF1B	21.75
STAT4	40.84	KLF6	28.08	PHX2B	21.12
ZNF716	40.69	HXB1	27.93	ATHB1	20.96
KLF14	39.94	GLI1	27.39	DEC	20.92
DREB2E	39.93	GCMA	27.03	ZNF709	20.88
FOXA1	38.43	ZNF676	26.96	YY1	20.87
USF	38.04	E47	26.62	EIF5A2	20.85
HME1	37.31	EHF	26.62	ZN148	20.75
ZN219	35.67	TCF12	26.59	PAZ5	20.35
D	35.16	RSRFC4	26.53	PO6F1	20.24
IKZF1	34.09	HIC1	25.64	FOXO4	19.97
YBR033W	34.01	EGR2	25.23	EBF	19.56

461

462

463 Reverse-forward orientation

TF	Score	TF	Score	TF	Score
RFX5	102.60	GABP1_GABP2	36.27	WT1	29.37
TCF1	75.65	RDS2	35.16	E74A	29.16
SRF	65.83	TCFAP2B	34.85	CREM	28.95
ZNF28	64.19	TBP	34.77	ZNF316	28.64
TCF2	62.76	ZNF670	34.63	HAND1_E47	28.37
ZNF225	62.48	ZBRK1	34.60	ZNF449	28.26
CEBPB	59.99	ZNF687	34.12	ZNF233	27.97
NFY	59.26	NFKB2	33.35	SOX2	26.77
ZNF286B	55.03	RFX3	33.04	RXRA	26.37
CTCF	53.46	NRF3	32.99	MYF6	26.15
PAX5	52.87	VND	32.75	MAFK	26.12
STAT1	52.41	RNF96	32.57	CEBPZ	26.03
ISL1	50.03	SP3	32.52	PO2F2	25.89
ETV7	47.38	FLI1	32.45	MADS10	25.62
KLF16	46.75	RBAK	32.22	CETS2	25.35
SP1	46.33	NR2F1	32.10	HSF1	25.06
SMAD3	45.57	NKX3	31.85	GCN4P	24.53
STAT5B	45.02	ZIC2	31.68	CRZ1	24.46
TEAD1	43.30	PARP	31.49	AHR	23.86
ZFP579	43.25	ZNF195	31.36	CRZ1P	23.80
ZBTB2	42.58	BDP1	31.14	PAX6	23.78
ZNF343	41.31	SOX9	31.03	TOE3	23.67
DOBOX5	39.35	SP8	30.79	IRF4	22.77
NANOG	39.14	EMBP1B	30.62	RAP1	22.54
LEU3	38.92	ZNF121	30.34	ETV3	22.20
NRF2	38.83	ZFP202	30.31	FOXP2	21.97
OBOX2	38.30	CD59	30.15	TATA	21.93
ZNF652	38.14	RXRB	29.87	CDX2	21.75
ZNF30	37.89	STAT5A	29.61	NFAC1	21.54
MYOG	36.64	SOX11	29.39	SP4	21.44

Table 2. Top 30 forward-reverse orientation of DNA binding motif sequences of transcription factors in MCF-7 and iPS. TF: DNA binding motif sequences of transcription factors. Score: $-\log_{10}$ (p-value).

MCF-7 iPS

TF	Score
CTCF	23.81
RAD21	16.51
SMC3	15.93
E2F	10.41
CAC	9.82
HOXA7	8.59
STA5B	7.60
REST	7.55
SIRT6	7.35
KLF3	6.82
YY1	5.59
BM9	5.44
DREB2E	5.29
PLAL1	5.27
ZN219	5.19
STAT1	4.73
SNT2	4.71
ETS	3.46
ZBTB6	3.42
PACC	3.39
XRCC4	3.23
DBP	2.83
SETB1	2.82
RREB1	2.80
E2F5	2.68
CLAMP	2.52
ZNF143	2.50
BTEB2	2.47
YKL222C	2.40
AP1	2.28

TF	Score
ZIC3	19.62
ZNF521	19.24
TFE3	16.93
ZNF317	16.74
NKX25	16.54
HES1	15.64
ZBTB24	14.51
AP1	13.99
SMC3	13.83
RAD21	13.24
YLL054C	12.80
ZNF836	11.89
TBX5	11.72
CTCF	11.54
STAT1	11.49
RUNX2	10.58
WT1	10.57
ZIC1	9.91
MAX	8.85
KLF3	8.60
KLF14	8.29
ZNF143	8.11
E2F5	7.96
SIRT6	7.86
HME1	7.79
YY1	7.31
STA5B	7.03
MAZ	6.69
ZN148	6.50
E47	6.03

Table 3. Biased orientation of DNA repeat sequences in monocytes. Score: $-\log_{10}$ (*p*-value).

DNA repeat seq.	Score
AluSg	7.52
L1ME4b	8.83
LTR16C	2.94

Table 4. Top 30 of co-locations of biased orientation of DNA binding motif sequences of transcription factors in monocytes. Co-locations of DNA motif sequences of CTCF with other biased orientation of DNA motif sequences were shown in the separate table. Motif 1,2: DNA binding motif sequences of transcription factors. # overlap: the number

483 of open chromatin regions including both Motif 1 and Motif 2. # motif 1: the number of
484 open chromatin regions including Motif 1. # motif 2: the number of open chromatin
485 regions including Motif 2. # others: the number of open chromatin regions not including
486 Motif 1 or Motif 2. Open chromatin regions overlapped with histone modification
487 (H3K27ac) were used (Total 26,095 regions).

Motif 1	Motif 2	# both	# motif 1	# motif 2	# others	*p*-value
CTCF	SMAD2	4527	781	4478	16309	0
CTCF	RREB1	4379	929	4757	16030	0
CTCF	ZN263	4357	951	4875	15912	0
CTCF	SP1	4321	987	4052	16735	0
CTCF	RAD21	3804	1504	225	20562	0
CTCF	KLF6	3695	852	3800	17748	0
CTCF	SMC3	1259	178	1977	22681	0
CTCF	RXRA	1220	259	2559	22057	0

488

Motif 1	Motif 2	# both	# motif 1	# motif 2	# others	*p*-value
SMAD2	ZN263	7427	1578	1805	15285	0
SMRC2	ZN263	7354	1192	1893	15656	0
RREB1	SP1	7208	1957	1165	15765	0
SIRT6	ZN263	7126	1154	2121	15694	0
MAZ	SP1	7052	909	1321	16813	0
MAZ	ZN263	7022	999	2225	15849	0
SP1	ZN263	7009	1364	2223	15499	0
SMAD2	SP1	6925	2080	1448	15642	0
MAZ	RREB1	6920	967	2245	15963	0
SIRT6	SMAD2	6766	1514	2239	15576	0
SIRT6	SMRC2	6668	1612	1878	15937	0
PLAL1	SMAD2	6637	1074	2368	16016	0
MAZ	SMAD2	6627	1334	2378	15756	0
RARG	SMAD2	6602	1012	2403	16078	0
PLAL1	ZN263	6571	1140	2661	15723	0
EGR2	RREB1	6565	787	2600	16143	0
KLF6	RREB1	6563	932	2573	16027	0
MAZ	SMRC2	6562	1459	1984	16090	0
RREB1	ZN263	6551	1607	2681	15256	0
KLF6	SP1	6541	954	1832	16768	0
PLAL1	RREB1	6496	1215	2640	15744	0
EGR2	SP1	6492	860	1881	16862	0
KLF6	SMAD2	6491	1004	2514	16086	0
RARG	ZN263	6390	1224	2842	15639	0
PLAL1	SP1	6318	1393	2055	16329	0
KLF6	ZN263	6304	1191	2928	15672	0
EGR2	SMAD2	6270	1082	2735	16008	0
EGR2	ZN263	6253	1099	2979	15764	0
EGR2	MAZ	6219	1133	1742	17001	0
RARG	RREB1	6219	1395	2917	15564	0

489

490

491

492 **Table 5.** Comparison of DNA binding motif sequences of transcription factors with
493 chromatin interaction data. TF: DNA binding motif sequences of transcription factors.

494 Score: $-\log_{10}$ (p-value).

495 Forward-reverse

TF	Score 1	Score 2
AP1	28.64	3.75
AT3G04100	24.51	7.07
BTEB2	8.64	2.69
CTCF	6.09	6.80
E2F	19.43	3.24
E2F5	10.58	2.02
EGR2	4.28	10.11
ETS	7.66	5.74
FOSB	7.98	1.99
FOXO4	7.98	5.84
GCM1	2.51	3.47
GCR	6.39	4.50
IRF7	16.36	2.03
KLF14	18.52	8.08
KLF3	6.42	6.43
MAZ	1.82	4.55
MIG2	42.64	6.22
PARG	5.38	4.21
PITX3	10.31	5.62
PLAL1	11.50	5.42
PO6F1	31.28	7.60
PTF1A	8.91	4.50
PU1	3.86	1.71
RAD21	9.28	5.16
RREB1	7.47	6.40
RUNX2	2.97	1.68
SETB1	1.63	2.08
SMAD2	4.10	3.48
SMC3	10.24	6.21
SP1	20.98	4.64
TCFAP2A	5.10	1.63
TF3B	3.21	5.75
TF3CB	20.26	4.66
TFE3	94.90	3.64
YY1	13.88	7.04
ZBTB6	15.08	3.93
ZIC1	5.37	1.63
ZN148	14.51	4.47
ZN219	15.44	7.99
ZNF28	11.88	3.36
ZNF460	14.14	1.63

Reverse-forward orientation

TF	Score 1	Score 2
ASCL2	6.23	3.80
ATF3	17.21	2.23
BC11A	1.65	4.07
BDP1	23.49	6.85
CTCF	12.32	11.74
E2F2	23.33	1.95
EBF1	14.83	7.11
GABP1_GABP2	6.43	2.80
IRF3	2.38	1.98
IRF8	4.30	4.62
NRF3	17.95	1.78
PAX5	48.74	2.74
POLR3A	41.85	9.98
RAD21	1.79	2.26
RNF96	7.69	1.93
RPN4	21.32	4.48
RXRA	29.89	11.87
SMAD4	41.27	2.44
SP2	2.09	11.15
SP4	12.74	1.84
SP8	20.03	2.17
TCFAP2A	8.90	3.53
TCFAP2E	5.72	1.84
ZFP202	23.15	4.85

509

510

511 **References**

512

513 Bailey SD, Zhang X, Desai K, Aid M, Corradin O, Cowper-Sal Lari R, Akhtar-Zaidi B,
514 Scacheri PC, Haibe-Kains B, Lupien M. 2015. ZNF143 provides sequence
515 specificity to secure chromatin interactions at gene promoters. *Nat Commun* **2**:
516 6186.

517 Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble
518 WS. 2009. MEME SUITE: tools for motif discovery and searching. *Nucleic*
519 *acids research* **37**: W202-208.
520 Barutcu AR, Lajoie BR, Fritz AJ, McCord RP, Nickerson JA, van Wijnen AJ, Lian JB,
521 Stein JL, Dekker J, Stein GS et al. 2016. SMARCA4 regulates gene expression
522 and higher-order chromatin structure in proliferating mammary epithelial cells.
523 *Genome research* **26**: 1188-1201.
524 de Wit E, Vos ES, Holwerda SJ, Valdes-Quezada C, Verstegen MJ, Teunissen H,
525 Splinter E, Wijchers PJ, Krijger PH, de Laat W. 2015. CTCF Binding Polarity
526 Determines Chromatin Looping. *Molecular cell* **60**: 676-684.
527 Grant CE, Bailey TL, Noble WS. 2011. FIMO: scanning for occurrences of a given
528 motif. *Bioinformatics (Oxford, England)* **27**: 1017-1018.
529 Guo Y, Xu Q, Canzio D, Shou J, Li J, Gorkin DU, Jung I, Wu H, Zhai Y, Tang Y et al.
530 2015. CRISPR Inversion of CTCF Sites Alters Genome Topology and
531 Enhancer/Promoter Function. *Cell* **162**: 900-910.
532 Javierre BM, Burren OS, Wilder SP, Kreuzhuber R, Hill SM, Sewitz S, Cairns J,
533 Wingett SW, Varnai C, Thiecke MJ et al. 2016. Lineage-Specific Genome
534 Architecture Links Enhancers and Non-coding Disease Variants to Target Gene
535 Promoters. *Cell* **167**: 1369-1384 e1319.
536 Jeong M, Huang X, Zhang X, Su J, Shamim M, Bochkov I, Reyes J, Jung H, Heikamp
537 E, Presser Aiden A et al. 2017. A Cell Type-Specific Class of Chromatin Loops
538 Anchored at Large DNA Methylation Nadirs. *bioRxiv*.
539 Ji X, Dadon DB, Abraham BJ, Lee TI, Jaenisch R, Bradner JE, Young RA. 2015.
540 Chromatin proteomic profiling reveals novel proteins associated with
541 histone-marked genomic regions. *Proc Natl Acad Sci U S A* **112**: 3841-3846.
542 Jolma A, Yan J, Whitington T, Toivonen J, Nitta KR, Rastas P, Morgunova E, Enge M,
543 Taipale M, Wei G et al. 2013. DNA-binding specificities of human transcription
544 factors. *Cell* **152**: 327-339.
545 Karolchik D, Barber GP, Casper J, Clawson H, Cline MS, Diekhans M, Dreszer TR,
546 Fujita PA, Guruvadoo L, Haeussler M et al. 2014. The UCSC Genome Browser
547 database: 2014 update. *Nucleic acids research* **42**: D764-770.
548 Kheradpour P, Kellis M. 2014. Systematic discovery and characterization of regulatory
549 motifs in ENCODE TF binding experiments. *Nucleic acids research* **42**:

550 2976-2987.
551 McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM,
552 Bejerano G. 2010. GREAT improves functional interpretation of cis-regulatory
553 regions. *Nature biotechnology* **28**: 495-501.
554 Mumbach MR, Satpathy AT, Boyle EA, Dai C, Gowen BG, Cho SW, Nguyen ML,
555 Rubin AJ, Granja JM, Kazane KR et al. 2017. Enhancer connectome in primary
556 human cells identifies target genes of disease-associated DNA elements. *Nature*
557 *genetics* **49**: 1602-1612.
558 Newburger DE, Bulyk ML. 2009. UniPROBE: an online database of protein binding
559 microarray data on protein-DNA interactions. *Nucleic acids research* **37**:
560 D77-82.
561 Osato N. 2018. Characteristics of functional enrichment and gene expression level of
562 human putative transcriptional target genes. *BMC Genomics* **19**: 957.
563 Phanstiel DH, Van Bortle K, Spacek D, Hess GT, Shamim MS, Machol I, Love MI,
564 Aiden EL, Bassik MC, Snyder MP. 2017. Static and Dynamic DNA Loops form
565 AP-1-Bound Activation Hubs during Macrophage Development. *Molecular cell*
566 **67**: 1037-1048.e1036.
567 Portales-Casamar E, Thongjuea S, Kwon AT, Arenillas D, Zhao X, Valen E, Yusuf D,
568 Lenhard B, Wasserman WW, Sandelin A. 2010. JASPAR 2010: the greatly
569 expanded open-access database of transcription factor binding profiles. *Nucleic*
570 *acids research* **38**: D105-110.
571 Ramani V, Cusanovich DA, Hause RJ, Ma W, Qiu R, Deng X, Blau CA, Disteche CM,
572 Noble WS, Shendure J et al. 2016. Mapping 3D genome architecture through in
573 situ DNase Hi-C. *Nat Protoc* **11**: 2104-2121.
574 Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn
575 AL, Machol I, Omer AD, Lander ES et al. 2014. A 3D map of the human
576 genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**:
577 1665-1680.
578 Saeed S, Quintin J, Kerstens HH, Rao NA, Aghajanirefah A, Matarese F, Cheng SC,
579 Ratter J, Berentsen K, van der Ent MA et al. 2014. Epigenetic programming of
580 monocyte-to-macrophage differentiation and trained innate immunity. *Science*
581 *(New York, NY)* **345**: 1251086.
582 Schreiber J, Libbrecht M, Bilmes J, Noble W. 2017. Nucleotide sequence and DNaseI

583 sensitivity are predictive of 3D chromatin architecture. *bioRxiv*.
584 Wang L, Wang S, Li W. 2012. RSeQC: quality control of RNA-seq experiments.
585 *Bioinformatics (Oxford, England)* **28**: 2184-2185.
586 Weintraub AS, Li CH, Zamudio AV, Sigova AA, Hannett NM, Day DS, Abraham BJ,
587 Cohen MA, Nabet B, Buckley DL et al. 2017. YY1 Is a Structural Regulator of
588 Enhancer-Promoter Loops. *Cell* **171**: 1573-1588 e1528.
589 Wingender E, Dietze P, Karas H, Knuppel R. 1996. TRANSFAC: a database on
590 transcription factors and their DNA binding sites. *Nucleic acids research* **24**:
591 238-241.
592 Xie Z, Hu S, Blackshaw S, Zhu H, Qian J. 2010. hPDI: a database of experimental
593 human protein-DNA interactions. *Bioinformatics (Oxford, England)* **26**:
594 287-289.
595 Zhang H, Li F, Jia Y, Xu B, Zhang Y, Li X, Zhang Z. 2017. Characteristic arrangement
596 of nucleosomes is predictive of chromatin interactions at kilobase resolution.
597 *Nucleic acids research* **45**: 12739-12751.
598 Zhang Y, An L, Xu J, Zhang B, Zheng WJ, Hu M, Tang J, Yue F. 2018. Enhancing Hi-C
599 data resolution with deep convolutional neural network HiCPlus. *Nat Commun*
600 **9**: 750.
601 Zhao Y, Stormo GD. 2011. Quantitative analysis demonstrates most transcription factors
602 require only simple models of specificity. *Nature biotechnology* **29**: 480-483.
603