**Title**

# Discovery of regulatory motifs associated with human insulator sites reveals the role of directional bias in insulator function

Naoki Osato[1*] and Michiaki Hamada [2,3,4]

[1] Waseda Research Institute for Science and Engineering, Waseda University, Tokyo 169-0072, Japan.

[2] Graduate School of Advanced Science and Engineering, Waseda University, Tokyo, Japan.

[3] Computational Bio Big-Data Open Innovation Laboratory (CBBD-OIL), National Institute of Advanced Industrial Science and Technology (AIST), Tokyo, Japan.

[4] Graduate School of Medicine, Nippon Medical School, Tokyo, Japan.

[*] Corresponding author.

Tel/Fax: +81-3-5286-3130

Email: naokiosato11@gmail.com, nosato@aoni.waseda.jp

## Abstract

Insulator proteins act as a barrier of enhancer–promoter interactions (EPIs). CTCF, a DNA-binding protein (DBP) that binds to chromatin interaction anchors, i.e., edges, is the main insulator protein in vertebrates. However, other DBPs involved in insulator function of EPIs are still unclear. Hence, we developed a systematic, comprehensive deep learning-based approach for identifying the DNA motifs of DBPs associated with insulator. We discovered 99 directional and minor nondirectional motifs in human fibroblast cells that corresponded to 23 DBPs related to insulator function, CTCF, and/or other types of chromosomal transcriptional regulation reported in previous studies. The estimated CTCF orientation bias was consistently proportional to CTCF orientation bias in chromatin interaction anchors. Furthermore, we observed that the motifs could regulate transcribed regions differentially repressed in alternative transcripts. These findings reveal that the directional bias of motifs is associated with insulator functions, chromosome conformation and alternative transcription, potentially mediated by structural interactions. These findings contribute to elucidate novel transcriptional regulations.

## Introduction

Chromatin interactions play an essential role in enhancer–promoter interactions (EPIs) and transcriptional regulation [1]. In this context, transcription factors (TFs) such

39 as CTCF and other DNA-binding proteins (DBPs) such as members of the cohesin
40 complex (RAD21 and SMC3) bind chromatin interaction anchors and facilitate the
41 formation of loop structures. In general, CTCF is considered to function as an insulator
42 by limiting the activity of enhancers within the loops[2]. CTCF indicates a characteristic
43 feature in its DNA-binding sites (DBSs), which show directional bias in favor of a
44 forward–reverse (FR) orientation (rather than a FF, RR, or RF orientation) at chromatin
45 interaction anchors[3]. The forward and reverse orientations of DNA-binding sites (DBSs)
46 are complementary in sequence and are relatively distinct in orientation according to
47 reference genome sequences. Moreover, although the FR and RF orientations of DBSs
48 are complementary, their frequencies are significantly different at chromatin interaction
49 anchors [3], suggesting that the directional bias of DBSs are associated with chromatin
50 interactions and insulator function.

51 While CTCF is the major insulator protein in vertebrates [4], regional micro-capture
52 (Micro-C) analyses revealed that microcompartment interactions are largely
53 independent of cohesin and transcription [5]. This finding suggests that other DBPs may
54 function as insulators or be associated with insulators. However, the experimental
55 identification of insulator-associated DBPs among many other types of DBPs through *in*
56 *situ* experiments is a laborious undertaking.

57 One possible approach for predicting insulator-associated DBPs is to analyze the
58 enrichment of DBSs of chromatin- and insulator-associated DBPs at chromatin

59 interaction anchors, considering that such DBSs may be enriched at these anchors. For
60 example, DBPs such as ZNF143 and YY1 are associated with chromatin interactions
61 and EPIs [1,6]. However, the binding site of ZNF143 is enriched at chromatin interaction
62 anchors, whereas the binding site of YY1 is not enriched in these anchors [6]. CTCF,
63 ZNF143, and YY1 participate in various biological functions in addition to chromatin
64 interactions. This multifunctionality of DBPs likely reduces the enrichment of motifs at
65 chromatin interaction anchors, making them more difficult to identify.

66 Another possible approach for predicting insulator-associated DBPs is to focus on
67 the directional bias of CTCF DBSs. When a DBP binds a single position in the genome,
68 the direction of its DBS may not be important [7], but when a DBP or a complex of DBPs
69 binds more than one genomic position through chromatin interactions, the binding of
70 the DBP or complex may be structurally arranged, contributing to the directional bias of
71 DBSs [8-10]. Some DBPs involved in chromatin interactions, such as YY1, are
72 multifunctional proteins, and some DBSs of multifunctional DBPs are involved in
73 chromatin interactions and insulator functions. Thus, the directional difference in the
74 DBSs of a multifunctional DBP may be reduced. The multifunctionality of DBPs may
75 make it difficult to identify their insulator-associated functions.

76 Computational methods have been developed to predict chromatin interactions
77 and associated DBPs [11] [12]. However, these methods were not designed to identify the
78 binding motifs of DBPs associated with insulator function. As for regulatory motifs, Xie

79 et al. [13] suggested that the development of systematic methods for identifying the
80 specific functions of regulatory motifs in a genome-wide manner is the next challenge
81 that must be overcome. As for the regulation of gene expression, insulators affect EPIs
82 and gene expression, and thus the use of gene expression data would improve the
83 prediction of DBPs associated with insulator function, compared with the enrichment
84 analysis or directional bias of DBSs alone. A novel computational method using deep
85 learning has been developed to predict gene expression levels in human tissues based on
86 the DBSs in promoter regions [14]. However, the method does not use the DBSs
87 associated with EPIs as input data, and was not designed to predict DBPs associated
88 with insulator function.

89 Hence, we aimed to search for insulator-associated DBPs using a computational
90 approach integrating and analyzing genome-wide experimental data. To predict
91 insulator-associated DBPs in the present study, we revised the previously developed
92 deep learning method with the information of EPIs based on eQTL data and the
93 enhancer–promoter association (EPA) rule (McLean et al. Supplementary Fig. 2) [15], and
94 predicted the directional bias of DBSs of CTCF and insulator-associated DBPs,
95 including those of known DBPs such as CTCF, RAD21, and SMC3. We show that 23 of
96 the predicted insulator-associated DBPs are associated with insulator function, CTCF,
97 and/or other types of transcriptional regulation affecting EPIs that have been identified
98 in previous studies. We reveal that directional DBSs are involved in insulator functions,

99 chromatin interactions and chromosome conformation. To validate these findings, we
100 demonstrate that the DBSs of the predicted insulator-associated DBPs are located within
101 potential insulator sites. Our method and analyses contribute to the identification of
102 insulator- and chromatin-associated DBSs that influence EPIs and reveal novel
103 structural and functional roles of DBPs and gene regulation.
104

105 ## Results

106 **Machine learning of EPIs to predict gene expression levels**

107 Insulator proteins act as a barrier of EPIs (Fig. 1a, b, c). To predict the EPIs
108 between DBPs that bind DBSs distal to transcription start sites (TSSs) and the
109 transcription initiation complex at TSSs, we identified the interactions between the
110 DBSs in putative enhancer regions and genes based on expression quantitative trait
111 locus (eQTL) data showing the association between a SNP (single nucleotide
112 polymorphism) and a gene expression level (Fig. 2a, b). To avoid the cellular
113 heterogeneity found in tissues, gene expression levels were analyzed in human cell lines
114 (human foreskin fibroblasts (HFFs), monocytes, human mammary epithelial cells
115 (HMECs), and neural progenitor cells). The predicted gene expression levels were
116 highly consistent with the actual expression levels (Spearman's rho = 0.71, 0.74, 0.73,
117 and 0.74 in the four cell types) without using the data on the RNA-binding proteins and
118 miRNAs obtained via the original deep learning method (Fig. 3a; Fig. S1a) [14].

119 Then, to investigate the directional bias of the DBSs of putative
120 insulator-associated DBPs, we searched the ChIP-seq peaks of candidate DBPs for the
121 associated binding motifs. First, the DBSs of a given DBP in open chromatin regions
122 were predicted using the Protein Interaction Quantitation (PIQ) tool and DNase-seq data
123 [16]. Then, to reduce the number of false-positive DBS predictions, the DBSs that
124 overlapped with the ChIP-seq peaks of the same DBP were selected. Among the 1,356
125 DBPs with available ChIP-seq data, the DBP-binding motif sequences of 715 DBPs
126 were obtained from databases and publications, specifically, TRANSFAC (2019.2),
127 JASPAR (2018), UniPROBE (2018), DNA motifs from high-throughput SELEX, DNA
128 motifs from ENCODE DBP ChIP-seq data, and HOCOMOCO versions 9 and 11 [17-22].
129 To confirm that the DBSs of insulator DBPs are located at the chromatin
130 interaction anchors of those DBPs, we examined the overlap between CTCF ChIP-seq
131 peaks and anchors identified to be involved in CTCF-mediated chromatin interactions in
132 HFF cells *in situ* by paired-end-tag sequencing (ChIA-pet). The number of CTCF
133 ChIP-seq peaks included in the Gene Transcription Regulation Database (GTRD)
134 (v20.06 release) was 2,433,551, and the number of chromatin interactions was
135 123,396,983. When the length of a chromatin interaction anchor was defined as 100 bp,
136 which is close to the length of 147 base pairs (bp) of DNA around a histone, 2,263,295
137 (93%) CTCF ChIP-seq peaks overlapped with the anchors of CTCF-mediated
138 chromatin interactions in HFF cells. These findings indicated that most of the DBSs of

139 CTCF were potentially involved in chromatin interactions. This suggested that the
140 CTCF DBSs predicted based on the experimental data of CTCF-binding sites, open
141 chromatin regions, and the DNA-binding motif of CTCF included the DBSs associated
142 with chromatin interactions and insulator functions.

143

144 **Comparison of the effect of insulator-associated DBPs on gene expression levels**

145 To investigate the effect of directional differences in the DBSs of putative
146 insulator-associated DBPs on EPIs and gene expression levels, we computed DeepLIFT
147 scores for use as inputs for deep learning analyses. The DeepLIFT method calculates a
148 score representing the contribution of the input data to the prediction. By comparing the
149 DeepLIFT scores of the DBSs to the ChIP-seq data, we compared the effects of putative
150 insulator sites, including insulator-associated DBSs, on the prediction of gene
151 expression levels according to the directional and positional differences among the
152 DBSs (Fig. 2d). At the anchors of distal CTCF-mediated chromatin interactions
153 identified *in situ* via ChIA-pet, 76.4% of the CTCF DBSs showed an FR orientation,
154 also known as a convergent orientation (Fig. 3b) [3].

155 To investigate the effect of CTCF DBS orientation on the DeepLIFT scores of the
156 DBSs in putative enhancer regions, we searched CTCF DBSs with an FR or RF
157 orientation in the upstream and downstream regions of TSSs and predicted EPIs based
158 on the enhancer–promoter association (EPA) rule (Fig. 2c). We then compared the

159 distribution of the DeepLIFT scores and mean DeepLIFT scores between the CTCF
160 DBSs with FR and RF orientations and observed a significant difference in the
161 distribution of DeepLIFT scores (Mann–Whitney *U* test, FDR-adjusted *p* value < 0.05
162 and *p* value < $10^{-3}$; Fig. 3c, d and Table S1). Our method was able to compare the effect
163 of directional difference of CTCF DBS on the prediction of gene expression level.

164 We performed the same analysis using the DBSs of RAD21, SMC3, and other
165 DBPs as putative insulator sites and found that the DBPs including RAD21 and SMC3
166 also showed a significant difference in the distribution of DeepLIFT scores between
167 DBSs with FR and RF orientations (Fig. 3c and Table S1). Other orientations of the
168 DBSs of putative insulator-associated DBPs, such as forward-forward (FF),
169 reverse-reverse (RR), and any direction (i.e., no directional bias), were also evaluated
170 and compared to examine the differences in the distribution of DeepLIFT scores (Table
171 S1). First, we selected 228 DBPs with at least 1,000 DBSs per DBP in HFF cells.
172 Among these, 92 DBPs, including CTCF, RAD21, SMC3, ZNF143, and YY1, showed
173 significant differences in the distribution of DeepLIFT scores between DBSs with
174 different orientations (e.g., FR and RF), and 79 DBPs, including YY1, presented a
175 significant difference in the distribution of DeepLIFT scores between DBSs with the
176 same orientation (e.g., FF and RR) (Fig. 3c; Fig. S1b; Table S1, DBP comp 1A).
177 Unexpectedly, we found that many DBPs, including known insulator- and
178 chromatin-associated DBPs, showed significant differences of DeepLIFT scores, and

179 that the known DBPs had the most significant differences in the distribution of
180 DeepLIFT scores, suggesting that our method worked properly. To reduce the number
181 of DBPs with a directional DBS bias and thus reduce the number of false-positive DBP
182 predictions, we set the *p* value threshold to $< 10^{-3}$. We confirmed that all the DBPs with
183 *p* values $< 10^{-3}$ constituted the subset of DBPs with false discovery rate (FDR)-adjusted
184 *p* values < 0.05 (included in Table S1).
185

186 **Discovery of directional and nondirectional DBSs of insulator-associated DBPs**

187 To investigate the directional bias of DBSs based on more stringent criteria, we
188 compared DBSs with four different orientations (FR, RF, FF, and RR), or no directional
189 bias based on pairwise comparisons of the directional biases of the DBSs. For the DBSs
190 with different orientations (e.g., FR and RF) associated with 23 DBPs, including CTCF,
191 RAD21, and SMC3, the distribution of DeepLIFT scores differed significantly among
192 those with the four different orientations and no directional bias. The distribution of
193 DeepLIFT scores differed significantly among the DBSs with the same orientation (e.g.,
194 FF and RR) for 17 DBPs. The distribution of DeepLIFT scores also differed
195 significantly among the nondirectional DBSs of six DBPs (*p* value $< 10^{-3}$, Table 1 and
196 Table S2, DBP list 1A; FDR-adjusted *p* value < 0.05, Table S3, DBP list 3A). We
197 confirmed that known insulator-associated DBPs such as CTCF, RAD21, and SMC3
198 showed FR orientation of DBSs, and discovered other DBPs with directional and

199 nondirectional bias. We also performed the analysis with other DBS selection criteria,
200 and discovered DBPs with directional and nondirectional bias that were not included in
201 those discovered with the first DBS selection criterion (Fig. 3e). (Data S1).

202 We compared the predicted directional biases of CTCF DBSs with the directional
203 biases identified based on experimental data on CTCF-mediated chromatin interactions.
204 In the analysis of ChIA-pet chromatin interaction data, the anchors of distal
205 CTCF-mediated chromatin interactions located in CTCF DBSs exhibited the following
206 orientations: 76.4% FR, 11.0% FF, 10.3% RR, and 2.3% RF (Fig. 3b; Guo Y et al. *Cell*
207 2015 [3], Fig. 4). Interestingly, the *p* values obtained from the statistical tests of the
208 DeepLIFT scores (Mann–Whitney *U* test) of CTCF, RAD21, and SMC3 DBSs with FR
209 vs. RF orientations were lower than those obtained for the FR vs. RR and FR vs. FF
210 comparisons. These results were consistent with the percent differences in the
211 orientations of CTCF DBSs at anchors of chromatin interactions: the percent difference
212 between the FR (76.4%) and RF (2.3%) orientations was larger than the percent
213 differences for FR (76.4%) vs. FF (11.0%) and RR (10.3%) (Fig. 3c, d). These data
214 indicated that the deep learning analysis accurately predicted the percent differences in
215 the FR, FF, RR, and RF orientations of CTCF DBSs at chromatin interaction anchors on
216 the basis of the DeepLIFT scores of the DBSs in putative enhancer regions.

217 A total of 99 insulator-associated DBPs were predicted based on the stringent
218 criterion of the DBSs of a given DBP showing either a specific directional bias or no

219 directional bias, as determined by comparing the four possible orientations (FR, RF, FF,
220 and RR) and the absence of directional bias (X). We found that 23 DBPs exhibited
221 insulator and CTCF-associated functions and/or chromatin regulatory functions that
222 affected EPIs, as reported in previous studies: 13 DBPs were associated with CTCF
223 (Table S6). The 23 DBPs were multifunctional proteins, and their known functions were
224 transcriptional regulator, repressor, chromatin regulator, and interaction with other
225 proteins including enzymes that facilitate epigenetic modifications. These findings
226 suggested that CTCF regulated EPIs by interacting with other DBPs, and that DBPs
227 with other functions were also associated with the regulation of EPIs and chromatin
228 conformation.
229
230 **Search for insulator-associated DBSs based on histone modifications and gene**
231 **expression**
232 To confirm the predicted insulator-associated DBPs via another method and select
233 their DBSs associated with insulator function, we first analyzed previously reported
234 histone modification data to identify differences between the upstream and downstream
235 regions of the DBSs. CTCF bound to sites located between repressed regions with
236 H3K27me3 marks and active regions with H2AK5ac marks, which showed different
237 signal intensities between the upstream and downstream regions of CTCF DBSs [23]. The
238 transcription factor USF1 exhibits a barrier function, in which histone modification

239 factors are recruited to separate adjacent regions into activating and repressive
240 chromatin domains [24]. We selected H3K27me3 histone modifications, associated with
241 the repression of gene expression, for comparison between the upstream and
242 downstream regions of the DBSs of predicted insulator-associated DBPs using
243 H3K27me3 ChIP-seq data because H3K27me3 histone marks occupy long genomic
244 regions and are thus easily compared between genomic regions. We searched for DBSs
245 where the mean H3K27me3 signal intensities differed by more than 2-fold between the
246 upstream and downstream regions of the DBS. The higher mean signal was found to be
247 greater than the first quartile of all H3K27me3 signal intensities because some
248 H3K27me3 signals exhibited a low intensity and appeared noisy. Thus, the criteria were
249 sufficient to clearly distinguish modification status among genomic regions.

250 We further counted the numbers of DBSs with different H3K27me3 signal
251 intensities between the upstream and downstream regions of a DBS and found that
252 known insulator-associated DBPs, such as CTCF, RAD21, and SMC3, generally
253 showed the greatest numbers of DBSs (Table S4). Via statistical analysis, we calculated
254 the median and mean numbers of DBSs showing different H3K27me3 signal intensities
255 between upstream and downstream regions for the DBSs of both predicted
256 insulator-associated DBPs and other DBPs. The median and mean numbers of DBSs
257 were significantly larger among predicted insulator-associated DBPs than other DBPs
258 (Mann–Whitney $U$ test, $p$ value $< 0.05$) (Table 2, Table S4, Fig. 4a, b). We confirmed

259 that a larger number of the DBSs of known insulator-associated DBPs were located
260 between activated and repressed regions than non-insulator-associated DBPs.

261 To identify additional genomic regions obviously separating transcriptionally
262 activated and repressed regions, we also examined the expression levels of transcripts
263 using the signal intensity indicated by RNA-seq data for identifying transcriptionally
264 active regions based on the available HFF cell data. As in the analysis of H3K27me3
265 signal intensity, we searched for DBSs of predicted insulator-associated DBPs with
266 mean transcript signal intensities differing more than 1.5-fold between upstream and
267 downstream regions, and the higher mean signal was greater than the first quartile of all
268 transcript signal intensities. We again found that a significantly larger number of the
269 DBSs of predicted insulator-associated DBPs were located between activated and
270 repressed regions than other DBPs (Table 2, Table S4, Fig. 4a, c).

271 In the analysis of chromatin interaction data, insulation scores were calculated
272 based on the analysis of the boundary regions of chromatin interactions [25]. We also
273 examined the overlap of DBSs of predicted insulator-associated DBPs with the
274 boundary regions, and again found that a significantly larger number of the DBSs of
275 predicted insulator-associated DBPs were located in the boundary regions than other
276 DBPs (Table 2, Fig. 4d).

277 To determine whether the DBSs of predicted insulator-associated DBPs
278 potentially act as insulators or cofactors of insulators, we investigated the genomic

279 locations of the DBSs on the UCSC Genome Browser (Fig. S2). We searched for DBSs
280 located between the H3K27me3 marks and transcribed regions of genes, which were
281 primarily located at the edges of H3K27me3 marks and were not found near the binding
282 sites of CTCF, RAD21, and SMC3. For each DBP, we looked for insulator sites
283 associated with the DBSs in the genome browser and selected primarily three DBSs for
284 each DBP (Fig. S2). We discovered DBSs in the expected genomic regions of all DBPs,
285 suggesting that these DBSs are associated with insulator functions. Genome browser
286 analysis is useful for selecting DBSs for experimental validation and further study.

287 While preparing Fig. S2, we noticed that several DBSs of insulator-associated
288 DBPs were found in the same genomic locus, close to each other on the genome. These
289 DBPs potentially act together to form an insulator site. In addition, some transcribed
290 regions in the genome were differentially repressed in alternative transcripts and
291 overlapped with a H3K27me3 mark that was separated with the DBS of
292 insulator-associated DBP between the repressed and active regions (Fig. 5, S2a, b).
293 Furthermore, the DBSs of insulator-associated DBPs were frequently located near
294 transcription start and termination sites. Alternative transcription start and termination
295 sites have been reported to explain the majority of tissue-specific exon usage rather than
296 alternative splicing[26]. The DBSs of insulator-associated DBPs could affect the form of
297 transcripts and the genomic location of transcription start and termination sites. The
298 differences in the DBSs and H3K27me3 marks between tissues could contribute to the

299 tissue-specific exon usage of alternative transcripts.

300

301 **Distribution of DBSs around chromatin interaction anchors**

302 In this study, we aimed to predict chromatin- and insulator-associated DBPs. To
303 estimate how many DBPs are involved in chromatin interactions and examine the
304 distribution and association of DBSs around chromatin interaction anchors, the DBSs of
305 DBPs in regions within ±500 bp of the centers of chromatin interaction anchors were
306 identified in HFF cells. Micro-C assays enabled us to identify chromatin interactions at
307 a higher resolution than can be achieved via Hi-C analysis. The distribution of CTCF,
308 RAD21, and SMC3 DBSs showed a characteristic pattern around chromatin interaction
309 anchors, with the highest peaks located around anchors (Fig. 6a, b, c). The peaks of the
310 DBSs of YY1 and ZNF143, which are known to be involved in chromatin interactions,
311 were distributed around chromatin interaction anchors. The DBSs of YY1 were also
312 broadly distributed in regions within ±500 bp of anchors, suggesting that this DBP is
313 involved in various functions in addition to chromatin interactions (Fig. 6d, e). MAZ,
314 which exhibits insulator function independent of CTCF[27], indicated the peak of
315 distribution of MAZ DBSs around anchors (Fig. 6f).

316 Among the 228 DBPs, more than 1,000 of the DBSs in open chromatin regions in
317 HFF cells overlapped with the ChIP-seq peaks of the DBPs, and the peaks of almost all
318 the DBPs (approximately 90%) were located around anchors. Furthermore, we found

319 DBPs with sharp peaks around anchors, DBPs with broad peaks around anchors, and
320 DBPs with peaks similar to those of CTCF, RAD21, and SMC3 around anchors,
321 suggesting that these DBPs may be associated with CTCF, RAD21, and/or SMC3
322 (Table S7; Fig. S3). Among the predicted insulator-associated DBPs, MeCP2 and
323 PRDM9 exhibited fewer than 1,000 DBSs in HFF cells, and when a DBP has fewer than
324 1,000 DBSs, the specific distribution of DBSs around chromatin interaction anchors
325 may not be evident (Fig. S3). Approximately 90% of the DBPs with more than 1,000
326 DBSs had peaks around anchors. This indicated that the DBPs interacted with other
327 genomic regions and that the DBSs of predicted insulator-associated DBPs were
328 involved in chromatin interactions.

329

330 **Co-localization of DBSs around chromatin interaction anchors**

331 Since the DBSs of the predicted insulator-associated DBPs may be located close
332 to those of known insulator-associated DBPs, such as CTCF, RAD21, and/or SMC3,
333 and, thus, be presumed to act as insulators, we next counted the number of DBSs in
334 each 10-bp bin within ±500 bp from the center of a chromatin interaction anchor and
335 compared these numbers between all possible pairs of DBPs. We found that the number
336 of DBSs of pairs of DBPs located within the same chromatin interaction anchor tended
337 to be the highest within the same bin distance from the anchor, suggesting that the
338 resolution of the chromatin interaction data was high enough to determine the distance

339 of DBSs of DBPs from anchors and distinguish two DNA-bound DBPs located close to
340 each other (Table S5). For this analysis, the DBSs that overlapped with the ChIP-seq
341 peaks of a DBP were selected to reduce the number of false-positive DBS predictions,
342 and the number of DBSs per DBP may therefore have been underestimated. Among the
343 228 DBPs, 91 pairs of 70 DBPs colocalized around chromatin interaction anchors when
344 at least 50% of the DBSs of a pair of DBPs colocalized in a bin with the same distance
345 from the anchor (Table S5). CTCF showed a high rate of DBS colocalization with both
346 RAD21 (39%) and SMC3 (81%) in bins within ±500 bp from the center of a chromatin
347 interaction anchor, whereas YY1 and ZNF143 showed low rates of DBS colocalization
348 with RAD21 (1.5% and 5.5%, respectively) and CTCF (2.3% and 4.8%, respectively)
349 (Table 3). For CTCFL (CCCTC-binding factor like), RXRA, PPARG, NEUROD1, and
350 ZBTB7A, more than 10% of DBSs colocalized with CTCF, RAD21, and/or SMC3 in
351 bins within the same distance from chromatin interaction anchors, suggesting that the
352 insulator function of the DBPs was associated with known insulator DBPs. For the other
353 DBPs among the 228 DBPs, less than 10% of DBSs colocalized with CTCF, RAD21,
354 and/or SMC within a bin. Most of the predicted insulator-associated DBPs did not have
355 DBS colocalization and were not in close proximity to known insulator-associated
356 DBPs such as CTCF, RAD21, and SMC3; that is, they were not within a 10-bp bin in
357 the regions ±500 bp from the center of an anchor.

358 To further examine the possibility that known insulator-associated DBPs interact

359 indirectly with the other predicted insulator-associated DBPs, we investigated the
360 overlap of DBSs with the DBP ChIP-seq peaks of CTCF, RAD21, and SMC3 and found
361 that a high percentage (> 98%) of the DBSs of the three DBPs overlapped with their
362 ChIP-seq peaks in all three examined combinations. However, the DBSs of almost all
363 other predicted insulator-associated DBPs overlapped with less than 50% of the
364 ChIP-seq peaks of the three known DBPs, even though the ChIP-seq data were
365 collected from various cell and tissue types. These results suggested that the predicted
366 insulator-associated DBPs play different functional roles than the known DBPs.

367 Although a DBS distribution pattern around chromatin interaction anchors similar
368 to those of CTCF, RAD21, and SMC3 was observed for DBPs such as BACH2, FOS,
369 ATF3, NFE2, and MAFK, less than 1% of the DBSs of the latter set of DBPs
370 colocalized with CTCF, RAD21, or SMC3 in a single bin. However, BACH2, FOS,
371 ATF3, NFE2, and MAFK had DBSs that colocalized with each other (Table 3). These
372 results suggested that the two different groups of colocalized DBPs were independently
373 involved in the same chromatin-related function such as chromatin remodeling [28 29 30].

374

375 **Discussion**

376 Here, we present the first, to our knowledge, computational method and
377 genome-wide analysis aimed at detecting insulator-associated DBPs. Transcriptional
378 regulation mediated by the binding of DBPs at sites located far from TSSs was

379 modelled using our deep learning method based on eQTL data. Through this approach,
380 the effects of the DBSs of insulator-associated DBPs on EPIs and gene expression were
381 estimated, and insulator-associated DBPs were predicted. We further identified DBPs,
382 including known insulator-associated DBPs that exhibit insulator-associated functions
383 and regulate transcription, thereby affecting EPIs reported in previous studies. The
384 DBSs of the predicted insulator-associated DBPs showed orientation bias of their DBSs;
385 unexpectedly, the significant differences in the orientation bias of CTCF DBSs based on
386 DeepLIFT scores (contribution scores) were found to be consistent with the percent
387 differences in the orientation of CTCF DBSs at anchors of CTCF-mediated chromatin
388 interactions. The DBSs of the predicted insulator-associated DBPs were located at
389 potential insulator sites more frequently than those of other DBPs, according to the
390 differences in H3K27me3 signal intensities and transcript levels. We further observed
391 that only a small percentage of the DBSs of the insulator-associated DBPs colocalized
392 with the DBSs of known insulator-associated DBPs such as CTCF, RAD21, and SMC3
393 around chromatin interaction anchors. We discovered that DBPs other than the known
394 insulator DBPs and the directionality of their DBSs are associated with insulator
395 functions. Our novel computational method enabled this study, which will contribute to
396 further promote the functional and structural analyses of insulator-associated DBPs
397 effectively through experimental analyses.

398 Our deep learning analysis accurately predicted the percent differences in the FR,

399 FF, RR, and RF orientations of CTCF DBSs at chromatin interaction anchors,
400 considering the genomic locations of DBSs of putative insulator-associated DBPs. The
401 other predicted insulator-associated DBPs also showed orientation bias of their DBSs.
402 These results suggest that the directional bias of DBSs of insulator-associated DBPs
403 may be involved in insulator function and chromatin regulation through structural
404 interactions among DBPs, other proteins, DNAs, and RNAs. For example, the
405 N-terminal amino acids of CTCF have been shown to interact with RAD21 in
406 chromatin loops [8-10]. For another example, although sequence-based predictions of EPIs
407 underestimated actual interactions, an activity-by-contact (ABC) approach, based on a
408 simple distance model, performed well, suggesting that factors in addition to sequence
409 influence EPIs [31,32]. Furthermore, CTCF-mediated chromatin looping induces phase
410 separation between the inner and outer regions of the loops, resulting in the formation of
411 transcriptional condensates and insulator function [33-35]. Therefore, structures that result
412 in phase separation, rather than functional differences in DBPs, may be responsible for
413 insulator functions. Other types of chromatin regulation are also expected to be related
414 to the structural interactions of molecules. Our analyses will allow us to reveal novel
415 structural roles of DBPs for insulator functions, chromatin regulation, and phase
416 separation and to advance their experimental analysis.

417 High-resolution chromatin interaction data from a Micro-C assay showed that
418 most of the predicted insulator-associated DBPs showed DBS distribution peaks around

419 chromatin interaction anchors, suggesting that these DBPs are involved in chromatin
420 interactions. However, due to the high resolution of a Micro-C assay, it is difficult to
421 obtain all chromatin interaction data at a reasonable and available cost, while a Region
422 Capture Micro-C assay focuses on a specific genomic region to increase the coverage of
423 chromatin interactions. The coverage of chromatin interactions will change according to
424 the experimental data of a Micro-C assay. In addition, chromatin interactions include
425 other types of interactions that are not associated with EPIs and insulator functions [36-38].
426 Therefore, we analyzed and predicted insulator-associated DBPs without using
427 chromatin interaction data. Our deep learning analysis successfully learned EPIs and
428 accurately predicted the percent differences in the FR, FF, RR, and RF orientations of
429 CTCF DBSs at chromatin interaction anchors.

430 The analysis of many contribution scores (i.e., DeepLIFT scores in this study)
431 for directional DBSs of putative insulator-associated DBPs in the prediction of
432 transcript expression levels successfully predicted the percent differences in the FR, FF,
433 RR, and RF orientations of CTCF DBSs at chromatin interaction anchors based on the
434 contribution scores of DBSs in putative enhancer regions. Some contribution scores
435 may not be accurate and thus the prediction of insulator-associated DBPs or DBSs
436 based on a small number of contribution scores may be erroneous. To compare
437 contribution scores computed based on different pairs of input and output data with our
438 deep learning model, we normalized contribution scores for each pair of input and

439 output data as previously reported [39]. The analysis of many contribution scores will be
440 used to estimate the effects of other molecules and functions involved in gene
441 regulation.

442 One of the next computational challenges will be to predict and narrow the
443 potential set of insulator-associated DBSs of multifunctional DBPs. Information on
444 histone modifications such as H3K27me3 and transcript levels will be beneficial for
445 comparing the results of such analyses to confirm the accuracy of a prediction and for
446 identifying insulator-associated DBSs in the genome for further experimental validation
447 and study. As a reference for an application of this study, Liu et al. focused on
448 CTCF-binding sites and performed various assays in human and mouse cell cultures and
449 in mice with the goal of applying these sites in gene therapy [40]. The analysis of
450 insulator-associated DBPs may contribute to the development of technologies to
451 regulate gene expression. In conclusion, our novel computational method discovered
452 insulator-associated DBPs, including known DBPs, and revealed that the directional
453 bias of insulator-associated DBPs is potentially involved in chromatin looping, phase
454 separation, transcriptional condensates and insulator functions through the structural
455 interactions of the DBPs. This will lead to a general and global understanding of
456 transcriptional regulation and gene expression.

457

458 **Methods**

459 **Prediction of gene expression levels based on DBSs**

460 **Gene expression data as output data for machine learning analysis**

461 To predict DBSs that also act as insulator sites in the human genome, we
462 modified a previously developed deep learning method designed to predict the transcript
463 expression level of genes in tissues and cell types based on the DBSs of DBPs in
464 promoter regions and other DNA regions (Fig. 2a; Fig. S4) [14]. To reduce the influence of
465 cellular heterogeneity observed when analyzing tissues, we used the gene expression
466 data from four distinct cell types where various experimental data are available in public
467 databases: HFFs (GSM941744), monocytes (Blueprint RNA-seq FPKM [fragments per
468 kilobase of exon per million reads mapped] data
469 'C0010KB1.transcript_quantification.rsem_grape2_crg.GRCh38.20150622.results' in
470 the Blueprint DCC portal at http://dcc.blueprint-epigenome.eu/#/files), human
471 mammary epithelial cells (HMECs) (Encyclopedia of DNA Elements at UCSC
472 http://genome.ucsc.edu/encode/,
473 'wgEncodeCshlLongRnaSeqHmecCellPapAlnRep1.bam'), and neural progenitor cells
474 (NPCs) (GSM915326, ENCFF529SIO). For gene expression data, FPKM values
475 representing the expression levels of genes were retrieved or calculated based on
476 GENCODE V19 gene annotations in the ENCODE portal database [41]
477 (https://www.encodeproject.org/) using the RSeQC tool [42] (http://rseqc.sourceforge.net/).
478 We calculated log2-fold-transformed FPKM values (adding 0.25 to prevent results of

479 negative infinity), median log2-FPKM values across the four cell types, and log2-fold
480 changes relative to the median for all cell types according to the original deep learning
481 method (transcriptome data processing as described by Tasaki et al.) [14].
482

483 **DBSs as input data for machine learning analysis**

484 To investigate the effect of insulator sites on EPIs and gene expression, the DBSs
485 in putative enhancer regions located away from promoter regions were added as input
486 data in the deep learning method. To identify the DBSs of DBPs in promoter and
487 enhancer regions, ChIP-seq data of the DBPs were used and were updated to GTRD
488 database release v20.06 (https://gtrd.biouml.org) [43]. The binding site of a DBP
489 regulating the expression level of each transcript was selected using eQTL data from
490 GTEx portal release v8 (https://gtexportal.org). If a DBS was located within a 100-bp
491 region around a single-nucleotide polymorphism (SNP) of an eQTL, we assumed that
492 the DBP regulated the expression of the transcript corresponding to the eQTL. If the
493 SNP of an eQTL was not located on the same chromosome as the transcribed region of
494 the eQTL (i.e., *trans*-eQTL), the eQTL data were not used. The genomic positions of
495 the eQTLs were converted to the corresponding positions in the hg19 version of the
496 human genome using the LiftOver tool
497 (http://hgdownload.soe.ucsc.edu/downloads.html#liftover).

498 The original input data of DBSs in promoter regions used for the deep learning

analysis consisted of the lengths of a DBP ChIP-seq peaks in each 100-base bin in a region between 2 kbp upstream and 1 kbp downstream of the transcription start site (TSS) of each gene or transcript [14]. To add the input data of DBSs of DBPs in more upstream and downstream regions of a TSS and reduce the number of bins, we first computed the length of a DBP ChIP-seq peak in a 100-base bin and then selected the maximum length in each interval of 100 bins in a region between 1 megabase upstream and downstream of a TSS. This step outputted the 200 longest DBP ChIP-seq peaks in the examined regions.

**Training step for machine learning analysis**

In the training of the modified deep learning model, the input data for promoters and putative enhancer regions were concatenated into a total of 230 bins of the lengths of 1,311 DBP ChIP-seq peaks, consisting of 30 bins for promoters and 200 bins for putative enhancers, because the DeepLIFT tool 'RevealCancel rule on dense layers and the Rescale rule on conv layers (genomics default)', with high scoring accuracy, [44] cannot be applied to a neural network with multiple input layers (Fig. S4). Ubuntu (version 20.04), Python (version 3.8.10), TensorFlow (version 2.6.0), NVIDIA CUDA toolkit (version 11.4.2), GeForce RTX 3080 GPU with 10 GB memory, and Intel Core i7-10700 CPU with 128 GB memory were used for training and prediction, and the original scripts were modified to run on newer versions of the software. The training

method was the same as the original, and the training time was two to four hours. In this study, we focused on the analysis of HFF cells and assessed the training process using all the input data as the test data, which slightly improved the prediction of gene expression. Therefore, we used the trained model for the analysis of insulator-associated DBPs.

**Prediction of insulator-associated DBPs affecting EPIs**

**DBSs of putative insulator-associated DBPs**

To identify the direction of DBP-binding motifs in the DBSs of predicted insulator-associated DBPs, we searched for DBP-binding motifs in open chromatin regions in HFF cells (GSE170964) using the PIQ tool [16] and the position frequency matrices of the DBP-binding motifs. The motif sequences were collected from databases and previous studies, specifically, TRANSFAC (2019.2), JASPAR (2018), UniPROBE (2018), DNA motifs from high-throughput SELEX, DNA motifs from ENCODE DBP ChIP-seq data, and HOCOMOCO versions 9 and 11 [17-22]. To reduce the number of false-positive DBS predictions, only DBSs that overlapped the ChIP-seq peaks of the same DBP (GTRD database release v20.06 [https://gtrd.biouml.org]) were used for further analyses. The genomic positions of ChIP-seq and DNase-seq data in public databases were converted to those provided in the hg19 version of the human genome using the LiftOver tool.

539 To confirm that CTCF DBSs were located at chromatin interaction anchors, we
540 examined the overlap between CTCF ChIP-seq peaks and the anchors of *in situ*
541 ChIA-pet chromatin interactions mediated by CTCF in human foreskin fibroblast (HFF)
542 cells (4D Nucleome data portal,
543 https://data.4dnucleome.org/files-processed/4DNFI73AFN12/). To eliminate DBSs that
544 overlapped EPI anchors, we obtained ChIP-seq peaks of histone modifications, such as
545 H3K4me3, H3K27ac, and H3K4me1, associated with promoter and enhancer regions in
546 HFF cells (4D Nucleome data portal, 4DNFIJOK1BWR, 4DNFI442BQLO, and
547 4DNFIAGV3X99).
548

549 **Preparation of input data considering the DBSs of an insulator-associated DBP**

550 To prepare the input data of the DBSs of DBPs in enhancers and promoters for
551 the deep learning analysis, first, a transcript regulated by a DBP bound to enhancers and
552 promoters was predicted based on the single nearest gene association rule (Fig. 2c;
553 McLean et al. [15] Fig. S2c), as we found that the association rule predicted EPIs more
554 accurately than the other rules in our previous study [15]. The maximum length of the
555 regulatory domain for each transcript was set to 1 Mbp upstream and downstream of the
556 TSS. Then, considering the genomic locations and effects of the insulator sites of a DBP
557 on EPIs, the length of a DBP ChIP-seq peak in a 100-base bin was calculated from a
558 TSS to the DBS of a predicted insulator-associated DBP. To reduce the computational

559 time of the analysis, we first calculated the length of a DBP ChIP-seq peak in the
560 regulatory domains of all transcripts in a 100-base bin. Then, for each interval of 100
561 bins between a TSS and a DBS of a predicted insulator-associated DBP in the
562 regulatory domain, we computed the maximum length of each DBP ChIP-seq peak in
563 100 bins on the basis of the precomputed length of the DBP ChIP-seq peak in each bin.
564 In the case of a set of 100 bins containing the DBS of a predicted insulator-associated
565 DBP, we computed the maximum lengths from the bin closest to a TSS to the first bin
566 containing the DBS of the insulator-associated DBP, taking into account the orientation
567 of the DBS.
568

569 **Prediction of insulator-associated DBPs using DeepLIFT tool**

570 The DeepLIFT tool (version 0.6.13.0) was used to calculate the contribution
571 (DeepLIFT score) of each DBS to the prediction of the expression level of a transcript [44].
572 The contribution scores of DBSs of DBPs in putative enhancers and promoters
573 upstream and downstream of each transcript were normalized by the sum of the absolute
574 values of all contribution scores for the transcript, as the range of contribution scores
575 changes according to the range of the predicted output (i.e., the expression level of a
576 transcript), as described in a previous study [39]. To estimate the effect of the DBSs of a
577 predicted insulator-associated DBP on the prediction of transcript levels, positive
578 normalized contribution scores for transcripts with DBSs in upstream or downstream

579 regions were subjected to comparisons between all combinations of two different
580 orientations (i.e., all combinations of two orientations among FR, RF, forward-forward
581 [FF], and reverse-reverse [RR] orientations, and no orientation [X]: FR vs. RF, FR vs. X,
582 RF vs. X, FF vs. RR, FF vs. X, RR vs. X, FF vs. FR, FF vs. RF, RR vs. FR, and RR vs.
583 RF) among the DBSs of a predicted insulator-associated DBP by using a statistical test
584 (Mann–Whitney *U* test, FDR-adjusted *p* value < 0.05 and *p* value < $10^{-3}$). For EPIs that
585 were common to the predictions based on two different DBS orientations of a predicted
586 insulator-associated DBP, the DBSs of DBPs in putative enhancers and promoters
587 among the common EPIs were omitted in the comparison of positive normalized
588 contribution scores between the two different orientations. Based on the results of the
589 statistical analysis and the mean differences in contribution scores, we predicted an
590 insulator-associated DBP on the basis of the DBS directional orientation bias or lack of
591 orientation bias; for example, to predict an insulator-associated DBP with a DBS in the
592 FR orientation, we selected a DBP for which the contribution scores among all
593 combinations of FR vs. RF, FR vs. FF, FR vs. RR, and FR vs. X were significant
594 (Mann–Whitney *U* test, FDR-adjusted *p* value < 0.05 and *p* value < $10^{-3}$) and showed
595 positive (high) mean differences among the contribution scores of all the combinations.
596
597 **Search for insulator-associated DBSs based on histone modifications and gene**
598 **expression levels**

599 The genomic positions in the bigWig file of H3K27me3 ChIP-seq data
600 (ENCFF854VTY) from HFF cells were converted to the corresponding positions in the
601 hg19 version of the human genome using the LiftOver tool and then converted to a
602 bedGraph file using the UCSC bigWigToBedGraph tool to visualize the data on the
603 UCSC Genome Browser through custom tracks
604 (http://www.genome.ucsc.edu/cgi-bin/hgCustom). The BAM file of RNA-seq data
605 (GSM941744) from HFF cells was converted to a WIG file using bam2wig.pl from
606 Bio::ToolBox. The genomic positions in the BED file of the boundaries identified
607 through a Micro-C assay (4D Nucleome data portal,
608 https://data.4dnucleome.org/files-processed/4DNFI5OZ7TBE/) were converted to the
609 corresponding positions in the hg19 version of the human genome.
610
611 **Distribution of DBSs around chromatin interaction anchors**
612 We examined the distribution of the DBSs of predicted insulator-associated DBPs
613 in regions within ±500 bp of the center of a chromatin interaction anchor using
614 chromatin interaction data from HFF cells obtained via a Micro-C assay (4D Nucleome
615 data portal, https://data.4dnucleome.org/files-processed/4DNFI1MCF9GZ/). The
616 number of DBSs of predicted insulator-associated DBPs that cooccurred with DBSs of
617 known insulator-associated DBPs, such as CTCF, RAD21, and SMC3, was counted in
618 each 10-bp bin in the regions within ±500 bp of the center of a chromatin interaction

619 anchor.

620

621 **Statistical analysis**

622 Comparison of the distribution of DeepLIFT scores between different orientations
623 of DBSs was performed using the Mann-Whitney *U* test with the
624 scipy.stats.mannwhitneyu library. Two-tailed *p* values were calculated. For multiple
625 comparisons, adjusted *p* values were calculated by false discovery rate (FDR) using the
626 Benjamini-Hochberg procedure. *Z* scores for the Mann-Whitney *U* test were not output
627 by the library. Therefore, we calculated *Z* scores by z-standardizing the *U* values. The
628 code is made available on GitHub at https://github.com/reposit2/insulator/tree/main.

629

630 **Data availability**

631 Human foreskin fibroblast (HFF) cell RNA-seq data were obtained from GSM941744.
632 Human monocyte RNA-seq data were obtained from the Blueprint DCC portal
633 http://dcc.blueprint-epigenome.eu/#/files at
634 'C0010KB1.transcript_quantification.rsem_grape2_crg.GRCh38.20150622.results'. The
635 RNA-seq data of human mammary epithelial cells (HMEC) were found on ENCODE at
636 UCSC http://genome.ucsc.edu/encode/ at
637 'wgEncodeCshlLongRnaSeqHmecCellPapAlnRep1.bam'. Human neural progenitor cell
638 (NPC) RNA-seq data were obtained from GSM915326. The DNase-seq data of HFF

639 cells were obtained from GSE170964. The *in situ* ChIA-pet chromatin interaction data
640 mediated by CTCF in HFF cells were available at the 4D Nucleome data portal,
641 https://data.4dnucleome.org/files-processed/4DNFI73AFN12/. The ChIP-seq data of
642 H3K4me3, H3K27ac, and H3K4me1 of HFF cells were found at the 4D Nucleome data
643 portal, 4DNFIJOK1BWR, 4DNFI442BQLO, and 4DNFIAGV3X99. The ChIP-seq data
644 of H3K27me3 in HFF cells were obtained from the ENCODE portal, ENCFF854VTY.
645 The chromatin interaction data and boundary data of HFF cells by the Micro-C were
646 obtained from the 4D Nucleome data portal,
647 https://data.4dnucleome.org/files-processed/4DNFI1MCF9GZ/ and
648 https://data.4dnucleome.org/files-processed/4DNFI5OZ7TBE/. ChIP-seq data of TFs
649 were downloaded from the v20.06 release of the GTRD database
650 (https://gtrd.biouml.org). eQTL data were downloaded from the v8 release of the GTEx
651 portal (https://gtexportal.org). DNA binding motifs of TFs were collected from
652 TRANSFAC (2019.2), JASPAR (2018) (https://jaspar.genereg.net/), UniPROBE (2018)
653 (http://the_brain.bwh.harvard.edu/uniprobe/), high-throughput SELEX
654 (https://doi.org/10.1016/j.cell.2012.12.009), transcription factor binding sequences of
655 ENCODE TF ChIP-seq data (http://compbio.mit.edu/encode-motifs/), and
656 HOCOMOCO version 9 and 11 (https://autosome.org/HOCOMOCOS/)
657 (https://hocomoco11.autosome.org/).
658

659 **Code availability**

660 Demonstration codes are made available at

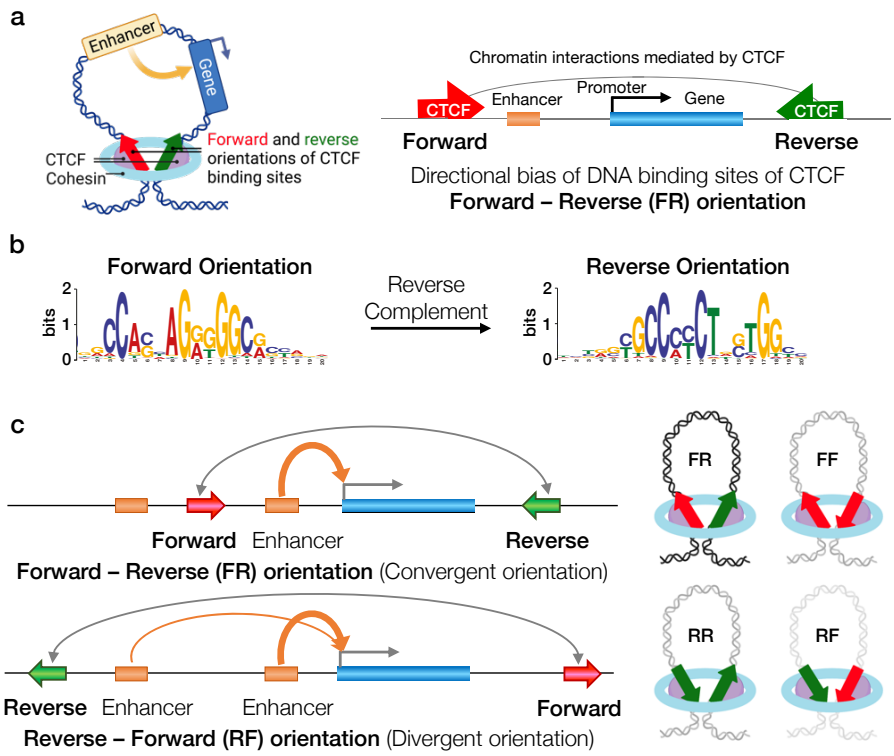
661 https://github.com/reposit2/insulator/tree/main and

662 https://doi.org/10.5281/zenodo.8216164.

663

664

665 **Figures**




666

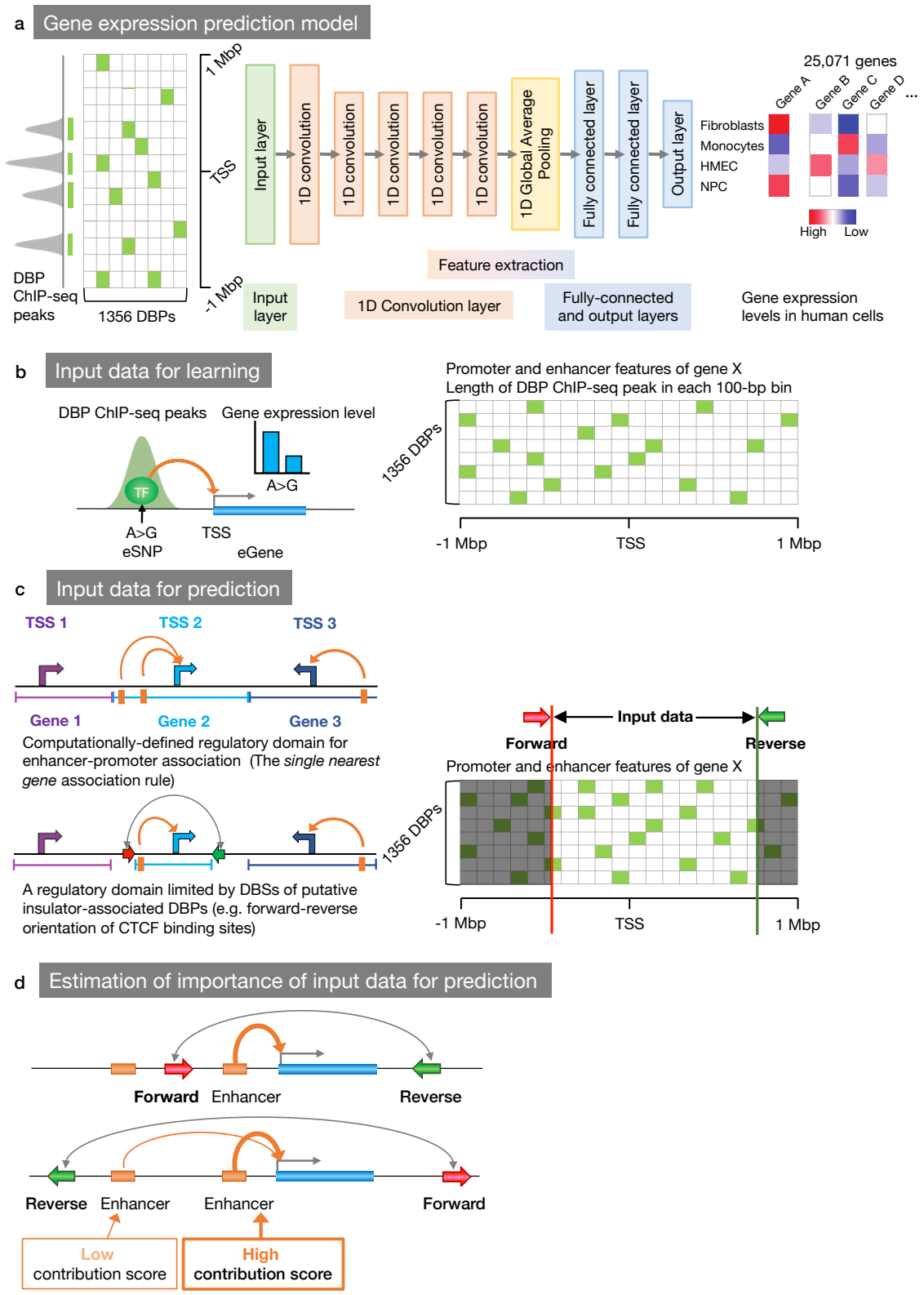
667 **Fig. 1. CTCF insulator function is involved in enhancer–promoter interactions. a**

668 CTCF DBSs with a forward–reverse orientation are frequently found at chromatin

669 interaction anchors. The forward and reverse orientations of DBSs have complementary

sequences and are relatively distinct in orientation according to reference genome sequences. **b** Position weight matrices (PWMs) of CTCF binding motifs. The forward and reverse orientations of the binding motifs are complementary on a DNA strand. A sequence logo is a graphical representation of the sequence conservation of nucleotides, and consists of a stack of letters at each position. The relative sizes of the letters indicate their frequency in the sequences (Wikipedia 'Sequence logo'). **c** CTCF can block the interaction between an enhancer and a promoter (or the transcriptional start site [TSS] of a gene), thereby restricting the activity of an enhancer to a certain functional domain (region between red and green arrows). The red and green arrows represent the DBSs of a putative insulator-associated DBP (e.g., forward–reverse orientation of the DBSs of CTCF in the figure). The numbers of DBSs with forward–reverse [FR], reverse–forward [RF], forward-forward [FF], and reverse-reverse [RR] orientations differ, as indicated by the intensity of the chromatin loop (right panel). Images of double-stranded DNA loops were adapted from "Transcriptional Regulation by CTCF and Cohesin" at BioRender.com (2023). Retrieved from https://app.biorender.com/biorender-templates.

686


a
Gene expression prediction model
1 Mbp
TSS
-1 Mbp
DBP ChIP-seq peaks
1356 DBPs
Input layer
1D convolution
1D convolution
1D convolution
1D convolution
1D convolution
1D Global Average Pooling
Fully connected layer
Fully connected layer
Output layer
25,071 genes
Gene A
Gene B
Gene C
Gene D
...
Fibroblasts
Monocytes
HMEC
NPC
High
Low
Feature extraction
Input layer
1D Convolution layer
Fully-connected and output layers
Gene expression levels in human cells
b
Input data for learning
DBP ChIP-seq peaks
Gene expression level
TF
A>G
A>G eSNP
TSS
eGene
Promoter and enhancer features of gene X
Length of DBP ChIP-seq peak in each 100-bp bin
1356 DBPs
-1 Mbp
TSS
1 Mbp
c
Input data for prediction
TSS 1
TSS 2
TSS 3
Gene 1
Gene 2
Gene 3
Computationally-defined regulatory domain for enhancer-promoter association (The single nearest gene association rule)
A regulatory domain limited by DBSs of putative insulator-associated DBPs (e.g. forward-reverse orientation of CTCF binding sites)
Input data
Forward
Reverse
Promoter and enhancer features of gene X
1356 DBPs
-1 Mbp
TSS
1 Mbp
d
Estimation of importance of input data for prediction
Forward
Enhancer
Reverse
Reverse
Enhancer
Enhancer
Forward
Low contribution score
High contribution score

687

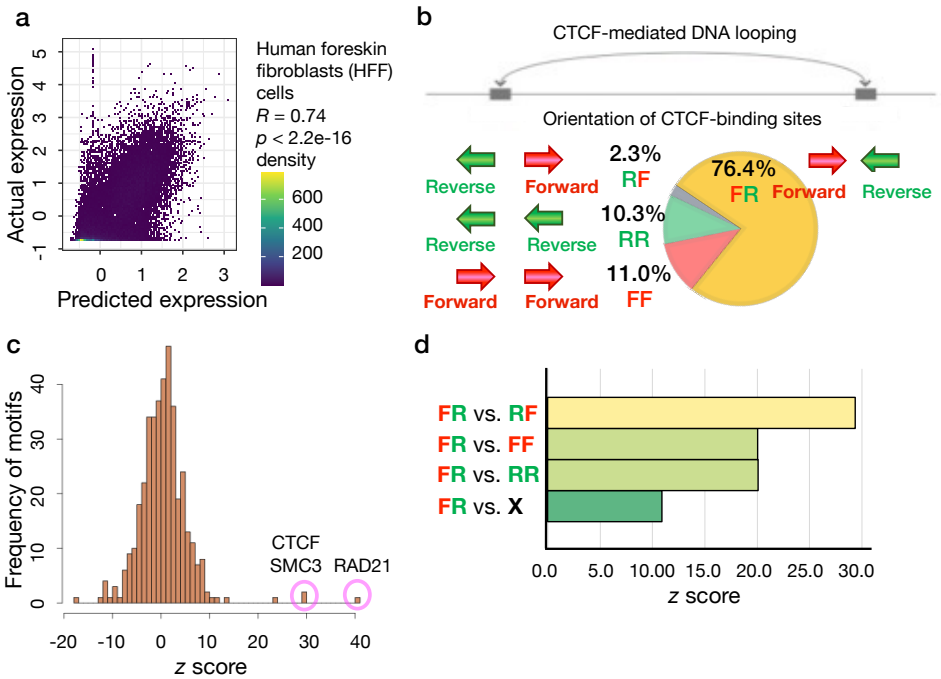
688

689 **Fig. 2. Overview of the prediction of insulator-associated DBPs. a** We used a
690 machine learning method named DEcode, with some modifications. The output data of
691 gene expression levels in human cells were predicted from the input data of genomic
692 locations of DBP binding at promoters and putative enhancers. **b** In the machine
693 learning phase, a DBP binding to a putative enhancer overlapping an eSNP (single
694 nucleotide polymorphism) was associated with a gene (eGene) that was potentially
695 regulated by the DBP based on eSNP and eGene of eQTL data. Based on the association
696 of each gene, the locations of DBP ChIP-seq peaks were binned (green box) and
697 summarized in a matrix as input data: Promoter consists of 30 bins of 100 bp between
698 -2 kbp and 1 kbp region from TSS. Enhancer regions consist of 200 bins of 10,000 bp
699 between -1 Mbp and 1 Mbp region from TSS, not including promoter. **c** In the phase of
700 putative insulator-associated DBP prediction, a DBP binding to a putative enhancer was
701 associated with a gene potentially regulated by the DBP based on a computationally
702 defined regulatory domain without using eQTL data. The *single nearest gene*
703 association rule defines a regulatory domain connecting the midpoints between the TSS
704 of a gene and the TSS of the nearest gene in the upstream and downstream regions. The
705 TSS and regulatory domain of each gene are indicated with a purple, light blue, or blue
706 arrow and a bracketed line, respectively. The regulatory domain of enhancer–promoter
707 association (EPA) was restricted by the DBSs (red and green arrows) of a putative
708 insulator-associated DBP in this study. Based on the association of each gene, the

709 locations of DBP ChIP-seq peaks were binned (green box in non-gray regions) and
710 summarized in a matrix and used as input data. **d** Estimation of the directional effect of
711 the DBSs (red and green arrows) of a putative insulator-associated DBP on the
712 prediction of gene expression levels. A DBP whose binding (orange box) is important
713 for predicting a gene expression level has a high contribution score.

714

715




716

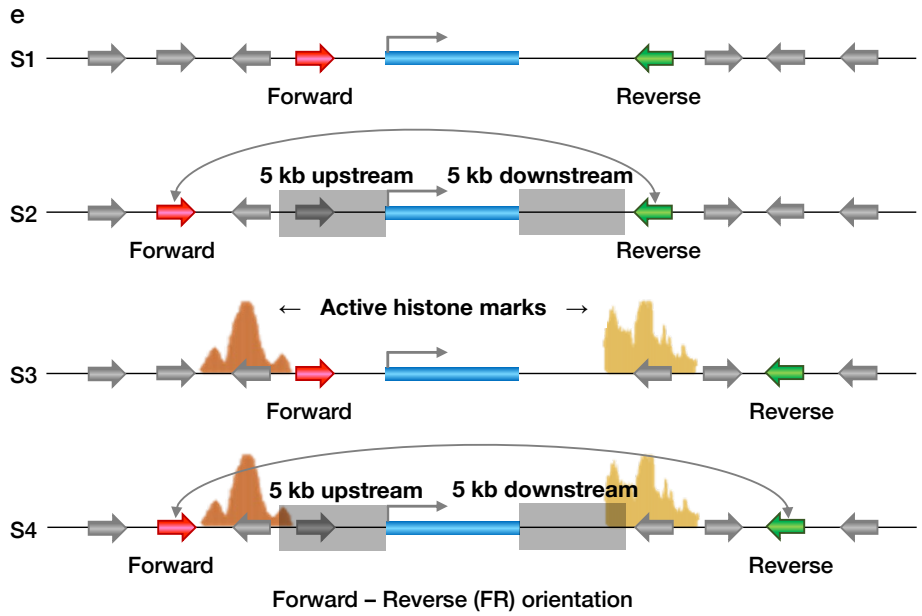
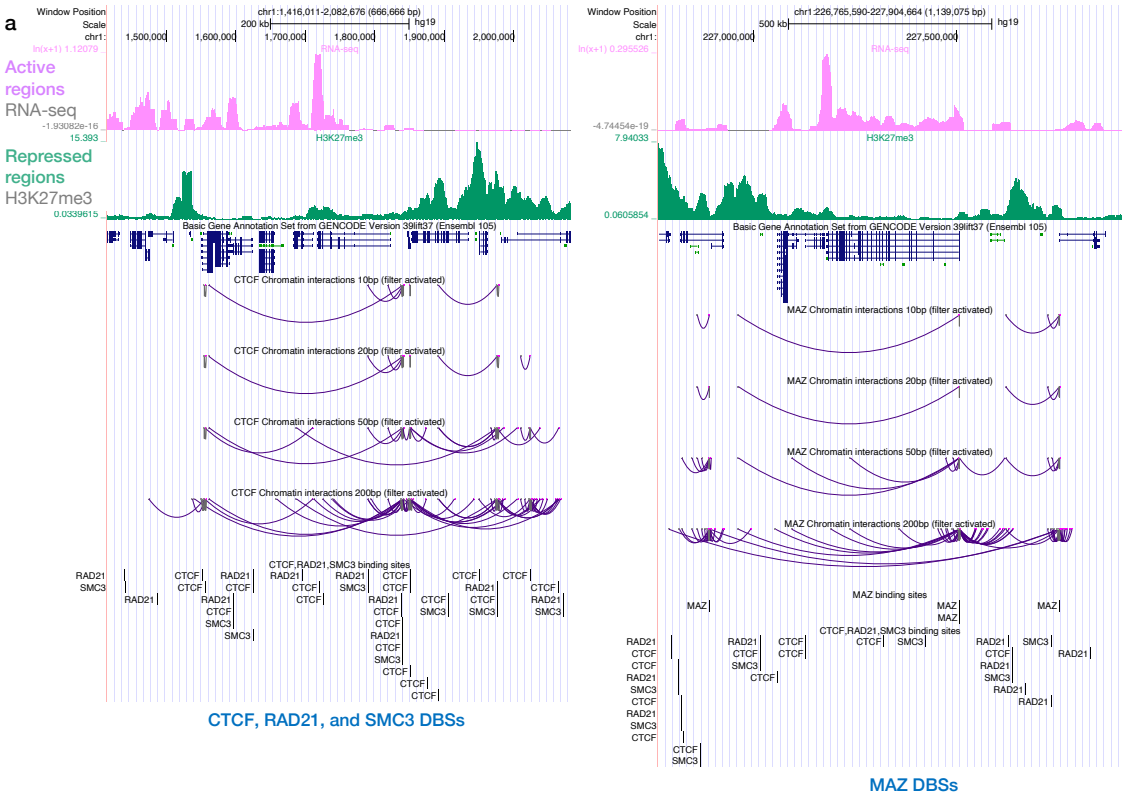



**Fig. 3. Estimation of the effect of directional regulatory motifs on gene expression.** **a** The predicted log2-TPM (transcripts per million) values of genes were compared with the actual expression levels in human foreskin fibroblast (HFF) cells using Spearman's rank correlation. **b** An FR orientation of CTCF-binding sites is common (76.4%) at chromatin interaction anchors. Forward-forward (FF), reverse-reverse (RR), and RF orientations occur at low frequencies (11.0%, 10.3%, and 2.3%, respectively) in chromatin interaction anchors. **c** Distribution of *z* scores from the Mann–Whitney *U* test for DeepLIFT scores between forward–reverse (FR) and reverse–forward (RF) orientations of DNA-binding sites of proteins (DBSs). The graph shows the results based on DBS Selection Criteria S2 in Fig. 3e and the Extended Data. **d** *Z* scores from the Mann–Whitney *U* test for DeepLIFT scores in pairwise comparisons of CTCF-binding sites with four different orientations (FR, RF, FF, and RR) and no

730 orientation (X, i.e., orientation is not distinguished). The differences in *z* scores
731 followed the same trend as the percentages of CTCF-binding site orientations in
732 chromatin interaction anchors. **e** DBS Selection Criteria S1: DBSs of predicted
733 insulator-associated DBPs; S2: DBSs of predicted insulator-associated DBPs, where the
734 DBSs overlap with a chromatin interaction anchor in a region between more than 5 kb
735 upstream and 5 kb downstream of a transcribed region; S3: DBSs of predicted
736 insulator-associated DBPs, where the DBSs do not overlap with the ChIP-seq peaks of
737 H3K4me3, H3K4me1, or H3K27ac; S4: DBSs of predicted insulator-associated DBPs,
738 where the DBSs overlap a chromatin interaction anchor in a region between more than 5
739 kb upstream and 5 kb downstream of a transcript and the DBSs do not overlap the
740 ChIP-seq peaks of H3K4me3, H3K4me1, or H3K27ac.
741



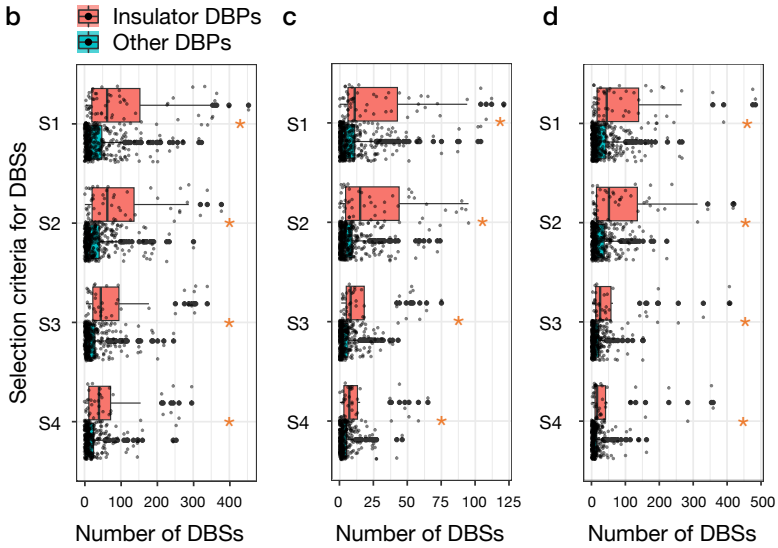




**Fig. 4. DBSs of predicted insulator-associated DBPs located at potential insulator sites. a** Distribution of H3K27me3 marks (green color) and transcript levels (pink color)

746 in upstream and downstream regions of a potential insulator site. Blue and green lines
747 below H3K27me3 marks represent transcripts of genes. The binding sites of CTCF,
748 RAD21, SMC3, and MAZ are shown as black bars. Chromatin interactions overlapped
749 with a 10bp, 20bp, 50bp, 200bp region centered on the binding sites of CTCF (left
750 panel) and MAZ (right panel) are shown as dark blue arcs, respectively. **b** Number of
751 DBSs of predicted insulator-associated DBPs, where H3K27me3 signal intensities
752 differed between the upstream and downstream regions of the DBS. * $p$ value < 0.01
753 (Table 2). **c** Number of DBSs of predicted insulator-associated DBPs, where
754 H3K27me3 signal intensities and transcript levels differed between the upstream and
755 downstream regions of the DBS. **d** Number of DBSs of predicted insulator-associated
756 DBPs in boundary regions assigned based on chromatin interaction data from human
757 foreskin fibroblast (HFF) cells. Homo sapiens (human) genome assembly GRCh37
758 (hg19) from Genome Reference Consortium was used in panels **a**.
759

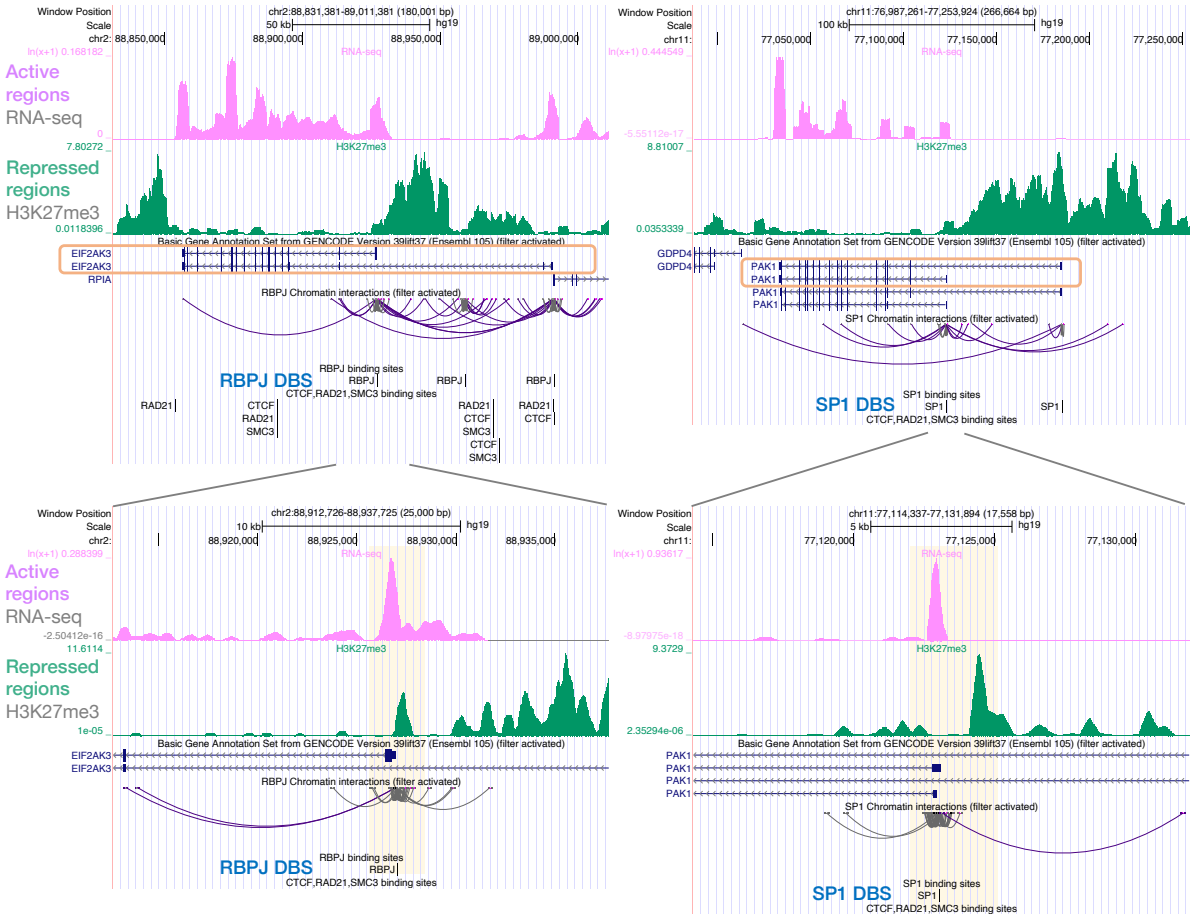



**Fig. 5. Insulator-associated DBPs potentially regulate alternative transcription.** The DBSs of insulator-associated DBPs were frequently located near transcription start and termination sites. Insulator-associated DBPs potentially regulate transcribed regions differentially repressed in alternative transcripts.

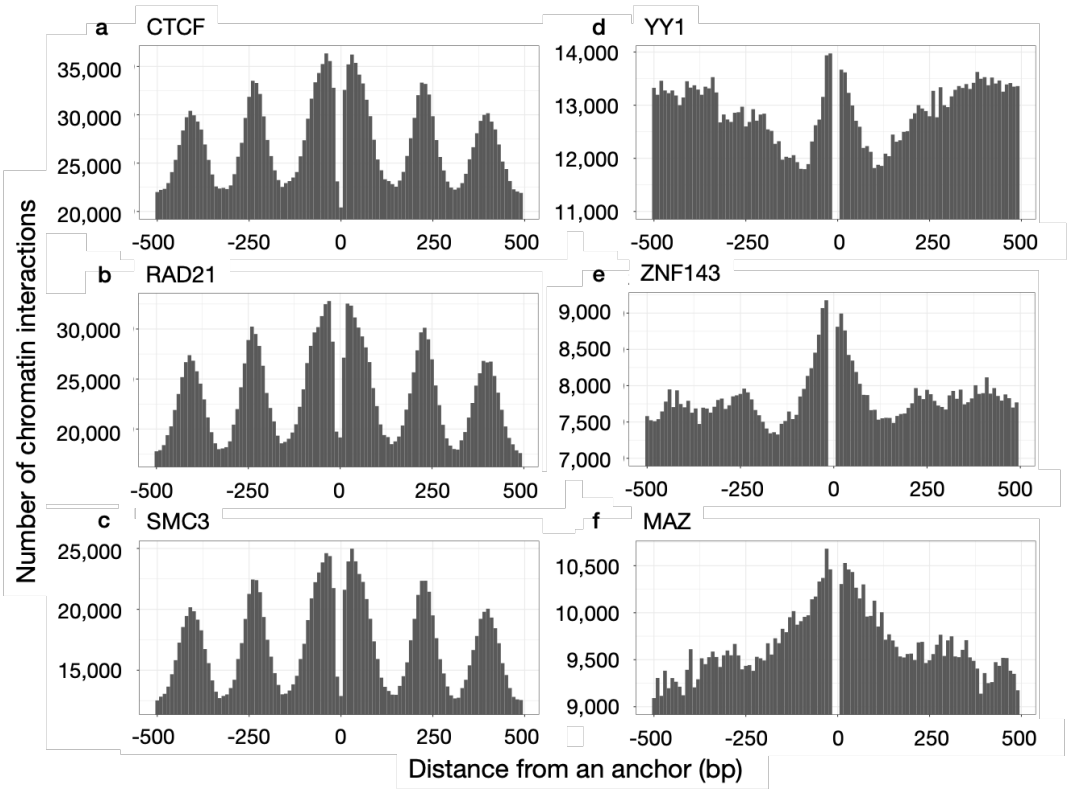



**Fig. 6. DBSs distributed around chromatin interaction anchor. a-f** The X-axis shows the distance of a DBS from a chromatin interaction anchor. The Y-axis shows the number of chromatin interaction anchors with DBSs within a ±500-bp region from the center of the interaction anchor.

# Tables

**Table 1. Predicted insulator-associated DBPs with directional or nondirectional DBSs.**

| FR orientation | | | | | Low | | | High |
|---|---|---|---|---|---|---|---|---|
| DBP | *z* score | | | | Mean difference | | | |
| | RF | FF | RR | X | RF | FF | RR | X |
| CEBPA | 5.64 | 3.99 | 5.17 | 13.79 | 1.92E-05 | 2.41E-05 | 1.33E-05 | 1.47E-04 |
| CTCF | 30.51 | 22.70 | 22.14 | 4.15 | 1.30E-04 | 1.07E-04 | 1.22E-04 | 1.54E-05 |
| CTCFL | 23.39 | 17.21 | 21.19 | 14.70 | 1.03E-04 | 7.86E-05 | 1.27E-04 | 5.62E-05 |
| EGR1 | 9.17 | 7.11 | 10.13 | 12.76 | 4.71E-05 | 3.81E-05 | 6.96E-05 | 7.41E-05 |
| EP300 | 6.81 | 4.23 | 3.94 | 3.49 | 5.24E-05 | 2.39E-05 | 5.02E-05 | 1.42E-05 |
| KLF9 | 9.11 | 6.92 | 4.52 | 17.23 | 7.56E-05 | 5.10E-05 | 6.36E-05 | 1.57E-04 |
| RAD21 | 47.70 | 34.64 | 37.59 | 24.35 | 1.97E-04 | 1.80E-04 | 1.93E-04 | 1.34E-04 |
| RBPJ | 7.52 | 4.04 | 7.55 | 5.77 | 9.29E-05 | 8.10E-05 | 1.12E-04 | 9.77E-05 |
| RUNX1 | 9.17 | 6.61 | 4.30 | 14.81 | 5.91E-05 | 4.42E-05 | 5.40E-05 | 1.87E-04 |
| SIN3A | 10.27 | 7.14 | 3.49 | 3.62 | 7.42E-05 | 6.50E-05 | 4.33E-05 | 2.03E-05 |
| SMC3 | 31.72 | 23.76 | 25.66 | 24.07 | 1.60E-04 | 1.36E-04 | 1.70E-04 | 1.67E-04 |
| ZEB1 | 3.53 | 4.32 | 3.85 | 7.06 | 2.67E-05 | 3.51E-05 | 3.69E-05 | 1.64E-04 |
| ZNF18 | 8.62 | 9.08 | 4.47 | 7.69 | 8.16E-05 | 1.17E-04 | 6.78E-05 | 3.27E-04 |

| FR orientation | | | | | Low | | | High |
|---|---|---|---|---|---|---|---|---|
| DBP | *z* score | | | | Mean difference | | | |
| | RF | FF | RR | X | RF | FF | RR | X |
| CTCF | 29.35 | 20.05 | 20.07 | 10.91 | 1.21E-04 | 9.10E-05 | 1.11E-04 | 4.12E-05 |
| CTCFL | 23.14 | 17.60 | 19.38 | 19.81 | 1.04E-04 | 8.77E-05 | 1.20E-04 | 9.44E-05 |
| EGR1 | 10.80 | 8.16 | 10.01 | 13.69 | 5.96E-05 | 3.84E-05 | 8.50E-05 | 1.08E-04 |
| EP300 | 9.77 | 7.57 | 7.91 | 10.66 | 6.67E-05 | 4.17E-05 | 7.85E-05 | 6.42E-05 |
| KLF1 | 8.29 | 5.23 | 10.08 | 10.85 | 8.32E-05 | 4.23E-05 | 1.39E-04 | 1.49E-04 |
| KLF9 | 11.70 | 7.66 | 5.01 | 12.95 | 8.88E-05 | 5.83E-05 | 6.72E-05 | 1.17E-04 |
| MAZ | 7.05 | 4.79 | 3.94 | 5.37 | 6.59E-05 | 2.47E-05 | 8.60E-05 | 6.87E-05 |
| RAD21 | 40.64 | 31.19 | 29.65 | 25.69 | 1.73E-04 | 1.64E-04 | 1.61E-04 | 1.47E-04 |
| RUNX1 | 8.47 | 5.66 | 3.50 | 13.54 | 6.70E-05 | 4.61E-05 | 6.52E-05 | 2.01E-04 |
| SIN3A | 13.18 | 6.66 | 4.99 | 4.91 | 1.02E-04 | 6.78E-05 | 6.17E-05 | 2.81E-05 |
| SMC3 | 29.76 | 22.20 | 24.11 | 29.30 | 1.56E-04 | 1.30E-04 | 1.66E-04 | 2.10E-04 |
| SP1 | 8.58 | 5.67 | 8.32 | 19.04 | 5.39E-05 | 4.78E-05 | 7.58E-05 | 2.02E-04 |
| SP2 | 8.80 | 7.55 | 3.56 | 7.37 | 7.07E-05 | 8.81E-05 | 3.55E-05 | 1.81E-04 |
| TAF1 | 6.57 | 4.74 | 5.15 | 4.91 | 1.20E-04 | 8.27E-05 | 1.35E-04 | 8.90E-05 |

**Predicted insulator-associated DBPs**

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| AHR | CTCF | FOS | JUND | MYB | PRDM9 | SMAD3 | TBX21 | ZBTB7A |
| ASCL1 | CTCFL | FOSL2 | KLF1 | MYCN | RAD21 | SMC3 | TCF12 | ZEB1 |
| ATF2 | DUX4 | FOXA3 | KLF16 | MYOD1 | RBPJ | SP1 | TCF3 | ZFP3 |
| ATF7 | EBF1 | FOXH1 | KLF5 | MYOG | REST | SP2 | TCF4 | ZIC2 |
| BATF | EGR1 | FOXK1 | KLF8 | NFYC | RUNX1 | SP5 | TEAD3 | ZNF143 |
| BATF3 | EGR2 | GATA2 | KLF9 | NR2C2 | RUNX3 | SPI1 | TFAP2C | ZNF18 |
| BCL6 | EGR3 | GRHL1 | MAFG | NR2F2 | RXRA | SPIB | TFAP4 | ZNF320 |
| CDX2 | ELK4 | GRHL2 | MAZ | NR3C1 | SCRT1 | SREBF2 | TP53 | ZNF329 |
| CEBPA | EP300 | HNF4A | MECP2 | NR5A2 | SIN3A | STAT1 | TRIM28 | ZNF35 |
| CLOCK | EPAS1 | JUN | MEF2B | PAX5 | SMAD1 | STAT3 | USF1 | ZNF449 |
| CREB5 | ETV1 | JUNB | MEIS2 | PBX2 | SMAD2 | TAF1 | VDR | ZZZ3 |

*Z* scores from the Mann–Whitney *U* test for DeepLIFT scores between forward-reverse (FR) and other orientations such as reverse-forward (RF), forward-forward (FF), reverse-reverse (RR), and non-orientation (X) of DNA-binding sites of proteins (DBSs). Mean differences of DeepLIFT scores. Color scales are shown for each DNA-binding protein (DBP). Top: DBS Selection Criteria S1 in Fig. 3e; Middle: DBS Selection Criteria S2 in Fig. 3e; Bottom: List of predicted insulator-associated DBPs. DBPs in a yellow cell are associated with insulators, CTCF, and/or other transcriptional regulations that have been shown to affect enhancer-promoter interactions (EPIs) in previous studies.

**Table 2. Number of DBSs of predicted insulator-associated DBPs where DBSs are located at insulator sites.**

| | | DBS Selection Criteria | | Median nunber of DBSs | Mean number of DBSs | Number of DBPs | *p* value |
|---|---|---|---|---|---|---|---|
| Experimental data for identifying insulator sites | H3K27me3 | S1 | Insulator-associated DBPs | 66 | 115.66 | 58 | 2.35E-08 |
| | | | Other DBPs | 16 | 38.14 | 354 | |
| | | S2 | Insulator-associated DBPs | 65 | 102.00 | 50 | 3.33E-08 |
| | | | Other DBPs | 14 | 31.73 | 346 | |
| | | S3 | Insulator-associated DBPs | 45 | 93.43 | 40 | 3.20E-08 |
| | | | Other DBPs | 12 | 25.98 | 298 | |
| | | S4 | Insulator-associated DBPs | 40 | 69.65 | 37 | 1.13E-06 |
| | | | Other DBPs | 11 | 22.45 | 274 | |
| | H3K27me3,Gene expression | S1 | Insulator-associated DBPs | 12 | 27.95 | 58 | 4.26E-06 |
| | | | Other DBPs | 4 | 10.04 | 354 | |
| | | S2 | Insulator-associated DBPs | 16 | 27.30 | 50 | 2.14E-06 |
| | | | Other DBPs | 4 | 8.28 | 346 | |
| | | S3 | Insulator-associated DBPs | 8 | 18.05 | 40 | 9.89E-06 |
| | | | Other DBPs | 2 | 5.09 | 298 | |
| | | S4 | Insulator-associated DBPs | 7 | 13.97 | 37 | 9.17E-05 |
| | | | Other DBPs | 2 | 4.46 | 274 | |
| | Boundary | S1 | Insulator-associated DBPs | 48 | 96.36 | 58 | 6.43E-07 |
| | | | Other DBPs | 14 | 30.88 | 354 | |
| | | S2 | Insulator-associated DBPs | 54 | 91.58 | 50 | 5.39E-07 |
| | | | Other DBPs | 13 | 26.70 | 346 | |
| | | S3 | Insulator-associated DBPs | 26 | 74.33 | 40 | 4.57E-07 |
| | | | Other DBPs | 7 | 15.17 | 298 | |
| | | S4 | Insulator-associated DBPs | 16 | 58.08 | 37 | 1.27E-05 |
| | | | Other DBPs | 7 | 13.75 | 274 | |

789

790 DBS Selection Criteria S1: DBSs of predicted insulator-associated DBPs; DBS
791 Selection Criteria S2: DBSs of predicted insulator-associated DBPs, where the DBSs
792 overlap with a chromatin interaction anchor in a region between more than 5 kb
793 upstream and 5 kb downstream of a transcribed region; DBS Selection Criteria S3:
794 DBSs of predicted insulator-associated DBPs, where the DBSs do not overlap with the
795 ChIP-seq peaks of H3K4me3, H3K4me1, or H3K27ac; DBS Selection Criteria S4:
796 DBSs of predicted insulator-associated DBPs, where the DBSs overlap a chromatin
797 interaction anchor in a region between more than 5 kb upstream and 5 kb downstream

798 of a transcript and the DBSs do not overlap the ChIP-seq peaks of H3K4me3,

799 H3K4me1, or H3K27ac. *P* values were calculated from the Mann–Whitney *U* test.

800

801 **Table 3. Pairs of DBSs colocalize at the same distance from chromatin interaction**

802 **anchors.**

| DBP1 | DBP2 | Position start | Position end | # chromatin interactions with DBP1 sites | # chromatin interactions with DBP2 sites | # chromatin interactions with DBP1 and 2 sites | Ratio (DBP1 and 2) / DBP1 | Ratio (DBP1 and 2) / DBP2 | Max ratio (DBP1 or DBP2) |
|---|---|---|---|---|---|---|---|---|---|
| CTCF | CTCFL | 30 | 39 | 36229 | 27159 | 19743 | 0.54 | 0.73 | 0.73 |
| CTCF | PPARG | 30 | 39 | 36229 | 2577 | 528 | 0.01 | 0.20 | 0.20 |
| CTCF | RAD21 | -10 | -1 | 23108 | 19746 | 7661 | 0.33 | 0.39 | 0.39 |
| CTCF | RXRA | 220 | 229 | 33327 | 13204 | 9102 | 0.27 | 0.69 | 0.69 |
| CTCF | SMC3 | -30 | -21 | 35558 | 24371 | 19701 | 0.55 | 0.81 | 0.81 |
| CTCF | ZBTB7A | 30 | 39 | 36229 | 5190 | 1539 | 0.04 | 0.30 | 0.30 |
| | | | | | | | | | |
| YY1 | CTCF | -50 | -41 | 12717 | 35247 | 291 | 0.02 | 0.01 | 0.02 |
| YY1 | RAD21 | -30 | -21 | 13931 | 32793 | 203 | 0.01 | 0.01 | 0.01 |
| | | | | | | | | | |
| ZNF143 | CTCF | 20 | 29 | 8993 | 35212 | 430 | 0.05 | 0.01 | 0.05 |
| ZNF143 | RAD21 | -30 | -21 | 9064 | 32793 | 500 | 0.06 | 0.02 | 0.06 |
| | | | | | | | | | |
| BACH2 | ATF3 | 290 | 299 | 2496 | 20052 | 634 | 0.25 | 0.03 | 0.25 |
| BACH2 | BACH1 | -280 | -271 | 2471 | 2863 | 730 | 0.30 | 0.25 | 0.30 |
| BACH2 | BATF | -250 | -241 | 2709 | 5924 | 486 | 0.18 | 0.08 | 0.18 |
| BACH2 | FOS | -240 | -231 | 2769 | 39887 | 884 | 0.32 | 0.02 | 0.32 |
| BACH2 | FOSB | -70 | -61 | 2774 | 16108 | 602 | 0.22 | 0.04 | 0.22 |
| BACH2 | FOSL1 | 240 | 249 | 2589 | 19417 | 687 | 0.27 | 0.04 | 0.27 |
| BACH2 | FOSL2 | 260 | 269 | 2649 | 37682 | 837 | 0.32 | 0.02 | 0.32 |
| BACH2 | JUN | 260 | 269 | 2649 | 36745 | 783 | 0.30 | 0.02 | 0.30 |
| BACH2 | JUNB | 440 | 449 | 2519 | 23889 | 535 | 0.21 | 0.02 | 0.21 |
| BACH2 | JUND | 260 | 269 | 2649 | 40876 | 925 | 0.35 | 0.02 | 0.35 |
| BACH2 | MAF | 430 | 439 | 2500 | 1520 | 367 | 0.15 | 0.24 | 0.24 |
| BACH2 | MAFF | -220 | -211 | 2676 | 3302 | 295 | 0.11 | 0.09 | 0.11 |
| BACH2 | MAFG | 180 | 189 | 2514 | 2738 | 385 | 0.15 | 0.14 | 0.15 |
| BACH2 | MAFK | -30 | -21 | 2726 | 7561 | 896 | 0.33 | 0.12 | 0.33 |
| BACH2 | MYC | 90 | 99 | 2674 | 6900 | 417 | 0.16 | 0.06 | 0.16 |
| BACH2 | NFE2 | -10 | -1 | 2399 | 8289 | 1284 | 0.54 | 0.15 | 0.54 |
| BACH2 | NFE2L2 | 440 | 449 | 2519 | 4908 | 928 | 0.37 | 0.19 | 0.37 |

803

# chromatin interactions with DBP1 and DBP2 sites shows the number of chromatin interaction anchors with DBP1 and DBP2 binding sites within ±500 bp of the anchor. Ratio (DBP1 and 2)/DBP2 is the ratio of the number of chromatin interaction anchors with both DBP1 and DBP2 binding sites to the number of anchors with DBP2 binding sites.

## References


1 Weintraub, A. S. *et al.* YY1 Is a Structural Regulator of Enhancer-Promoter Loops. *Cell* **171**, 1573-1588 e1528, doi:10.1016/j.cell.2017.11.008 (2017).

2 Rao, S. S. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665-1680, doi:10.1016/j.cell.2014.11.021 (2014).

3 Guo, Y. *et al.* CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function. *Cell* **162**, 900-910, doi:10.1016/j.cell.2015.07.038 (2015).

4 Ortabozkoyun, H. *et al.* CRISPR and biochemical screens identify MAZ as a cofactor in CTCF-mediated insulation at Hox clusters. *Nature genetics* **54**, 202-212, doi:10.1038/s41588-021-01008-5 (2022).

5 Goel, V. Y., Huseyin, M. K. & Hansen, A. S. Region Capture Micro-C reveals coalescence of enhancers and promoters into nested microcompartments. *Nature genetics*, doi:10.1038/s41588-023-01391-1 (2023).

6 Bailey, S. D. *et al.* ZNF143 provides sequence specificity to secure chromatin interactions at gene promoters. *Nat Commun* **2**, 6186, doi:10.1038/ncomms7186 (2015).

7 Lis, M. & Walther, D. The orientation of transcription factor binding site motifs

in gene promoter regions: does it matter? *BMC Genomics* **17**, 185, doi:10.1186/s12864-016-2549-x (2016).

8 Nora, E. P. *et al.* Molecular basis of CTCF binding polarity in genome folding. *Nat Commun* **11**, 5612, doi:10.1038/s41467-020-19283-x (2020).

9 Li, Y. *et al.* The structural basis for cohesin-CTCF-anchored loops. *Nature* **578**, 472-476, doi:10.1038/s41586-019-1910-z (2020).

10 Pugacheva, E. M. *et al.* CTCF mediates chromatin looping via N-terminal domain-dependent cohesin retention. *Proc Natl Acad Sci U S A* **117**, 2020-2031, doi:10.1073/pnas.1911708117 (2020).

11 Tao, H. *et al.* Computational methods for the prediction of chromatin interaction and organization using sequence and epigenomic profiles. *Brief Bioinform* **22**, doi:10.1093/bib/bbaa405 (2021).

12 Zhang, K., Li, N., Ainsworth, R. I. & Wang, W. Systematic identification of protein combinations mediating chromatin looping. *Nat Commun* **7**, 12249, doi:10.1038/ncomms12249 (2016).

13 Xie, X. *et al.* Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature* **434**, 338-345, doi:10.1038/nature03441 (2005).

14 Tasaki, S., Gaiteri, C., Mostafavi, S. & Wang, Y. Deep learning decodes the principles of differential gene expression. *Nat Mach Intell* **2**, 376-386, doi:10.1038/s42256-020-0201-6 (2020).

15 McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nature biotechnology* **28**, 495-501, doi:10.1038/nbt.1630 (2010).

16 Sherwood, R. I. *et al.* Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. *Nature biotechnology* **32**, 171-178, doi:10.1038/nbt.2798 (2014).

17 Wingender, E., Dietze, P., Karas, H. & Knuppel, R. TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic acids research* **24**, 238-241 (1996).

18 Newburger, D. E. & Bulyk, M. L. UniPROBE: an online database of protein

860 binding microarray data on protein-DNA interactions. *Nucleic acids research* **37**,
861 D77-82, doi:10.1093/nar/gkn660 (2009).
862 19 Portales-Casamar, E. *et al.* JASPAR 2010: the greatly expanded open-access
863 database of transcription factor binding profiles. *Nucleic acids research* **38**,
864 D105-110, doi:10.1093/nar/gkp950 (2010).
865 20 Jolma, A. *et al.* DNA-binding specificities of human transcription factors. *Cell*
866 **152**, 327-339, doi:10.1016/j.cell.2012.12.009 (2013).
867 21 Kheradpour, P. & Kellis, M. Systematic discovery and characterization of
868 regulatory motifs in ENCODE TF binding experiments. *Nucleic acids research*
869 **42**, 2976-2987, doi:10.1093/nar/gkt1249 (2014).
870 22 Kulakovskiy, I. V. *et al.* HOCOMOCO: towards a complete collection of
871 transcription factor binding models for human and mouse via large-scale
872 ChIP-Seq analysis. *Nucleic acids research* **46**, D252-D259,
873 doi:10.1093/nar/gkx1106 (2018).
874 23 Cuddapah, S. *et al.* Global analysis of the insulator binding protein CTCF in
875 chromatin barrier regions reveals demarcation of active and repressive domains.
876 *Genome research* **19**, 24-32, doi:10.1101/gr.082800.108 (2009).
877 24 Huang, S., Li, X., Yusufzai, T. M., Qiu, Y. & Felsenfeld, G. USF1 recruits
878 histone modification complexes and is critical for maintenance of a chromatin
879 barrier. *Mol Cell Biol* **27**, 7991-8002, doi:10.1128/MCB.01326-07 (2007).
880 25 Crane, E. *et al.* Condensin-driven remodelling of X chromosome topology
881 during dosage compensation. *Nature* **523**, 240-244, doi:10.1038/nature14450
882 (2015).
883 26 Reyes, A. & Huber, W. Alternative start and termination sites of transcription
884 drive most transcript isoform differences across human tissues. *Nucleic acids*
885 *research* **46**, 582-592, doi:10.1093/nar/gkx1165 (2018).
886 27 Xiao, T., Li, X. & Felsenfeld, G. The Myc-associated zinc finger protein (MAZ)
887 works together with CTCF to control cohesin positioning and genome
888 organization. *Proc Natl Acad Sci U S A* **118**, doi:10.1073/pnas.2023127118
889 (2021).

28 Ishihara, K., Oshimura, M. & Nakao, M. CTCF-dependent chromatin insulator is linked to epigenetic remodeling. *Molecular cell* **23**, 733-742, doi:10.1016/j.molcel.2006.08.008 (2006).

29 Vierbuchen, T. *et al.* AP-1 Transcription Factors and the BAF Complex Mediate Signal-Dependent Enhancer Selection. *Molecular cell* **68**, 1067-1082 e1012, doi:10.1016/j.molcel.2017.11.026 (2017).

30 Gong, Q. H., McDowell, J. C. & Dean, A. Essential role of NF-E2 in remodeling of chromatin structure and transcriptional activation of the epsilon-globin gene in vivo by 5' hypersensitive site 2 of the beta-globin locus control region. *Mol Cell Biol* **16**, 6055-6064, doi:10.1128/MCB.16.11.6055 (1996).

31 Karollus, A., Mauermeier, T. & Gagneur, J. Current sequence-based models capture gene expression determinants in promoters but mostly ignore distal enhancers. *Genome Biol* **24**, 56, doi:10.1186/s13059-023-02899-9 (2023).

32 Fulco, C. P. *et al.* Activity-by-contact model of enhancer-promoter regulation from thousands of CRISPR perturbations. *Nature genetics* **51**, 1664-1669, doi:10.1038/s41588-019-0538-0 (2019).

33 Lee, R. *et al.* CTCF-mediated chromatin looping provides a topological framework for the formation of phase-separated transcriptional condensates. *Nucleic acids research* **50**, 207-226, doi:10.1093/nar/gkab1242 (2022).

34 Amankwaa, B., Schoborg, T. & Labrador, M. <i>Drosophila</i> insulator proteins exhibit in vivo liquid-liquid phase separation properties. *Life Sci Alliance* **5**, doi:10.26508/lsa.202201536 (2022).

35 Nuebler, J., Fudenberg, G., Imakaev, M., Abdennur, N. & Mirny, L. A. Chromatin organization by an interplay of loop extrusion and compartmental segregation. *Proc Natl Acad Sci U S A* **115**, E6697-E6706, doi:10.1073/pnas.1717730115 (2018).

36 Phillips, J. E. & Corces, V. G. CTCF: master weaver of the genome. *Cell* **137**, 1194-1211, doi:10.1016/j.cell.2009.06.001 (2009).

37 Kumar, S., Kaur, S., Seem, K., Kumar, S. & Mohapatra, T. Understanding 3D Genome Organization and Its Effect on Transcriptional Gene Regulation Under

Environmental Stress in Plant: A Chromatin Perspective. *Front Cell Dev Biol* **9**, 774719, doi:10.3389/fcell.2021.774719 (2021).

38 Dekker, J. & Misteli, T. Long-Range Chromatin Interactions. *Cold Spring Harb Perspect Biol* **7**, a019356, doi:10.1101/cshperspect.a019356 (2015).

39 Kim, D. S. *et al.* The dynamic, combinatorial cis-regulatory lexicon of epidermal differentiation. *Nature genetics* **53**, 1564-1576, doi:10.1038/s41588-021-00947-3 (2021).

40 Liu, M. *et al.* Genomic discovery of potent chromatin insulators for human gene therapy. *Nature biotechnology* **33**, 198-203, doi:10.1038/nbt.3062 (2015).

41 Sloan, C. A. *et al.* ENCODE data at the ENCODE portal. *Nucleic acids research* **44**, D726-732, doi:10.1093/nar/gkv1160 (2016).

42 Wang, L., Wang, S. & Li, W. RSeQC: quality control of RNA-seq experiments. *Bioinformatics (Oxford, England)* **28**, 2184-2185, doi:10.1093/bioinformatics/bts356 (2012).

43 Kolmykov, S. *et al.* GTRD: an integrated view of transcription regulation. *Nucleic acids research* **49**, D104-D111, doi:10.1093/nar/gkaa1057 (2021).

44 Shrikumar, A., Greenside, P. & Kundaje, A. in *Proceedings of the 34th International Conference on Machine Learning* Vol. 70 (eds Precup Doina & Teh Yee Whye) 3145--3153 (PMLR, Proceedings of Machine Learning Research, 2017).

## Acknowledgements

Supercomputing resources were provided by the Human Genome Center of the Institute of Medical Science at the University of Tokyo and ROIS National Institute of Genetics. This work was partially supported by JSPS KAKENHI (Grant Number 16K00387), JST CREST (Grant Number JPMJCR15G1), Osaka University start-up research grant,

KIOXIA young scientist research grant at Waseda University, Sumitomo basic science research projects to NO, AMED [grant nos. JP21ae0121049, 21479280], and JSPS KAKENHI [grant nos. 22H04925, 20H00624, 17K20032] to MH.

## Author contributions

Conceptualization: NO

Methodology: NO

Investigation: NO

Visualization: NO, MH

Supervision: NO

Writing—original draft: NO

Writing—review & editing: NO, MH

## Competing interests

The authors declare no competing financial interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1101/573595.

**Correspondence** and requests for materials should be addressed to Naoki Osato.