**Title**

# Systematic discovery of directional regulatory motifs associated with human insulator sites

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

[1] Department of Life Science and Technology, Institute of Science Tokyo, Tokyo, 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: +81-3-5734-3777

Email: naokiosato11@gmail.com, osato.n.a71d@m.isct.ac.jp

## Abstract

Insulator proteins act as a barrier of enhancer–promoter interactions (EPIs). The main insulator protein in vertebrates is CTCF, a DNA-binding protein (DBP). However, other DBPs associated with insulators of EPIs are still unclear. Hence, we developed a systematic, comprehensive deep learning-based approach for identifying the DNA motifs of DBPs associated with insulators. We discovered 97 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 observed in chromatin interactions. The motifs were significantly more abundant at insulator sites separated by repressive and active regions, at boundary sites identified by chromatin interaction data, and at splice sites than other DBPs. Finally, the insulator-pairing model explains that homologous and heterologous insulator-insulator pairing interactions are orientation-dependent. These findings contribute to elucidate novel transcriptional regulations.

## Introduction

Enhancer-promoter interactions (EPIs), partitioned by chromatin interactions, regulate the transcription of genes[1]. Chromatin interactions form loop structures bundled by transcription factors (TFs) and DNA-binding proteins (DBPs) such as CTCF

37 and members of the cohesin complex (RAD21 and SMC3). In general, CTCF is
38 considered to function as an insulator by holding together chromatin loops and limiting
39 the activity of enhancers within the loops[2]. A pair of CTCFs indicate a characteristic
40 feature in their DNA-binding sites, which show directional bias in favor of a
41 forward–reverse (FR) orientation (rather than a FF, RR, or RF orientation)[3]. The
42 forward and reverse orientations of DNA-binding sites are complementary in sequence
43 and are relatively distinct in orientation according to reference genome sequences.
44 Moreover, although the FR and RF orientations of DNA-binding sites are
45 complementary, their frequencies are significantly different at chromatin loops[3],
46 suggesting that the directional bias of DNA-binding sites are associated with chromatin
47 loops.

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

54 To identify insulator-associated DBPs, one possible approach is to analyze the
55 enrichment of DNA-binding sites of chromatin- and insulator-associated DBPs at
56 chromatin interaction sites. For example, DBPs such as ZNF143 and YY1 are

57 associated with chromatin interactions and EPIs [1,6]. However, the binding site of
58 ZNF143 is enriched at chromatin interaction sites, whereas the binding site of YY1 is
59 not enriched in these sites [6]. CTCF, ZNF143, and YY1 participate in various biological
60 functions in addition to chromatin interactions. This multifunctionality of DBPs likely
61 reduces the enrichment of motifs at chromatin interaction sites, making them more
62 difficult to identify.

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

74 To predict chromatin interactions and associated DBPs, computational methods
75 have been developed [11 12]. However, these methods were not designed to identify the
76 binding motifs of DBPs associated with insulators. As for the binding motifs of DBPs,

77 Xie et al. suggested that the next challenge is to develop systematic methods for
78 identifying their specific functions in a genome-wide manner [13]. As for the regulatory
79 function of gene expression, insulators affect EPIs and gene expression, and thus the use
80 of gene expression data would improve the identification of DBPs associated with
81 insulators, compared to solely relying on the enrichment analysis or directional bias of
82 DNA-binding sites. To analyze gene expression levels in human tissues, a novel
83 computational method was developed that uses deep learning to estimate expression
84 levels based on the DNA-binding sites in promoter regions [14]. However, the method
85 does not use the DNA-binding sites associated with EPIs as input data, and was not
86 designed to identify DBPs associated with insulators.

87 Hence, we aim to search for insulator-associated DBPs using a computational
88 approach integrating and analyzing genome-wide experimental data. To identify
89 insulator-associated DBPs in the present study, we revise the previously developed deep
90 learning method with the information of EPIs based on eQTL data and the
91 enhancer–promoter association rule (McLean et al. Supplementary Fig. 2) [15]. We
92 examine the directional bias of DNA-binding sites of CTCF and insulator-associated
93 DBPs, including those of known DBPs such as RAD21 and SMC3, using contribution
94 scores based on the deep learning method, and show that the directional bias of CTCF
95 DNA-binding sites is consistently proportional to the CTCF orientation bias observed in
96 chromatin interactions. Furthermore, statistical testing the contribution scores between

97 the directional and non-directional DNA-binding sites of insulator-associated DBPs
98 revealed that the directional sites contributed more significantly to the prediction of
99 gene expression levels than the non-directional sites. We identify 23 insulator-associated
100 DBPs that are associated with insulator function, CTCF, and/or other types of
101 transcriptional regulation affecting EPIs in previous studies. To validate these findings,
102 we demonstrate that the DNA-binding sites of the identified insulator-associated DBPs
103 are located within potential insulator sites, and some of the DNA-binding sites in the
104 insulator site are found without the nearby DNA-binding sites of CTCF and cohesin.
105 Homologous and heterologous insulator-insulator pairing interactions are
106 orientation-dependent, as suggested by the insulator-pairing model based on
107 experimental analysis in flies. Our method and analyses contribute to the identification
108 of insulator- and chromatin-associated DNA-binding sites that influence EPIs and
109 reveal novel functional roles and molecular mechanisms of DBPs associated with
110 transcriptional condensation, phase separation and transcriptional regulation.
111

112 ## Results

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

114 Insulator proteins act as a barrier of EPIs (Fig. 1a, b, c). To predict the EPIs
115 between DNA-binding sites distal to transcription start sites (TSSs) and the
116 transcription initiation complex at TSSs, we identified the interactions between the

117 DNA-binding sites in putative enhancers and genes based on expression quantitative
118 trait locus (eQTL) data showing the association between a SNP (single nucleotide
119 polymorphism) and a gene expression level (Fig. 2a, b). The eQTL data were derived
120 from the GTEx v8 dataset, after quality control, consisting of 838 donors and 17,382
121 samples from 52 tissues and two cell lines [16]. If a DNA-binding site was located within
122 a 100-bp region around a single-nucleotide polymorphism (SNP) of an eQTL, we
123 assumed that the DBP regulated the expression of the transcript corresponding to the
124 eQTL. To predict gene expression levels based on DNA-binding sites of DBPs around
125 the TSS of a gene, we modified a machine learning tool reported in a previous study to
126 train with DNA-binding sites distal as well as proximal to the TSS [14]. DNA-binding
127 sites of DBPs were obtained from ChIP-seq data from the Gene Transcription
128 Regulation Database (GTRD, v20.06 release). To avoid the cellular heterogeneity found
129 in tissues, we used gene expression data from human cell lines, including human
130 foreskin fibroblasts (HFFs), monocytes, human mammary epithelial cells (HMECs),
131 and neural progenitor cells. The machine learning tool was trained to predict gene
132 expression levels in the cell lines based on DNA-binding sites of DBPs around TSSs of
133 genes. Through the training, the tool learned to select the binding sites of DNA-binding
134 proteins from ChIP-seq assays that were suitable for predicting gene expression levels
135 in the cell types. The binding sites of a DNA-binding protein tend to be observed in
136 common across multiple cell and tissue types. Therefore, ChIP-seq data and eQTL data

137 in different cell and tissue types were used as input data for learning, and then the tool
138 selected the data suitable for predicting gene expression levels in the cell types, even if
139 the data were not obtained from the same cell types. The predicted expression levels
140 were highly consistent with the actual expression levels (Spearman's rho = 0.74, 0.71,
141 0.74, and 0.73 for all 27,428 genes, and 0.70, 0.67, 0.69, and 0.70 for the 2,705 genes as
142 test data in the four cell types) without using the data on the RNA-binding proteins and
143 miRNAs obtained via the original machine learning method (Fig. 3a; Fig. S2a) [14].

144 Then, to investigate the DNA-binding sites of putative insulator-associated DBPs,
145 we obtained ChIP-seq data from the GTRD (v20.06 release), and searched the ChIP-seq
146 peaks of candidate DBPs for associated binding motifs. First, the DNA-binding sites of
147 a given DBP in open chromatin regions were predicted using the Protein Interaction
148 Quantitation (PIQ) tool with DNase-seq data and DBP-binding motif sequences, which
149 were obtained from databases and publications, specifically, TRANSFAC (2019.2),
150 JASPAR (2018), UniPROBE (2018), DNA motifs from high-throughput SELEX, DNA
151 motifs from ENCODE DBP ChIP-seq data, and HOCOMOCO versions 9 and 11 [17 18-23].
152 Then, to reduce the number of false-positive predictions of DNA-binding sites, we
153 selected the DNA-binding sites that overlapped with the ChIP-seq peaks of the same
154 DBP. Among the 1,310 DBPs with available ChIP-seq data, the DBP-binding motif
155 sequences were available for 715 DBPs.

156

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

158 To investigate the effect of directional differences in the DNA-binding sites of
159 putative insulator-associated DBPs on EPIs and gene expression levels, we computed
160 DeepLIFT scores of the input data (i.e., each binding site of the ChIP-seq data of
161 DNA-binding proteins) in the deep leaning analysis on gene expression levels [24].
162 DeepLIFT compares the importance of each input for predicting gene expression levels
163 to its 'reference or background level' and assigns contribution scores according to the
164 difference. DeepLIFT calculates a metric to measure the difference between an input
165 and the reference of the input. The DeepLIFT calculates a score representing the
166 contribution of the input data to the prediction. By comparing the DeepLIFT scores of
167 the DNA-binding sites of the ChIP-seq data, we compared the effects of putative
168 insulator sites with insulator-associated DNA-binding sites on the prediction of gene
169 expression levels according to the directional and positional differences among the
170 DNA-binding sites (Fig. 2d).

171 At the sites of distal CTCF-mediated chromatin interactions identified *in situ* via
172 ChIA-PET, 76.4% of the CTCF DNA-binding sites showed an FR orientation, also
173 known as a convergent orientation (Fig. 3b) [3]. To investigate the effect of orientation of
174 CTCF DNA-binding sites on the DeepLIFT scores of the DNA-binding sites in putative
175 enhancer and open chromatin regions, we searched for CTCF DNA-binding sites with
176 an FR or RF orientation in the upstream and downstream regions of TSSs and predicted

177 EPIs based on the enhancer–promoter association rule (Fig. 2c). We then compared the
178 distribution of the DeepLIFT scores and mean DeepLIFT scores between the CTCF
179 DNA-binding sites with FR and RF orientations and observed a significant difference in
180 the distribution of DeepLIFT scores (Mann–Whitney *U* test, FDR-adjusted *p* value <
181 0.05 and *p* value < $10^{-3}$; Fig. 3c, d and Table S1). Our method was able to compare the
182 effect of directional difference of CTCF DNA-binding site on the prediction of gene
183 expression level. We also performed the analysis with other selection criteria of
184 DNA-binding sites (Supplementary Fig. 1) (Data S1).

185 We performed the same analysis using the DNA-binding sites of RAD21, SMC3,
186 and other DBPs as putative insulator sites and found that the DBPs including RAD21
187 and SMC3 also showed a significant difference in the distribution of DeepLIFT scores
188 between DNA-binding sites with FR and RF orientations (Fig. 3c and Table S1). Other
189 orientations of the DNA-binding sites of putative insulator-associated DBPs, such as
190 forward-forward (FF), reverse-reverse (RR), and any direction (i.e., no directional bias),
191 were also evaluated and compared to examine the differences in the distribution of
192 DeepLIFT scores (Table S1). First, we selected 228 DBPs with at least 1,000
193 DNA-binding sites per DBP in HFF cells. Among these, 92 DBPs, including CTCF,
194 RAD21, SMC3, ZNF143, and YY1, showed significant differences in the distribution of
195 DeepLIFT scores between DNA-binding sites with different orientations (e.g., FR and
196 RF), and 79 DBPs, including YY1, presented a significant difference in the distribution

197 of DeepLIFT scores between DNA-binding sites with the same orientation (e.g., FF and
198 RR) (Fig. 3c; Fig. S2b; Table S1, DBP list 1A). We observed that the *z* score distribution
199 of the Mann-Whitney *U* test of DeepLIFT scores between the directional (e.g., FR, RF,
200 FF, and RR) and non-directional (X) DNA-binding sites shifted to higher *z* scores on the
201 X-axis of the graphs in Fig. 3c than between the two types of directional DNA-binding
202 sites, suggesting that directional DNA-binding sites contribute more to the prediction of
203 gene expression levels than non-directional DNA-binding sites. Unexpectedly, we found
204 that many DBPs, including known insulator- and chromatin-associated DBPs, showed
205 significant differences of DeepLIFT scores, and that the known DBPs had the most
206 significant differences in the distribution of DeepLIFT scores, suggesting that our
207 method worked properly. To reduce the number of DBPs with a directional bias of
208 DNA-binding sites and thus reduce the number of false-positive DBP predictions, we
209 set the *p* value threshold to $< 10^{-3}$. We confirmed that all the DBPs with *p* values $< 10^{-3}$
210 constituted the subset of DBPs with false discovery rate (FDR)-adjusted *p* values $< 0.05$
211 (included in Table S1).
212
213 **Discovery of directional and nondirectional DNA-binding sites of**
214 **insulator-associated DBPs**
215 To investigate the directional bias of DNA-binding sites based on more stringent
216 criteria, we compared DNA-binding sites with four different orientations (FR, RF, FF,

217 and RR), or no directional bias based on pairwise comparisons of the directional biases
218 of the DNA-binding sites. For the DNA-binding sites with different orientations (e.g.,
219 FR and RF) associated with 23 DBPs, including CTCF, RAD21, and SMC3, the
220 distribution of DeepLIFT scores differed significantly among those with the four
221 different orientations and no directional bias. The distribution of DeepLIFT scores
222 differed significantly among the DNA-binding sites with the same orientation (e.g., FF
223 and RR) for 17 DBPs. The distribution of DeepLIFT scores also differed significantly
224 among the nondirectional DNA-binding sites of six DBPs ($p$ value < $10^{-3}$, Table 1 and
225 Table S2, DBP list 1A; FDR-adjusted $p$ value < 0.05, Table S3, DBP list 1A). We
226 confirmed that known insulator-associated DBPs such as CTCF, RAD21, and SMC3
227 showed FR orientation of DNA-binding sites, and discovered other DNA-binding sites
228 with directional and nondirectional bias.

229 We compared the predicted directional biases of CTCF DNA-binding sites with
230 the directional biases identified based on experimental data on CTCF-mediated
231 chromatin interactions. In the analysis of ChIA-PET chromatin interaction data, the
232 orientations of CTCF DNA-binding sites that overlapped with distal CTCF-mediated
233 chromatin interaction sites were as follows: 76.4% FR, 11.0% FF, 10.3% RR, and 2.3%
234 RF (Fig. 3b; Guo Y et al. *Cell* 2015 [3], Fig. 4). Interestingly, the $p$ values obtained from
235 the statistical tests of the DeepLIFT scores (Mann–Whitney $U$ test) of CTCF, RAD21,
236 and SMC3 DNA-binding sites with FR vs. RF orientations were lower than those

237 obtained for the FR vs. RR and FR vs. FF comparisons. These results were consistent
238 with the percent differences in the orientations of CTCF DNA-binding sites at
239 chromatin interaction sites: the percent difference between the FR (76.4%) and RF
240 (2.3%) orientations was larger than the percent differences for FR (76.4%) vs. FF
241 (11.0%) and RR (10.3%) (Fig. 3c, d). These data indicated that the deep learning
242 analysis accurately predicted the percent differences in the FR, FF, RR, and RF
243 orientations of CTCF DNA-binding sites at chromatin interaction sites on the basis of
244 the DeepLIFT scores of the DNA-binding sites in putative enhancer and open chromatin
245 regions.

246 A total of 99 insulator-associated DBPs were predicted based on the stringent
247 criterion of the DNA-binding sites of a given DBP showing either a specific directional
248 bias or no directional bias, as determined by comparing the four possible orientations
249 (FR, RF, FF, and RR) and the absence of directional bias (X). Among 99
250 insulator-associated DBPs, EP300 and SIN3A do not directory interact with DNA, and
251 thus 97 insulator-associated DBPs directory bind to DNA. Most of the
252 insulator-associated DBPs showed a directional bias. Non-directional motifs were
253 identified for only 6 DBPs such as FOS, FOSL2, JUN, JUNB, and JUND with
254 palindromic regulatory motifs and SPI1 interacting with the DBPs. We found that 23
255 DBPs exhibited insulator and CTCF-associated functions and/or chromatin regulatory
256 functions that affected EPIs, as reported in previous studies: 13 DBPs were associated

257 with CTCF (Table S7). The 23 DBPs were multifunctional proteins, and their known
258 functions were transcriptional regulator, repressor, chromatin regulator, and interaction
259 with other proteins including enzymes that facilitate epigenetic modifications. These
260 findings suggested that CTCF regulated EPIs by interacting with other DBPs, and that
261 DBPs with other functions were also associated with the regulation of EPIs and
262 chromatin conformation.

263 To further investigate whether insulator-associated DBPs are involved in a pair of
264 chromatin interaction sites, we searched for the DNA-binding sites of
265 insulator-associated DBPs located at a pair of 400-bp and 200-bp long chromatin
266 interaction sites that were more than 5-kb apart and did not overlap with 2-kb upstream
267 and downstream of TSSs, considering the FR, RF, FF, or RR orientation of the pair of
268 DNA-binding sites (Fig. 3e), since 400-bp long chromatin interaction sites would
269 include the cluster of DNA-binding sites of insulator-associated DBPs at insulator sites
270 (explained in the next section and Fig. 4f). For the 400-bp long chromatin interaction
271 sites, the DNA-binding sites of 44 (45%) and 97 (100%) insulator-associated DBPs,
272 excluding EP300 and SIN3A, were identified at a pair of chromatin interaction sites or
273 at either site of 30,369 chromatin interactions, respectively. For the 200-bp long
274 chromatin interaction sites, the DNA-binding sites of 30 (31%) and 97 (100%)
275 insulator-associated DBPs were identified at a pair of chromatin interaction sites or at
276 either site of 10,384 chromatin interactions, respectively, although long-range chromatin

277 interactions would not be fully identified due to the limited coverage of chromatin
278 interaction data from the comparison of results based on the distance between a pair of
279 chromatin interaction sites [25]. CTCF, RAD21, and SMC3 showed directional
280 DNA-binding sites as well as CTCF ChIA-PET chromatin interaction data.
281
282 **Search for insulator-associated DNA-binding sites based on histone modifications**
283 **and gene expression**
284 To confirm the predicted insulator-associated DBPs via another method and select
285 their DNA-binding sites associated with insulators, we first analyzed previously
286 reported histone modification data to identify differences between the upstream and
287 downstream regions of the DNA-binding sites. CTCF bound to sites located between
288 repressed regions with H3K27me3 marks and active regions with H2AK5ac marks,
289 which showed different signal intensities between the upstream and downstream regions
290 of CTCF DNA-binding sites [26]. The transcription factor USF1 exhibits a barrier
291 function, in which histone modification factors are recruited to separate adjacent regions
292 into activating and repressive chromatin domains [27]. We selected H3K27me3 histone
293 modifications, associated with the repression of gene expression, for comparison
294 between the upstream and downstream regions of the DNA-binding sites of predicted
295 insulator-associated DBPs using H3K27me3 ChIP-seq data because H3K27me3 histone
296 marks occupy long genomic regions and are thus easily compared, between genomic

297 regions. We searched for DNA-binding sites where the mean H3K27me3 signal
298 intensities differed by more than 2-fold between the upstream and downstream regions
299 of the DNA-binding site. The higher mean signal was found to be greater than the first
300 quartile of all H3K27me3 signal intensities because some H3K27me3 signals had a low
301 intensity and appeared noisy. Thus, the criteria were sufficient to clearly distinguish
302 modification status among genomic regions.

303 We further counted the numbers of DNA-binding sites with different H3K27me3
304 signal intensities between the upstream and downstream regions of a DNA-binding site
305 and found that known insulator-associated DBPs, such as CTCF, RAD21, and SMC3,
306 generally showed the greatest numbers of DNA-binding sites (Table S4). Via statistical
307 analysis, we calculated the median and mean numbers of DNA-binding sites showing
308 different H3K27me3 signal intensities between upstream and downstream regions for
309 the DNA-binding sites of both predicted insulator-associated DBPs and other DBPs.
310 The median and mean numbers of DNA-binding sites were significantly larger among
311 predicted insulator-associated DBPs than other DBPs (Mann–Whitney $U$ test, $p$ value <
312 0.05) (Table 2, Table S4, Fig. 4a, b). We confirmed that a larger number of the
313 DNA-binding sites of known insulator-associated DBPs were located between activated
314 and repressed regions than non-insulator-associated DBPs.

315 To identify additional genomic regions obviously separating transcriptionally
316 activated and repressed regions, we also examined the expression levels of transcripts

317 using the signal intensity indicated by RNA-seq data for identifying transcriptionally
318 active regions based on the available HFF cell data. As in the analysis of H3K27me3
319 signal intensity, we searched for DNA-binding sites of predicted insulator-associated
320 DBPs with mean transcript signal intensities differing more than 1.5-fold between
321 upstream and downstream regions, and the higher mean signal was greater than the first
322 quartile of all transcript signal intensities. We again found that a significantly larger
323 number of the DNA-binding sites of predicted insulator-associated DBPs were located
324 between activated and repressed regions than other DBPs (Table 2, Table S4, Fig. 4a, c).

325 In the analysis of chromatin interaction data, insulation scores were calculated
326 based on the analysis of the boundary regions of chromatin interactions [28]. We also
327 examined the overlap of DNA-binding sites of predicted insulator-associated DBPs with
328 the boundary regions, and again found that a significantly larger number of the
329 DNA-binding sites of predicted insulator-associated DBPs were located in the boundary
330 regions than other DBPs (Table 2, Fig. 4d).

331 To determine whether the DNA-binding sites of predicted insulator-associated
332 DBPs potentially act as insulators or cofactors of insulators, we investigated the
333 genomic locations of the DNA-binding sites on the UCSC Genome Browser (Fig. 4a,
334 S3). We searched for DNA-binding sites located at boundary sites identified by Micro-C
335 chromatin interaction data from 4D Nucleome Data Portal. We also searched for
336 DNA-binding sites located between the H3K27me3 marks and transcribed regions of

genes, which were primarily located at the edges of H3K27me3 marks and were not found near the binding sites of CTCF, RAD21, and SMC3. For each DBP, we looked for insulator sites associated with the DNA-binding sites in the genome browser and selected primarily three DNA-binding sites for each DBP (Fig. S3a for boundary sites of chromatin interactions and Fig. S3f for histone marks and transcribed regions as insulator sites). We discovered DNA-binding sites in the expected genomic regions of all DBPs, suggesting that these DNA-binding sites are associated with insulator functions. Interestingly, MyoD1, which is an insulator-associated DBP, bound to an insulator and boundary site and located near SELENON (SEPN1) gene. MyoD1 and SELENON are involved in the differentiation of human skeletal muscle, suggesting that the insulator-associated DBP affects the expression of SELENON and the differentiation (Fig. 4a). The same analysis was performed using H3K9me3 marks, instead of H3K27me3 (Fig. S3b). We found that the distribution of H3K9me3 signal was different from that of H3K27me3 in some regions, and discovered the insulator-associated DNA-binding sites close to the edges of H3K9me3 regions (Fig. S3b). The insulator-associated DNA-binding sites potentially regulate boundary sites associated with other histone modification marks, including H3K9me3 marks. We also found that the multiple DNA-binding sites of insulator-associated DNA-binding proteins were located close to each other at insulator sites and were associated with distinct chromatin interactions (Fig. 4f, S3c, S3d). Topologically associated domains

357 (TADs) of chromosome are structured by architectural/insulator proteins clustered
358 around their boundaries, maintaining the megabase-scale chromatin structure, while
359 intra-TAD looping is mediated by discrete motifs for insulator factor binding [29].

360 When comparing insulator sites identified by histone modification marks and
361 transcribed regions with boundary sites based on chromatin interaction data in the
362 genome browser, we observed that the insulator sites did not always overlap with 5-kb
363 long boundary sites at the same location, some insulator sites were more than 5-kb away
364 from a boundary site. Overall, among 20,837 DNA-binding sites of the 97
365 insulator-associated proteins found at insulator sites identified by H3K27me3 histone
366 modification marks (type 1 insulator sites), 1,315 (6%) overlapped with 264 of 17,126
367 5kb long boundary sites, and 6,137 (29%) overlapped with 784 of 17,126 25kb long
368 boundary sites in HFF cells. Among 5,205 DNA-binding sites of the 97
369 insulator-associated DNA-binding proteins found at insulator sites identified by
370 H3K27me3 histone modification marks and transcribed regions (type 2 insulator sites),
371 383 (7%) overlapped with 74 of 17,126 5-kb long boundary sites, 1,901 (37%)
372 overlapped with 306 of 17,126 25-kb long boundary sites. Although CTCF-binding sites
373 separate active and repressive domains [26], the limited number of DNA-binding sites of
374 insulator-associated proteins found at type 1 and 2 insulator sites overlapped boundary
375 sites identified by chromatin interaction data. Furthermore, by analyzing the regulatory
376 regions of genes, the DNA-binding sites of the 97 insulator-associated DNA-binding

377 proteins were found (1) at the type 1 insulator sites (based on H3K27me3 marks) in the
378 regulatory regions of 3,170 genes, (2) at the type 2 insulator sites (based on H3K27me3
379 marks and gene expression levels) in the regulatory regions of 1,044 genes, and (3) at
380 insulator sites as boundary sites identified by chromatin interaction data in the
381 regulatory regions of 6,275 genes. The boundary sites showed the highest number of
382 overlaps with the DNA-binding sites. Comparing the insulator sites identified by (1)
383 and (3), 1,212 (38%) genes have both types of insulator sites. Comparing the insulator
384 sites between (2) and (3), 389 (37%) genes have both types of insulator sites. From the
385 comparison of insulator and boundary sites, we found that (1) or (2) types of insulator
386 sites overlapped or were close to boundary sites identified by chromatin interaction
387 data.

388 While preparing Fig. 4a and S3, we noticed that several DNA-binding sites of
389 insulator-associated DBPs were found in the same genomic locus, close to each other on
390 the genome. These DBPs potentially act together to form an insulator site. In addition,
391 some transcribed regions in the genome were differentially repressed in alternative
392 transcripts and overlapped with a H3K27me3 mark that was separated with the
393 DNA-binding site of an insulator-associated DBP between the repressive and active
394 regions (Fig. 5, S3e). Furthermore, the DNA-binding sites of insulator-associated DBPs
395 were frequently located near transcription start and termination sites. Alternative
396 transcription start and termination sites have been reported to explain the majority of

397 tissue-specific exon usage rather than alternative splicing[30]. We performed the statistical
398 test to estimate the enrichment of insulator-associated DNA-binding sites compared to
399 the other DNA-binding proteins, and found that the insulator-associated DNA-binding
400 sites were significantly more abundant at splice sites than the DNA-binding sites of the
401 other proteins (Fig 4e and Table 2; Mann–Whitney $U$ test, $p$ value $< 0.05$). The
402 comparison between the splice sites of both ends of first and last introns and those of
403 other introns showed the similar statistical significance of enrichment and number of
404 splice sites with the insulator-associated DNA-binding proteins (Table 2 and Table S9).
405 The DNA-binding sites of insulator-associated DBPs could affect the form of transcripts
406 and the genomic location of transcription start and termination sites. The differences in
407 the DNA-binding sites and H3K27me3 marks between tissues could contribute to the
408 cell and tissue type specific exon usage of alternative transcripts.
409

410 **Distribution of DNA-binding sites around chromatin interaction sites**

411 In this study, we aimed to predict chromatin- and insulator-associated DBPs. To
412 estimate how many DBPs are involved in chromatin interactions and examine the
413 distribution and association of DNA-binding sites around chromatin interaction sites,
414 the DNA-binding sites of DBPs in regions within ±500 bp of the centers of chromatin
415 interaction sites were identified in HFF cells. As an improved method for chromatin
416 interaction analysis, Micro-C assays enabled us to identify chromatin interactions at a

417 higher resolution than can be achieved via Hi-C analysis. Base pair resolution was
418 reported using Micro Capture-C [31]. The distribution of CTCF, RAD21, and SMC3
419 DNA-binding sites showed a characteristic pattern around chromatin interaction sites,
420 with the highest peaks located around interaction sites (Fig. 6a, b, c). The peaks of the
421 DNA-binding sites of YY1 and ZNF143, which are known to be involved in chromatin
422 interactions, were distributed around chromatin interaction sites. The DNA-binding sites
423 of YY1 were also broadly distributed in regions within ±500 bp of interaction sites,
424 suggesting that this DBP is involved in various functions in addition to chromatin
425 interactions (Fig. 6d, e). MAZ, which exhibits insulator function independent of CTCF
426 [32], indicated the peak of distribution of MAZ DNA-binding sites around interaction sites
427 (Fig. 6f). For the insulator-associated DBPs other than CTCF, RAD21, and SMC3, the
428 DNA-binding sites that do not overlap with those of CTCF, RND21, and SMC3 were
429 used to examine their distribution around interaction sites.

430 Among the 228 DNA-binding sites, more than 1,000 of the DNA-binding sites in
431 open chromatin regions in HFF cells overlapped with the ChIP-seq peaks of the DBPs,
432 and the peaks of almost all the DBPs (approximately 90%) were located around
433 interaction sites. Furthermore, we found DBPs with sharp peaks around interaction sites,
434 DBPs with broad peaks around interaction sites, and DBPs with peaks similar to those
435 of CTCF, RAD21, and SMC3 around interaction sites, suggesting that these DBPs may
436 be associated with CTCF, RAD21, and/or SMC3 (Table S8; Fig. S4). Among the

437 predicted insulator-associated DBPs, MeCP2 and PRDM9 exhibited fewer than 1,000
438 DNA-binding sites in HFF cells, and when a DBP has fewer than 1,000 DNA-binding
439 sites, the specific distribution of DNA-binding sites around chromatin interaction sites
440 may not be evident (Fig. S4). Approximately 90% of the DBPs with more than 1,000
441 DNA-binding sites had peaks around interaction sites. This indicated that the DBPs
442 interacted with other genomic regions and that the DNA-binding sites of predicted
443 insulator-associated DBPs were involved in chromatin interactions.
444
445 **Co-localization of DNA-binding sites around chromatin interaction sites**
446 Since the DNA-binding sites of the predicted insulator-associated DBPs may be
447 located close to those of known insulator-associated DBPs, such as CTCF, RAD21,
448 and/or SMC3, and, thus, be presumed to act as insulators, we next counted the number
449 of DNA-binding sites in each 10-bp bin within ±500 bp from the center of a chromatin
450 interaction site and compared these numbers between all possible pairs of DBPs. For
451 this analysis, the DNA-binding sites that overlapped with the ChIP-seq peaks of a DBP
452 were selected to reduce the number of false-positive DNA-binding site predictions, and
453 the number of DNA-binding sites per DBP may therefore have been underestimated.
454 Among the 228 DBPs (Selection Criteria S1 in Fig. S1), 245 pairs of 135 DBPs
455 colocalized around chromatin interaction sites when at least 30% of the DNA-binding
456 sites of a pair of DBPs colocalized in a bin with the same distance from the interaction

457 site (Table S5). CTCF showed a high rate of colocalization of DNA-binding sites with
458 both RAD21 (39%) and SMC3 (81%) in bins within ±500 bp from the center of a
459 chromatin interaction site, whereas YY1 and ZNF143 showed low rates of
460 colocalization of DNA-binding site with RAD21 (1.5% and 5.5%, respectively) and
461 CTCF (2.3% and 4.8%, respectively) (Table 3). For CTCFL (CCCTC-binding factor
462 like), RXRA, PPARG, NEUROD1, and ZBTB7A, more than 10% of DNA-binding sites
463 colocalized with CTCF, RAD21, and/or SMC3 in bins within the same distance from
464 chromatin interaction sites, suggesting that the insulator function of the DBPs was
465 associated with known insulator DBPs. For the other DBPs among the 228 DBPs, less
466 than 10% of DNA-binding sites colocalized with CTCF, RAD21, and/or SMC within a
467 bin. Most of the predicted insulator-associated DBPs did not have the colocalization of
468 DNA-binding sites and were not in close proximity to known insulator-associated DBPs
469 such as CTCF, RAD21, and SMC3; that is, they were not within a 10-bp bin in the
470 regions ±500 bp from the center of an interaction site.

471 We found that the multiple DNA-binding sites of insulator-associated
472 DNA-binding proteins were located close to each other at insulator sites (Fig. S3d). If
473 we identify the co-localization of the DNA-binding sites located at boundary sites of
474 chromatin interactions compared to non-boundary sites, the co-localization of
475 DNA-binding sites is potentially involved in boundary and insulator functions. Among
476 the 201 DBPs (Selection Criteria S2 in Fig. S1), we identified 197 pairs of 101 DBPs

477 that colocalized within a 10-bp bin in the regions ±500 bp from the center of a
478 chromatin interaction site in a 5-kb long boundary site compared to non-boundary sites,
479 when at least 30% of the DNA-binding sites of a pair of DBPs colocalized in a bin with
480 the same distance from the interaction site (Table S6). This suggested that the
481 insulator-associated DNA-binding proteins interact with each other at chromatin
482 interaction sites at insulator sites like CTCF and cohesin.

483 To further examine the possibility that known insulator-associated DBPs interact
484 indirectly with the other predicted insulator-associated DBPs, we investigated the
485 overlap of DNA-binding sites with the DBP ChIP-seq peaks of CTCF, RAD21, and
486 SMC3 and found that a high percentage (> 98%) of the DNA-binding sites of the three
487 DBPs overlapped with their ChIP-seq peaks in all three examined combinations.
488 However, the DNA-binding sites of almost all other predicted insulator-associated
489 DBPs overlapped with less than 50% of the ChIP-seq peaks of the three known DBPs,
490 even though the ChIP-seq data were collected from various cell and tissue types. These
491 results suggested that the predicted insulator-associated DBPs play different functional
492 roles than the known DBPs.

493 Although a DNA-binding-site distribution pattern around chromatin interaction
494 sites similar to those of CTCF, RAD21, and SMC3 was observed for DBPs such as
495 BACH2, FOS, ATF3, NFE2, and MAFK, less than 1% of the DNA-binding sites of the
496 latter set of DBPs colocalized with CTCF, RAD21, or SMC3 in a single bin (Fig. S4c).

497 However, BACH2, FOS, ATF3, NFE2, and MAFK had DNA-binding sites that
498 colocalized with each other (Table 3). These results suggested that the two different
499 groups of colocalized DBPs were independently involved in the same chromatin-related
500 function such as chromatin remodeling [33 34 35].
501

502 **Discussion**

503 Here, we present the first, to our knowledge, computational method and
504 genome-wide analysis aimed at detecting insulator-associated DBPs. Transcriptional
505 regulation mediated by the binding of DBPs at sites located far from TSSs was
506 modelled using our deep learning method based on eQTL data. Through this approach,
507 the effects of the DNA-binding sites of insulator-associated DBPs on EPIs and gene
508 expression were estimated, and insulator-associated DBPs were predicted. We further
509 identified DBPs, including known insulator-associated DBPs that exhibit
510 insulator-associated functions and regulate transcription, thereby affecting EPIs reported
511 in previous studies. The DNA-binding sites of the predicted insulator-associated DBPs
512 showed orientation bias of their DNA-binding sites; unexpectedly, the significant
513 differences in the orientation bias of CTCF DNA-binding sites based on DeepLIFT
514 scores (contribution scores) were found to be consistent with the percent differences in
515 the orientation of CTCF DNA-binding sites at CTCF-mediated chromatin interaction
516 sites. Furthermore, statistical testing of the contribution scores revealed that the

517 directional sites of DBPs contributed more significantly to the prediction of gene
518 expression levels than the non-directional sites. The DNA-binding sites of the predicted
519 insulator-associated DBPs were located at potential insulator sites more frequently than
520 those of other DBPs, according to the differences in H3K27me3 signal intensities and
521 transcript levels, and according to insulator sites as boundary sites identified by
522 chromatin interactions. We further observed that only a small percentage of the
523 DNA-binding sites of the insulator-associated DBPs colocalized with the DNA-binding
524 sites of known insulator-associated DBPs such as CTCF, RAD21, and SMC3 around
525 chromatin interaction sites. We discovered that DBPs other than the known insulator
526 DBPs and the directionality of their DNA-binding sites are associated with insulators.

527 Our deep learning analysis accurately predicted the percent differences in the FR,
528 FF, RR, and RF orientations of CTCF DNA-binding sites at chromatin interaction sites,
529 considering the genomic locations of DNA-binding sites of putative insulator-associated
530 DBPs. The other predicted insulator-associated DBPs also showed orientation bias of
531 their DNA-binding sites. These results suggest that the directional bias of DNA-binding
532 sites of insulator-associated DBPs would be involved in insulator function and
533 chromatin regulation through structural interactions among DBPs, other proteins, DNAs,
534 and RNAs. For example, the N-terminal amino acids of CTCF have been shown to
535 interact with RAD21 in chromatin loops [8-10]. Cohesin complexes that interact with
536 convergent CTCF sites, that is, the N-terminus of CTCF, might be protected from

537 WAPL, but those that interact with divergent CTCF sites, that is, the C-terminus of
538 CTCF, might not be protected from WAPL, which could release cohesin from chromatin
539 and thus disrupt cohesin-mediated chromatin loops [36]. For another example, although
540 sequence-based predictions of EPIs underestimated actual interactions, an
541 activity-by-contact (ABC) approach, based on a simple distance model, performed well,
542 suggesting that factors in addition to sequence influence EPIs [37,38]. Furthermore,
543 CTCF-mediated chromatin looping induces the formation of transcriptional condensates
544 and phase separation between the inner and outer regions of the loops, resulting in
545 insulator function [39-41]. Therefore, structures that result in phase separation, rather than
546 functional differences in DBPs, may be responsible for insulator functions. CEBPA in
547 the identified insulator-associated DNA-binding proteins was also reported to be
548 involved in transcriptional condensates and phase separation [42]. FOXA1 pioneer factor
549 functions as an initial chromatin-binding and chromatin-remodeling factor and has been
550 reported to form biomolecular condensates [43].

551 Other types of chromatin regulation are also expected to be related to the
552 structural interactions of molecules. As the boundary-pairing (insulator-pairing) model,
553 molecules bound to insulators physically pair with their partners, either head-to-head or
554 head-to-tail, with different degrees of specificity at the termini of TADs in flies (Fig. 7).
555 Although the experiments do not reveal how partners find each other, the mechanism
556 unlikely requires loop extrusion, as presumed in CTCF and cohesin-mediated chromatin

557 looping. Homologous and heterologous insulator-insulator pairing interactions are
558 central to the architectural functions of insulators. The manner of insulator-insulator
559 interactions is orientation-dependent [44-46].

560 Regional micro-capture (Micro-C) analyses revealed that microcompartment
561 interactions are largely independent of cohesin and transcription [5]. The depletion of
562 cohesin causes a subtle reduction in longer-range enhancer-promoter interactions and
563 that CTCF depletion can cause rewiring of regulatory contacts [47]. Another group
564 reported that enhancer-promoter interactions and transcription are largely maintained
565 upon depletion of CTCF, cohesin, WAPL or YY1. Instead, cohesin depletion decreased
566 transcription factor binding to chromatin. Thus, cohesin may allow transcription factors
567 to find and bind their targets more efficiently [48]. Furthermore, the loop extrusion is not
568 essential for enhancer-promoter interactions, but contributes to their robustness and
569 specificity and to precise regulation of gene expression [47]. The 'loop extrusion'
570 hypothesis is motivated by in vitro observations. The experiment in yeast, in which
571 cohesin variants that are unable to extrude DNA loops but retain the ability to
572 topologically entrap DNA, suggested that in vivo chromatin loops are formed
573 independently of loop extrusion. Instead, transcription promotes loop formation and acts
574 as an extrinsic motor that extends these loops and defines their final positions [49]. Maz
575 and MyoD1 among the identified insulator-associated DNA-binding proteins form loops
576 without CTCF [4,32,50]. Furthermore, protein post-transcriptional modifications (PTMs)

577 facilitate control over the molecular valency and strength of protein-protein interactions
578 [51]. O-GlcNAcylation as a PTM inhibits CTCF binding to chromatin [52]. We found that
579 the identified insulator-associated DNA-binding proteins tend to form a cluster at
580 potential insulator sites (Supplementary Fig. 3d). These proteins may interact and
581 actively regulate chromatin interactions, transcriptional condensation, and phase
582 separation through PTMs. Our analyses will contribute to reveal novel functional and
583 structural roles of DBPs for insulator functions, chromatin regulation, and phase
584 separation and to advance their experimental analysis.

585 High-resolution chromatin interaction data from a Micro-C assay indicated that
586 most of the predicted insulator-associated DBPs showed the peak of the
587 DNA-binding-site distribution around chromatin interaction sites, suggesting that these
588 DBPs are involved in chromatin interactions and that the chromatin interaction data has
589 a high degree of resolution. Base pair resolution was reported using Micro Capture-C [31].
590 However, due to the high resolution of a Micro-C assay, it is difficult to obtain all
591 chromatin interaction data at a reasonable and available cost, while Micro Capture-C
592 and Region Capture Micro-C assays focus on a specific genomic region to increase the
593 coverage of chromatin interactions. The coverage of chromatin interactions will change
594 according to the experimental data of a Micro-C assay. In addition, chromatin
595 interactions include other types of interactions that are not associated with EPIs and
596 insulator functions [53-55]. Therefore, we analyzed and predicted insulator-associated

597 DBPs without using chromatin interaction data. Our deep learning analysis successfully
598 learned EPIs that affect gene expression levels and accurately predicted the percent
599 differences in the FR, FF, RR, and RF orientations of CTCF DNA-binding sites at
600 chromatin interaction sites.

601 When comparing insulator sites separated by active and repressive domains with
602 boundary sites identified by chromatin interaction data, we identified insulator sites
603 based on the total signal levels of H3K27me3 histone modification marks and the
604 expression level of transcripts in a region. Some insulator sites may not have smooth
605 continuous distributions of histone marks and expression levels, which may cause the
606 difference between the insulator and boundary sites. Another reason for the difference
607 may be that boundary sites are more closely associated with topologically associated
608 domains (TADs) of chromosome than are insulator sites [29]. Boundary sites are regions
609 identified based on the separation of numerous chromatin interactions. On the other
610 hand, we found that the multiple DNA-binding sites of insulator-associated
611 DNA-binding proteins were located close to each other at insulator sites and were
612 associated with distinct nested and focal chromatin interactions, as reported by Micro-C
613 assay [5]. These interactions may be transient and relatively weak, such as tissue/cell type,
614 conditional or lineage-specific interactions [56].

615 AI (Artificial Intelligence) is considered as a black box, since the reason and
616 cause of prediction are difficult to know. To solve this issue, tools and methods have

been developed to know the reason and cause. These technologies are called Explainable AI. DeepLIFT is considered to be a tool for Explainable AI. However, DeepLIFT does not answer the reason and cause for a prediction. It calculates scores representing the contribution of the input data to the prediction. The analysis of many contribution scores (i.e., DeepLIFT scores in this study) for directional DNA-binding sites of putative insulator-associated DBPs in the prediction of transcript expression levels successfully predicted the percent differences in the FR, FF, RR, and RF orientations of CTCF DNA-binding sites at chromatin interaction sites based on the contribution scores of DNA-binding sites in putative enhancer and open chromatin regions. Some contribution scores may not be accurate and thus the prediction of insulator-associated DBPs or DNA-binding sites based on a small number of contribution scores may be erroneous. To compare contribution scores computed based on different pairs of input and output data with our deep learning model, we normalized contribution scores for each pair of input and output data as previously reported [57]. The analysis of many contribution scores will be used to estimate the effects of other molecules and functions involved in gene regulation.

One of the next computational challenges will be to predict and narrow the potential set of insulator-associated DNA-binding sites of multifunctional DBPs. To confirm the machine learning prediction and to identify insulator-associated DNA-binding sites in the genome for further experimental study, information on histone

637 modifications such as H3K27me3 and transcript levels will be beneficial, and other
638 epigenomic data can be used for this purpose. As a reference for an application of this
639 study, Liu et al. focused on CTCF-binding sites and performed various assays in human
640 and mouse cell cultures and in mice with the goal of applying these sites in gene therapy
641 [58]. The analysis of insulator-associated DBPs may contribute to the development of
642 technologies to regulate gene expression. In conclusion, our novel computational
643 method discovered insulator-associated DBPs, including known DBPs, and revealed
644 that the directional bias of insulator-associated DBPs is potentially involved in
645 chromatin looping, phase separation, transcriptional condensates and insulator functions
646 through the structural interactions of the DBPs. This will lead to a general and global
647 understanding of transcriptional regulation and gene expression.
648

649 # Methods

650 **Prediction of gene expression levels based on DNA-binding sites**

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

652 To predict DNA-binding sites that also act as insulator sites in the human genome,
653 we modified a previously developed deep learning method designed to predict the
654 transcript expression levels of genes in tissues and cell types based on the DNA-binding
655 sites of DBPs in promoter regions and other DNA regions (Fig. 2a; Fig. S4) [14]. To
656 reduce the influence of cellular heterogeneity observed when analyzing tissues, we used

657 the gene expression data from four distinct cell types where various experimental data
658 are available in public databases: HFFs (GSM941744), monocytes (Blueprint RNA-seq
659 FPKM [fragments per kilobase of exon per million reads mapped] data
660 'C0010KB1.transcript_quantification.rsem_grape2_crg.GRCh38.20150622.results' in
661 the Blueprint DCC portal at http://dcc.blueprint-epigenome.eu/#/files), human
662 mammary epithelial cells (HMECs) (Encyclopedia of DNA Elements at UCSC
663 http://genome.ucsc.edu/encode/,
664 'wgEncodeCshlLongRnaSeqHmecCellPapAlnRep1.bam'), and neural progenitor cells
665 (NPCs) (GSM915326, ENCFF529SIO). For gene expression data, FPKM values
666 representing the expression levels of genes were retrieved or calculated based on
667 GENCODE V19 gene annotations in the ENCODE portal database [59]
668 (https://www.encodeproject.org/) using the RSeQC tool [60] (http://rseqc.sourceforge.net/).
669 We calculated log2-fold-transformed FPKM values (adding 0.25 to prevent results of
670 negative infinity), median log2-FPKM values across the four cell types, and log2-fold
671 changes relative to the median for all cell types according to the original deep learning
672 method (transcriptome data processing as described by Tasaki et al.) [14].
673

674 **DNA-binding sites as input data for machine learning analysis**

675 To investigate the effect of insulator sites on EPIs and gene expression, the
676 DNA-binding sites in putative enhancer and open chromatin regions located away from

677 promoter regions were added as input data in the deep learning method. To identify the
678 DNA-binding sites of DBPs in promoter and enhancer regions, ChIP-seq data of the
679 DBPs were used and were updated to GTRD database release v20.06
680 (https://gtrd.biouml.org) [61]. The binding site of a DBP regulating the expression level of
681 each transcript was selected using eQTL data (GTEx_Analysis_v8_eQTL.tar) from
682 GTEx portal release v8 (https://gtexportal.org). If a DNA-binding site was located
683 within a 100-bp region around a single-nucleotide polymorphism (SNP) of an eQTL,
684 we assumed that the DBP regulated the expression of the transcript corresponding to the
685 eQTL. If the SNP of an eQTL was not located on the same chromosome as the
686 transcribed region of the eQTL (i.e., *trans*-eQTL), the eQTL data were not used. The
687 genomic positions of the eQTLs were converted to the corresponding positions in the
688 hg19 version of the human genome using the LiftOver tool
689 (http://hgdownload.soe.ucsc.edu/downloads.html#liftover).

690 The original input data of DNA-binding sites in promoter regions used for the
691 deep learning analysis consisted of the lengths of a DBP ChIP-seq peak in each
692 100-base bin in a region between 2 kbp upstream and 1 kbp downstream of the
693 transcription start site (TSS) of each gene or transcript [14]. To add the input data of
694 DNA-binding sites of DBPs in more upstream and downstream regions of a TSS and
695 reduce the number of bins, we first computed the length of a DBP ChIP-seq peak in a
696 100-base bin and then selected the maximum length in each interval of 100 bins in a

697 region between 1 Mbp upstream and downstream of a TSS. This step outputted the 200
698 longest DBP ChIP-seq peaks in the examined regions.
699
700 **Training step for machine learning analysis**
701 In the training of the modified deep learning model, we used the same test,
702 training, and validation data as in the original method: all 2,705 genes encoded on
703 chromosome 1 as test data, training data (22,251 genes), and validation data (2,472
704 genes) randomly selected from the other genes [14]. The input data for promoters and
705 putative enhancer regions were concatenated into a total of 230 bins of the lengths of
706 1,310 DBP ChIP-seq peaks, consisting of 30 bins for promoters and 200 bins for
707 putative enhancers because the DeepLIFT tool 'RevealCancel rule on dense layers and
708 the Rescale rule on conv layers (genomics default)', with high scoring accuracy, [24]
709 cannot be applied to a neural network with multiple input layers (Fig. S5). Ubuntu
710 (version 20.04), Python (version 3.8.10), TensorFlow (version 2.6.0), NVIDIA CUDA
711 toolkit (version 11.4.2), GeForce RTX 3080 GPU with 10 GB memory, and Intel Core
712 i7-10700 CPU with 128 GB memory were used for training and prediction, and the
713 original scripts were modified to run on newer versions of the software. The training
714 method was the same as the original, and the training time was two to four hours. In this
715 study, we focused on the analysis of HFF cells and assessed the training process using
716 all the input data as the test data, which slightly improved the prediction of gene

717 expression. Therefore, we used the trained model for the analysis of insulator-associated
718 DBPs.
719
720 **Prediction of insulator-associated DBPs affecting EPIs**
721 **DNA-binding sites of putative insulator-associated DBPs**
722 To identify the direction of DBP-binding motifs in the DNA-binding sites of
723 predicted insulator-associated DBPs, we searched for DBP-binding motifs in open
724 chromatin regions in HFF cells (GSE170964) using the PIQ tool [17] and the position
725 frequency matrices of the DBP-binding motifs. The motif sequences were collected
726 from databases and previous studies, specifically, TRANSFAC (2019.2), JASPAR
727 (2018), UniPROBE (2018), DNA motifs from high-throughput SELEX, DNA motifs
728 from ENCODE DBP ChIP-seq data, and HOCOMOCO versions 9 and 11 [18-23]. To
729 reduce the number of false-positive prediction of DNA-binding sites, only
730 DNA-binding sites that overlapped the ChIP-seq peaks of the same DBP (GTRD
731 database release v20.06 [https://gtrd.biouml.org]) were used for further analyses. The
732 genomic positions of ChIP-seq and DNase-seq data in public databases were converted
733 to those provided in the hg19 version of the human genome using the LiftOver tool.
734 To confirm that CTCF DNA-binding sites were located at chromatin interaction
735 sites, we examined the overlap between CTCF ChIP-seq peaks and the sites of *in situ*
736 ChIA-PET chromatin interactions mediated by CTCF in human foreskin fibroblast

737 (HFF) cells (4D Nucleome data portal,
738 https://data.4dnucleome.org/files-processed/4DNFI73AFN12/). To eliminate
739 DNA-binding sites that overlapped EPI, we obtained ChIP-seq peaks of histone
740 modifications, such as H3K4me3, H3K27ac, and H3K4me1, associated with promoter
741 and enhancer regions in HFF cells (4D Nucleome data portal, 4DNFIJOK1BWR,
742 4DNFI442BQLO, and 4DNFIAGV3X99).
743
744 **Preparation of input data considering the DNA-binding sites of an**
745 **insulator-associated DBP**
746 To prepare the input data of the DNA-binding sites of DBPs in enhancers and
747 promoters for the deep learning analysis, first, a transcript regulated by a DBP bound to
748 enhancers and promoters was predicted based on the single nearest gene association rule
749 (Fig. 2c; McLean et al. [15] Fig. S2c), as we found that the association rule predicted EPIs
750 more accurately than the other rules in our previous study [15]. The maximum length of
751 the regulatory domain for each transcript was set to 1 Mbp upstream and downstream of
752 the TSS. Then, considering the genomic locations and effects of the insulator sites of a
753 DBP on EPIs, the length of a DBP ChIP-seq peak in a 100-base bin was calculated from
754 a TSS to the DNA-binding site of a predicted insulator-associated DBP. To reduce the
755 computational time of the analysis, we first calculated the length of a DBP ChIP-seq
756 peak in the regulatory domains of all transcripts in a 100-base bin. Then, for each

757 interval of 100 bins between a TSS and a DNA-binding site of a predicted
758 insulator-associated DBP in the regulatory domain, we computed the maximum length
759 of each DBP ChIP-seq peak in 100 bins on the basis of the precomputed length of the
760 DBP ChIP-seq peak in each bin. In the case of a set of 100 bins containing the
761 DNA-binding site of a predicted insulator-associated DBP, we computed the maximum
762 lengths from the bin closest to a TSS to the first bin containing the DNA-binding site of
763 the insulator-associated DBP, taking into account the orientation of the DNA-binding
764 site.
765
766 **Prediction of insulator-associated DBPs using DeepLIFT tool**
767 The DeepLIFT tool (version 0.6.13.0) was used to calculate the contribution
768 (DeepLIFT score) of each DNA-binding site to the prediction of the expression level of
769 a transcript [24]. The contribution scores of DNA-binding sites of DBPs in putative
770 enhancers and promoters upstream and downstream of each transcript were normalized
771 by the sum of the absolute values of all contribution scores for the transcript, as the
772 range of contribution scores changes according to the range of the predicted output (i.e.,
773 the expression level of a transcript), as described in a previous study [57]. To estimate the
774 effect of the DNA-binding sites of a predicted insulator-associated DBP on the
775 prediction of transcript levels, positive normalized contribution scores for transcripts
776 with DNA-binding sites in upstream or downstream regions were subjected to

777 comparisons between all combinations of two different orientations (i.e., all
778 combinations of two orientations among FR, RF, forward-forward [FF], and
779 reverse-reverse [RR] orientations, and no orientation [X]: FR vs. RF, FR vs. X, RF vs.
780 X, FF vs. RR, FF vs. X, RR vs. X, FF vs. FR, FF vs. RF, RR vs. FR, and RR vs. RF)
781 among the DNA-binding sites of a predicted insulator-associated DBP by using a
782 statistical test (Mann–Whitney *U* test, FDR-adjusted $p$ value < 0.05 and $p$ value < $10^{-3}$).
783 For EPIs that were common to the predictions based on two different orientations of
784 DNA-binding sites of a predicted insulator-associated DBP, the DNA-binding sites of
785 DBPs in putative enhancers and promoters among the common EPIs were omitted in
786 the comparison of positive normalized contribution scores between the two different
787 orientations. Based on the results of the statistical analysis and the mean differences in
788 contribution scores, we predicted an insulator-associated DBP on the basis of the
789 directional bias or lack of orientation bias of DNA-binding sites; for example, to predict
790 an insulator-associated DBP with a DNA-binding site in the FR orientation, we selected
791 a DBP for which the contribution scores among all combinations of FR vs. RF, FR vs.
792 FF, FR vs. RR, and FR vs. X were significant (Mann–Whitney *U* test, FDR-adjusted $p$
793 value < 0.05 and $p$ value < $10^{-3}$) and showed positive (high) mean differences among
794 the contribution scores of all the combinations.
795
796 **Search for insulator-associated DNA-binding sites based on histone modifications**

797 **and gene expression levels**

798 The genomic positions in the BAM files of H3K27me3 and H3K9me3 ChIP-seq
799 data (ENCFF267JBO, ENCFF528UZJ) from HFF cells were converted to the
800 corresponding positions in the hg19 version of the human genome using the CrossMap
801 tool [62] and then converted to WIG files using bam2wig.pl from Bio::ToolBox
802 (https://github.com/tjparnell/biotoolbox) to visualize the data on the UCSC Genome
803 Browser through custom tracks (http://www.genome.ucsc.edu/cgi-bin/hgCustom) and a
804 track hub (https://genome.ucsc.edu/cgi-bin/hgHubConnect). The BAM file of RNA-seq
805 data (GSM941744) of HFF cells was converted to a WIG file using bam2wig.pl in
806 Bio::ToolBox. The genomic positions in the BED file of the boundaries and the
807 distribution of insulation scores identified through a Micro-C assay (4D Nucleome data
808 portal, https://data.4dnucleome.org/files-processed/4DNFI5OZ7TBE/ and
809 4DNFIPMHBA41) were converted to the corresponding positions in the hg19 version
810 of the human genome.

811

812 **Distribution of DNA-binding sites around chromatin interaction sites**

813 We examined the distribution of the DNA-binding sites of predicted
814 insulator-associated DBPs in regions within ±500 bp of the center of a chromatin
815 interaction site using chromatin interaction data from HFF cells obtained via a Micro-C
816 assay (4D Nucleome data portal,

817 https://data.4dnucleome.org/files-processed/4DNFI1MCF9GZ/). For the
818 insulator-associated DBPs other than CTCF, RAD21, and SMC3, the DNA-binding
819 sites that do not overlap with those of CTCF, RND21, and SMC3 were used to examine
820 their distribution around interaction sites. The number of DNA-binding sites of
821 predicted insulator-associated DBPs that cooccurred with DNA-binding sites of known
822 insulator-associated DBPs, such as CTCF, RAD21, and SMC3, was counted in each
823 10-bp bin in the regions within ±500 bp of the center of a chromatin interaction site.
824
825 **Insulator-associated DNA-binding sites near splice sites**
826 Insulator-associated DNA-binding sites were searched within 200 bp of the center
827 of splice sites using the file of the genomic positions of alternative transcripts in the
828 UCSC Genome Browser
829 (https://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/all_mrna.txt.gz ;
830 https://genome.ucsc.edu/cgi-bin/hgTables?db=hg19&hgta_group=rna&hgta_track=mrn
831 a&hgta_table=all_mrna&hgta_doSchema=describe+table+schema). We performed the
832 statistical test to estimate the enrichment of insulator-associated DNA-binding proteins
833 at the splicing sites compared with the other DNA-binding proteins (Mann–Whitney *U*
834 test, $p$ value $< 0.05$).
835
836 **Statistical analysis**

837 Comparison of the distribution of DeepLIFT scores between different orientations
838 of DNA-binding sites was performed using the Mann-Whitney *U* test with the
839 scipy.stats.mannwhitneyu library. Two-tailed *p* values were calculated. For multiple
840 comparisons, adjusted *p* values were calculated by false discovery rate (FDR) using the
841 Benjamini-Hochberg procedure. *Z* scores for the Mann-Whitney *U* test were not output
842 by the library. Therefore, we calculated *z* scores by z-standardizing the *U* values. The
843 code is made available on GitHub at https://github.com/reposit2/insulator/tree/main.
844

845 **Data availability**

846 Human foreskin fibroblast (HFF) cell RNA-seq data were obtained from GSM941744.
847 Human monocyte RNA-seq data were obtained from the Blueprint DCC portal
848 http://dcc.blueprint-epigenome.eu/#/files at
849 'C0010KB1.transcript_quantification.rsem_grape2_crg.GRCh38.20150622.results'. The
850 RNA-seq data of human mammary epithelial cells (HMEC) were found on ENCODE at
851 UCSC http://genome.ucsc.edu/encode/ at
852 'wgEncodeCshlLongRnaSeqHmecCellPapAlnRep1.bam'. Human neural progenitor cell
853 (NPC) RNA-seq data were obtained from GSM915326. The DNase-seq data of HFF
854 cells were obtained from GSE170964. The *in situ* ChIA-PET chromatin interaction data
855 mediated by CTCF in HFF cells were available at the 4D Nucleome data portal,
856 https://data.4dnucleome.org/files-processed/4DNFI73AFN12/. The ChIP-seq data of

857 H3K4me3, H3K27ac, and H3K4me1 of HFF cells were found at the 4D Nucleome data
858 portal, 4DNFIJOK1BWR, 4DNFI442BQLO, and 4DNFIAGV3X99. The ChIP-seq data
859 of H3K27me3 and H3K9me3 in HFF cells were obtained from the ENCODE portal,
860 ENCFF267JBO and ENCFF528UZJ. The chromatin interaction data and boundary data
861 of HFF cells by the Micro-C were obtained from the 4D Nucleome data portal,
862 https://data.4dnucleome.org/files-processed/4DNFI1MCF9GZ/ and
863 https://data.4dnucleome.org/files-processed/4DNFI5OZ7TBE/. ChIP-seq data of TFs
864 were downloaded from the v20.06 release of the GTRD database
865 (https://gtrd.biouml.org). eQTL data (GTEx_Analysis_v8_eQTL.tar) were downloaded
866 from the v8 release of the GTEx portal (https://gtexportal.org). DNA binding motifs of
867 TFs were collected from TRANSFAC (2019.2), JASPAR (2018)
868 (https://jaspar.genereg.net/), UniPROBE (2018)
869 (http://the_brain.bwh.harvard.edu/uniprobe/), high-throughput SELEX
870 (https://doi.org/10.1016/j.cell.2012.12.009), transcription factor binding sequences of
871 ENCODE TF ChIP-seq data (http://compbio.mit.edu/encode-motifs/), and
872 HOCOMOCO version 9 and 11 (https://autosome.org/HOCOMOCOS/)
873 (https://hocomoco11.autosome.org/). The genomic positions of alternative transcripts in
874 the UCSC Genome Browser
875 (https://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/all_mrna.txt.gz).
876

877 **Code availability**

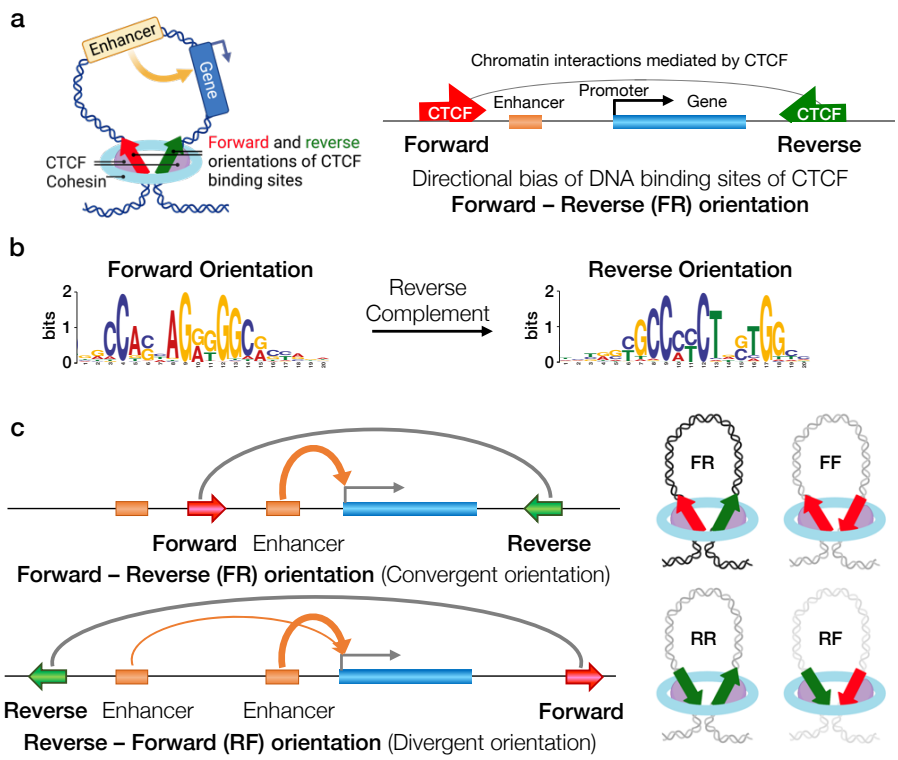
878 Demonstration codes are made available at

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

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

881

882 **Figures**




883

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

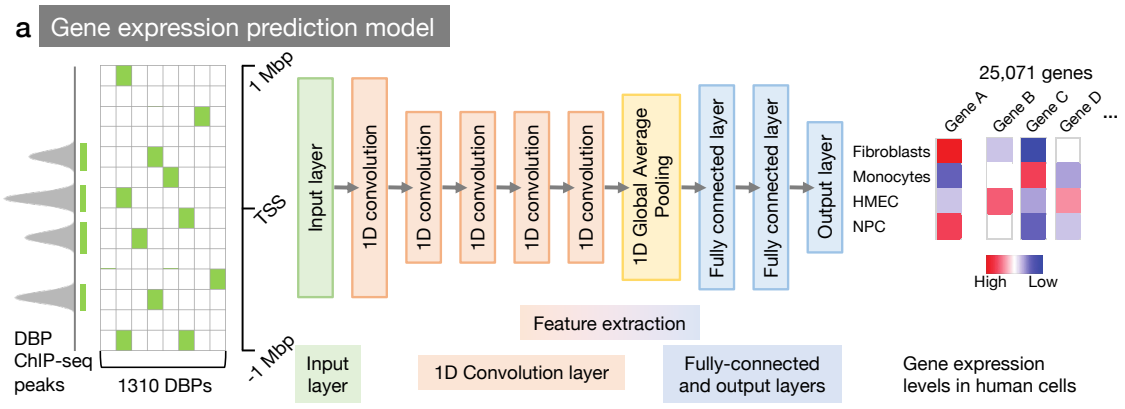
885 CTCF DNA-binding sites with a forward–reverse orientation are frequently found at

886 chromatin interaction sites. The forward and reverse orientations of DNA-binding sites

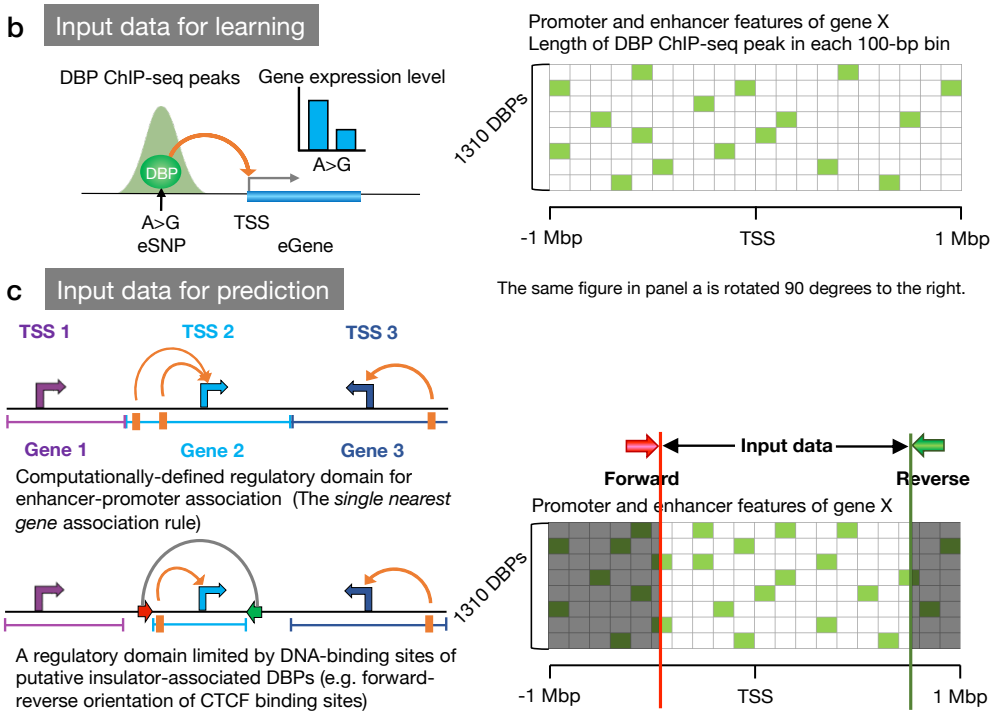
887 have complementary sequences and are relatively distinct in orientation according to

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 DNA-binding sites of a putative insulator-associated DBP (e.g., forward–reverse orientation of the DNA-binding sites of CTCF in the figure). The numbers of DNA-binding sites 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.

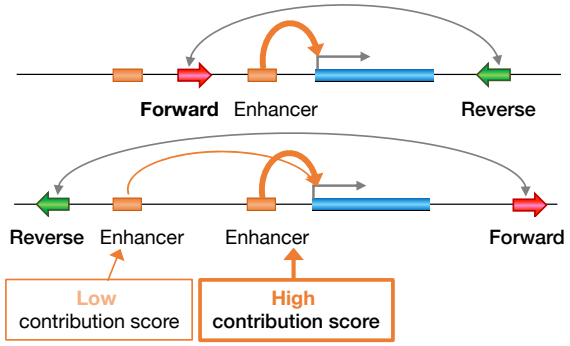
905


a Gene expression prediction model
DBP ChIP-seq peaks
1310 DBPs
1 Mbp
TSS
-1 Mbp
Input layer
1D convolution
1D convolution
1D convolution
1D convolution
1D convolution
1D Global Average Pooling
Fully connected layer
Fully connected layer
Output layer
Fibroblasts
Monocytes
HMEC
NPC
25,071 genes
Gene A
Gene B
Gene C
Gene D
...
High
Low
Feature extraction
Input layer
1D Convolution layer
Fully-connected and output layers
Gene expression levels in human cells

906

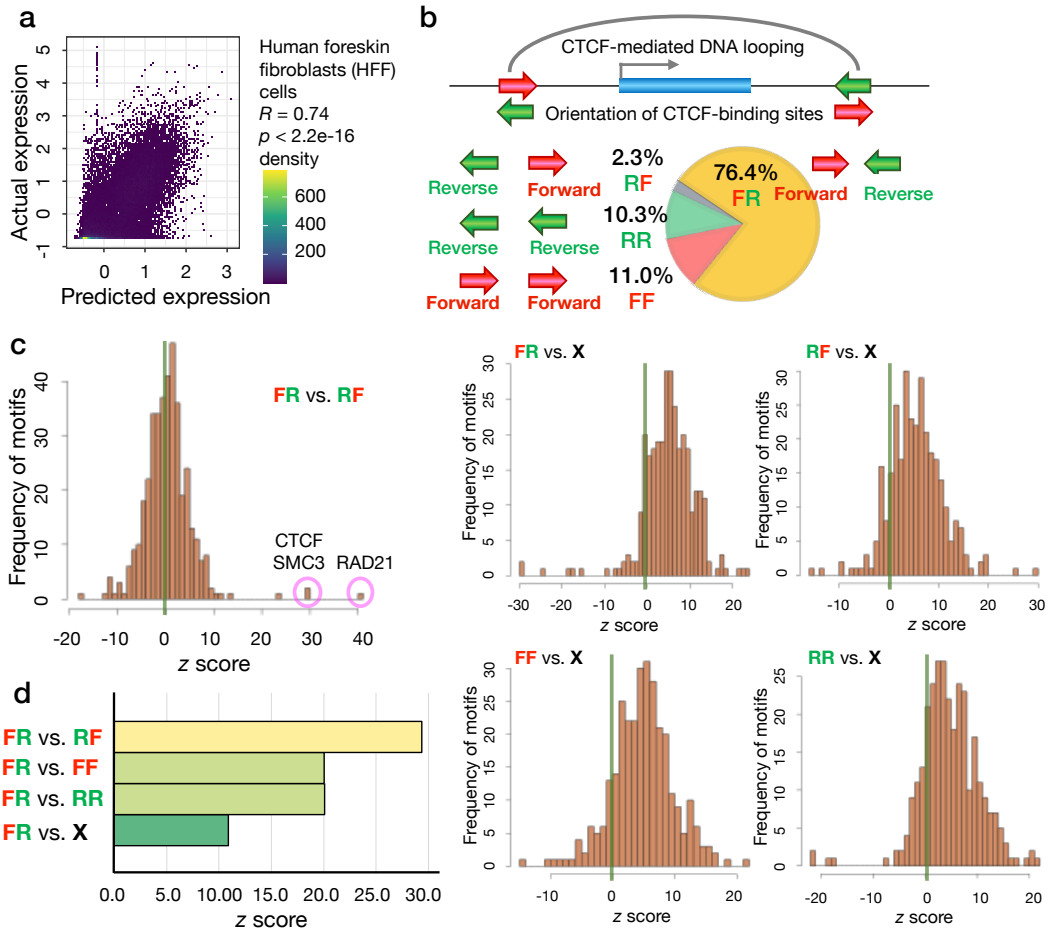

b Input data for learning
DBP ChIP-seq peaks
Gene expression level
DBP
A>G
A>G eSNP
TSS
eGene
Promoter and enhancer features of gene X
Length of DBP ChIP-seq peak in each 100-bp bin
1310 DBPs
-1 Mbp
TSS
1 Mbp
The same figure in panel a is rotated 90 degrees to the right.
c Input data for prediction
TSS 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 DNA-binding sites 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
1310 DBPs
-1 Mbp
TSS
1 Mbp

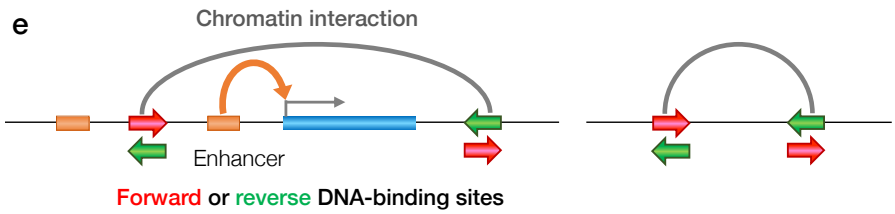
907


d Estimation of importance of input data for prediction
Forward
Enhancer
Reverse
Reverse
Enhancer
Enhancer
Forward
Low contribution score
High contribution score

908 **Fig. 2. Overview of the prediction of insulator-associated DBPs. a** We used a
909 machine learning method named DEcode, with some modifications. The output data of
910 gene expression levels in human cells were predicted from the input data of genomic
911 locations of DBP binding at promoters and putative enhancers. **b** In the machine
912 learning phase, a DBP binding to a putative enhancer overlapping an eSNP (single
913 nucleotide polymorphism) was associated with a gene (eGene) that was potentially
914 regulated by the DBP based on eSNP and eGene of eQTL data. Based on the association
915 of each gene, the locations of DBP ChIP-seq peaks were binned (green box) and
916 summarized in a matrix as input data: Promoter consists of 30 bins of 100 bp between
917 -2 kbp and 1 kbp region from TSS. Enhancer-gene regulatory interaction regions consist
918 of 200 bins of 10 kbp between -1 Mbp and 1 Mbp region from TSS, not including
919 promoter. **c** In the phase of putative insulator-associated DBP prediction, a DBP binding
920 to a putative enhancer was associated with a gene potentially regulated by the DBP
921 based on a computationally defined regulatory domain without using eQTL data. The
922 *single nearest gene* association rule defines a regulatory domain connecting the
923 midpoints between the TSS of a gene and the TSS of the nearest gene in the upstream
924 and downstream regions. The TSS and regulatory domain of each gene are indicated
925 with a purple, light blue, or blue arrow and a bracketed line, respectively. The regulatory
926 domain of enhancer–promoter association was restricted by the DNA-binding sites (red
927 and green arrows) of a putative insulator-associated DBP in this study. Based on the

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






936

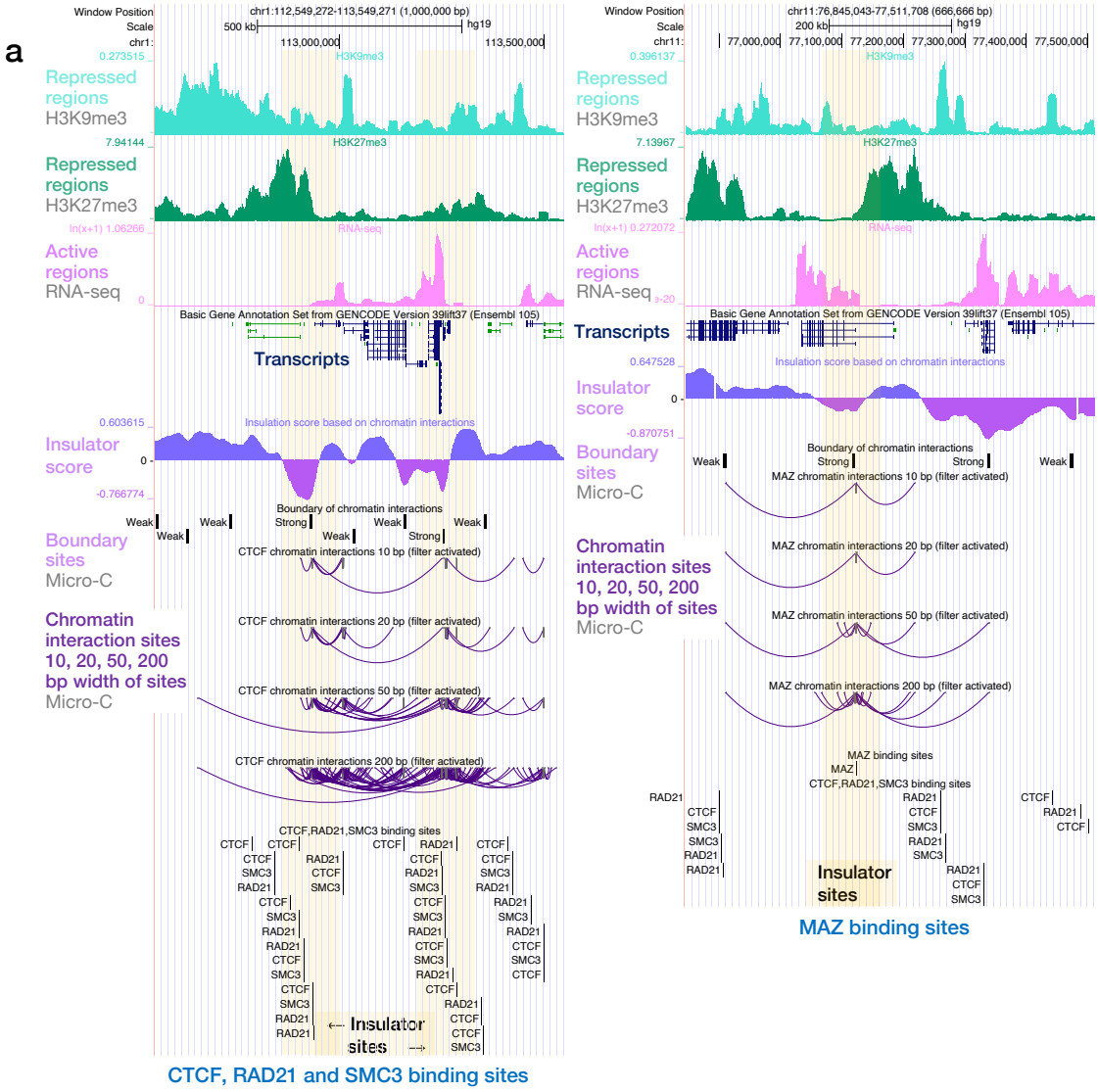
937 **Fig. 3. Estimation of the effect of directional regulatory motifs on gene expression.**
938 **a** The predicted log2-TPM (transcripts per million) values of genes were compared with
939 the actual expression levels in human foreskin fibroblast (HFF) cells using Spearman's
940 rank correlation. **b** An FR orientation of CTCF-binding sites is common (76.4%) at
941 chromatin interaction sites. Forward-forward (FF), reverse-reverse (RR), and RF
942 orientations occur at low frequencies (11.0%, 10.3%, and 2.3%, respectively) in
943 chromatin interaction sites (Guo Y et al. *Cell* 2015) [3]. **c** Distribution of *z* scores from the
944 Mann–Whitney *U* test for DeepLIFT scores between forward–reverse (FR) and
945 reverse–forward (RF) orientations of DNA-binding sites of proteins. Distribution of *z*
946 scores between no orientation (X) and another orientation (FR, RF, FF, or RR) is shown
947 in the right panel. The graphs show the results based on Selection Criteria S2 in
948 Supplementary Fig. 1 and the Extended Data. **d** *Z* scores from the Mann–Whitney *U* test
949 for DeepLIFT scores in pairwise comparisons of CTCF-binding sites with four different
950 orientations (FR, RF, FF, and RR) and no orientation (X, i.e., orientation is not
951 distinguished). The differences in *z* scores followed the same trend as the percentages of
952 CTCF-binding site orientations in chromatin interaction sites. **e** The DNA-binding sites

953 of insulator-associated proteins were found at a pair of long-range chromatin interaction

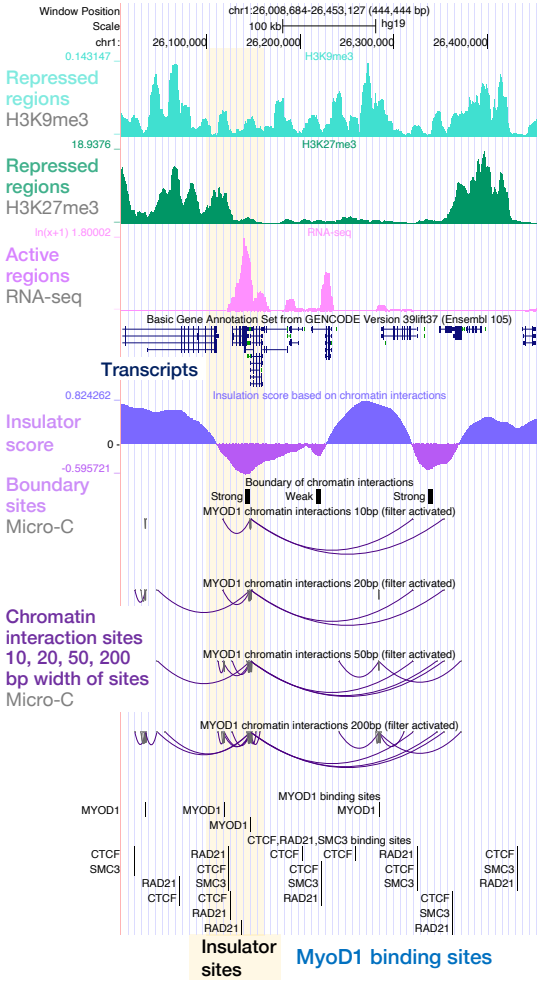
954 sites or at either site.

955

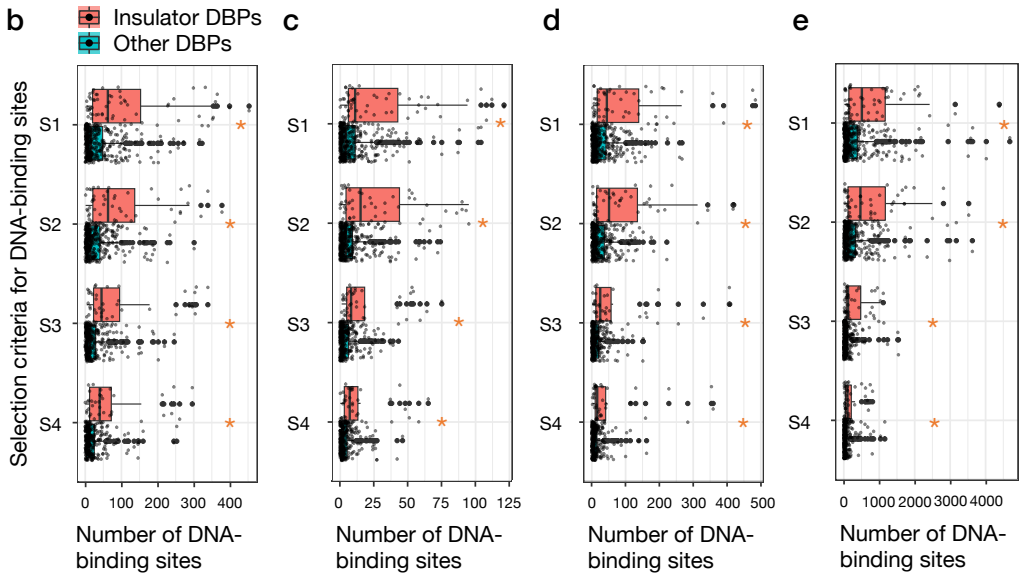
956




957


Window Position
chr1:26,008,684-26,453,127 (444,444 bp)
Scale
100 kb
hg19
chr1:
26,100,000
26,200,000
26,300,000
26,400,000
0.143147
H3K9me3
Repressed regions H3K9me3
18.9376
H3K27me3
Repressed regions H3K27me3
ln(x+1) 1.80002
RNA-seq
Active regions RNA-seq
Basic Gene Annotation Set from GENCODE Version 39lift37 (Ensembl 105)
Transcripts
0.824262
Insulation score based on chromatin interactions
Insulator score
0
-0.595721
Boundary sites Micro-C
Boundary of chromatin interactions
Strong
Weak
Strong
MYOD1 chromatin interactions 10bp (filter activated)
MYOD1 chromatin interactions 20bp (filter activated)
Chromatin interaction sites 10, 20, 50, 200 bp width of sites Micro-C
MYOD1 chromatin interactions 50bp (filter activated)
MYOD1 chromatin interactions 200bp (filter activated)
MYOD1 binding sites
MYOD1
CTCF,RAD21,SMC3 binding sites
CTCF
SMC3
RAD21
Insulator sites
MyoD1 binding sites

958



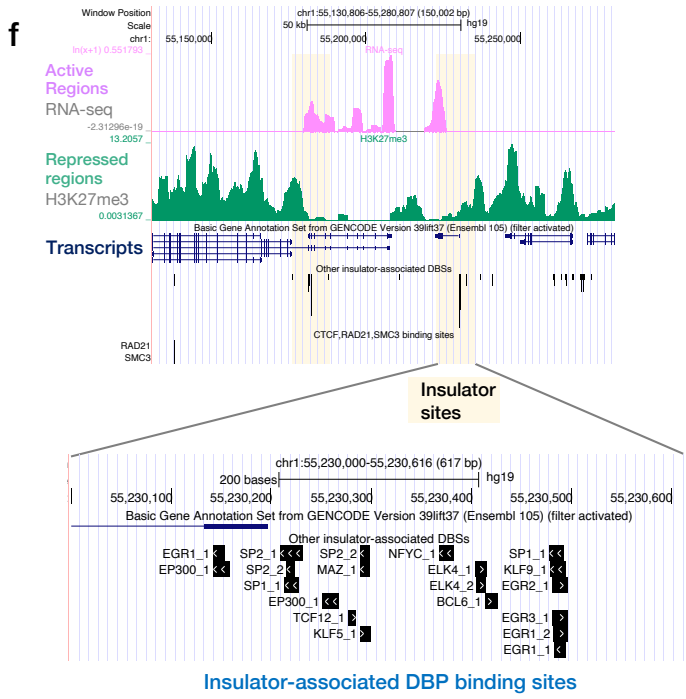



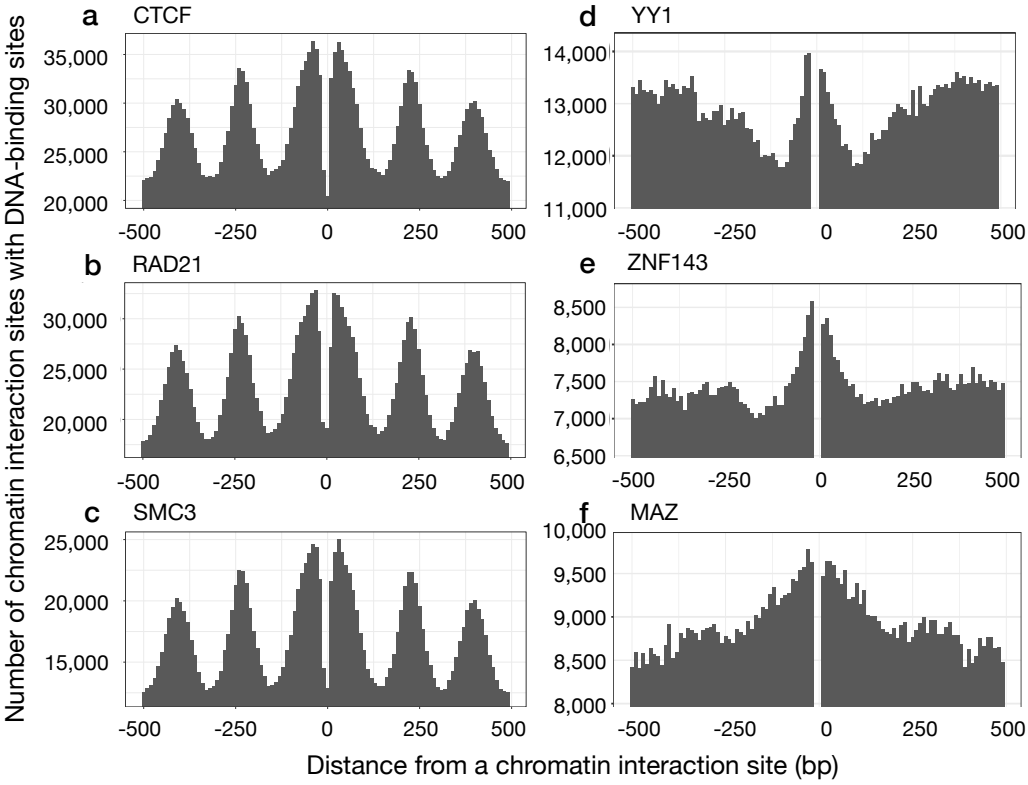

**Fig. 4. DNA-binding sites of predicted insulator-associated DBPs located at potential insulator sites. a** Distribution of histone modification marks H3K27me3 (green color) and H3K9me3 (turquoise color) and transcript levels (pink color) in the

964 upstream and downstream regions of a potential insulator site (light orange color).
965 These histone modification marks are observed in transcriptionally repressed regions.
966 Insulator sites would be located at the boundary between transcriptionally active and
967 repressed regions (light orange vertical bars). Blue and green lines in the GENCODE
968 track represent transcripts of genes. Insulator score and boundary sites were estimated
969 from chromatin interactions, which are available in 4D Nucleome Data Portal. Dark
970 blue arcs show chromatin interactions within a 10 bp, 20 bp, 50 bp, and 200 bp of the
971 binding sites of CTCF (upper left panel), MAZ (upper right panel), and MyoD1 (lower
972 left panel), respectively. The binding sites of CTCF, RAD21, SMC3 (upper left panel),
973 MAZ (upper right panel), and MyoD1 (lower left panel) are shown as black bars. Homo
974 sapiens (human) genome assembly GRCh37 (hg19) from Genome Reference
975 Consortium was used in panels **a**. **b** Number of DNA-binding sites of predicted
976 insulator-associated DBPs, where H3K27me3 signal intensities differed between the
977 upstream and downstream regions of the DNA-binding site. * $p$ value < 0.01 (Table 2).
978 Boxplots show the median as the centerline, the boxes define the interquartile range
979 (IQR), the whiskers extend up to 1.5 times the IQR, and all points represent outliers
980 from that range. All data points are displayed. **c** Number of DNA-binding sites of
981 predicted insulator-associated DBPs, where H3K27me3 signal intensities and transcript
982 levels differed between the upstream and downstream regions of the DNA-binding site.
983 **d** Number of DNA-binding sites of predicted insulator-associated DBPs in boundary

regions assigned based on chromatin interaction data from human foreskin fibroblast (HFF) cells. **e** Number of DNA-binding sites of predicted insulator-associated DBPs that overlap within 200 bp centered on splice sites. The DNA-binding sites of insulator-associated DNA-binding proteins were found within 200 bp centered on splice sites more significantly than the other DNA-binding proteins. **f** Cluster of DNA-binding sites of insulator-associated DBPs.

**Fig. 5. Insulator-associated DBPs potentially regulate alternative transcription.** Distribution of H3K27me3 histone modification mark (green color) and transcript levels (pink color) are shown. The histone modification mark is observed in transcriptionally

995 repressed regions. Insulator sites would be located at the boundary between
996 transcriptionally active and repressed regions (light orange vertical bars). Blue lines in
997 the GENCODE track represent transcripts of genes. Chromatin interactions overlapped
998 with a 200-bp region centered on the binding sites of RBPJ (left panel) and SP1 (right
999 panel) are shown as dark blue arcs, respectively. The binding sites of RBPJ (left panel),
1000 SP1 (right panel), and CTCF, RAD21, SMC3 (both panels) are shown as black bars.
1001 The DNA-binding sites of insulator-associated DBPs were frequently located near
1002 transcription start and termination sites. Insulator-associated DBPs potentially regulate
1003 transcribed regions differentially repressed in alternative transcripts.
1004




1005

1006 **Fig. 6. DNA-binding sites distributed around a chromatin interaction site. a-f** The

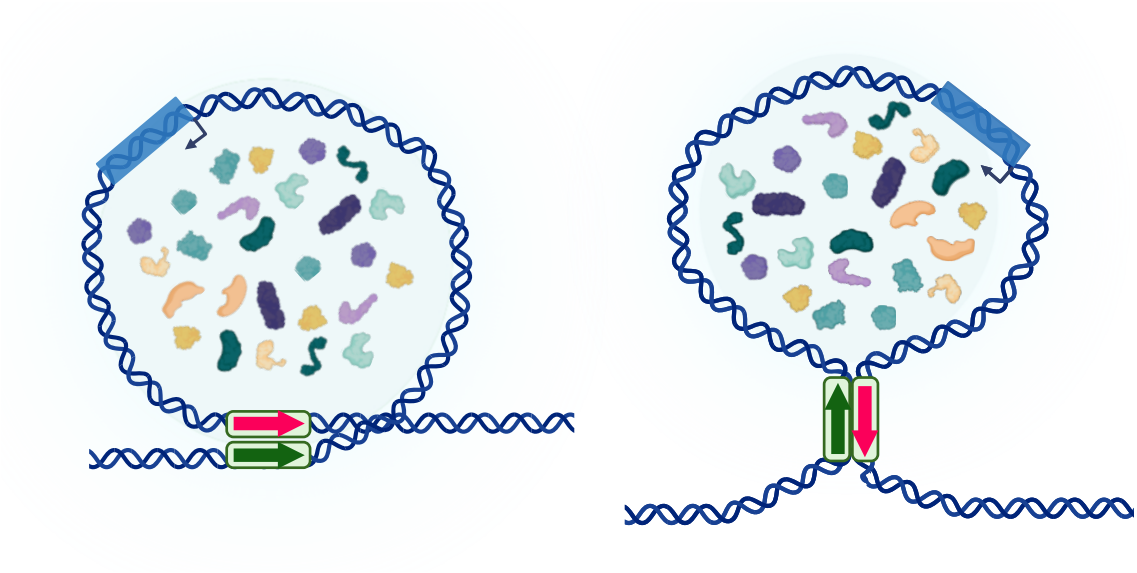
1007 X-axis shows the distance of a DNA-binding site from a chromatin interaction site. The

1008 Y-axis shows the number of chromatin interaction sites with DNA-binding sites within a

1009 ±500-bp region from the center of the interaction site. For YY1, ZNF143, and MAZ,

1010 DNA-binding sites overlapping with those of CTCF, RAD21, and SMC3 were removed.

1011



1012

1013 **Fig. 7. Boundary-pairing (insulator-pairing) model for the regulation of**

1014 **enhancer-promoter interactions.** Molecules bound to insulators physically pair with

1015 their partners, either head-to-head or head-to-tail, with different degrees of specificity at

1016 the termini of TADs in flies [44-46]. Although the experiments do not reveal how partners

1017 find each other, the mechanism unlikely requires loop extrusion. Homologous and

1018 heterologous insulator-insulator pairing interactions are central to the architectural

1019 functions of insulators. The manner of insulator-insulator interactions is

1020 orientation-dependent. Homologous and heterologous interactions between the

1021 insulator-associated DNA-binding proteins potentially occur with the same mechanism

1022 in humans. Chromatin looping could induce the formation of transcriptional

1023 condensates and phase separation as well as cohesin-CTCF-anchored chromatin looping

1024 [39-41]. The figure was created with BioRender.com.

1025

1026 **Tables**

1027 **Table 1. Predicted insulator-associated DBPs with directional or nondirectional**

1028 **DNA-binding sites.**

| 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 |

1030 *Z* scores from the Mann–Whitney *U* test for DeepLIFT scores between forward-reverse
1031 (FR) and other orientations such as reverse-forward (RF), forward-forward (FF),
1032 reverse-reverse (RR), and non-orientation (X) of DNA-binding sites of proteins. Mean
1033 differences of DeepLIFT scores. Color scales are shown for each DNA-binding protein
1034 (DBP). Top: Selection Criteria S1 in Supplementary Fig. 1; Middle: Selection Criteria
1035 S2 in Supplementary Fig. 1; Bottom: List of predicted insulator-associated DBPs. DBPs
1036 in a yellow cell are associated with insulators, CTCF, and/or other transcriptional
1037 regulations that have been shown to affect enhancer-promoter interactions (EPIs) in
1038 previous studies.
1039
1040 **Table 2. Number of DNA-binding sites of predicted insulator-associated DBPs at**
1041 **insulator sites and splice sites.**

| | | Selection Criteria | | Median nunber of DNA sites | Mean number of DNA sites | Number of DBP | *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 of interactions | 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 | |
| Overlap with splice sites | | S1 | Insulator-associated DBPs | 535 | 774.45 | 58 | 1.02E-06 |
| | | | Other DBPs | 132 | 350.11 | 354 | |
| | | S2 | Insulator-associated DBPs | 472 | 748.72 | 50 | 6.28E-06 |
| | | | Other DBPs | 122 | 290.21 | 346 | |
| | | S3 | Insulator-associated DBPs | 120 | 305.7 | 40 | 2.13E-07 |
| | | | Other DBPs | 35 | 99.23 | 298 | |
| | | S4 | Insulator-associated DBPs | 103 | 198.24 | 37 | 1.45E-05 |
| | | | Other DBPs | 30 | 81.01 | 274 | |

Selection Criteria S1: DNA-binding sites of predicted insulator-associated DBPs; Selection Criteria S2: DNA-binding sites of predicted insulator-associated DBPs, where the DNA-binding sites overlap with a chromatin interaction site in a region between more than 5-kb upstream and 5-kb downstream of a transcribed region; Selection Criteria S3: DNA-binding sites of predicted insulator-associated DBPs, where the

1048 DNA-binding sites do not overlap with the ChIP-seq peaks of H3K4me3, H3K4me1, or
1049 H3K27ac; Selection Criteria S4: DNA-binding sites of predicted insulator-associated
1050 DBPs, where the DNA-binding sites overlap a chromatin interaction site in a region
1051 between more than 5-kb upstream and 5-kb downstream of a transcript and the
1052 DNA-binding sites do not overlap the ChIP-seq peaks of H3K4me3, H3K4me1, or
1053 H3K27ac. *P* values were calculated from the Mann–Whitney *U* test.
1054
1055 **Table 3. Pairs of DNA-binding sites colocalize at the same distance from chromatin**
1056 **interaction sites.**

| DBP1 | DBP2 | Distance to start (bp) | Distance to end (bp) | # 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 |

Distance to start and end (bp) shows the genomic distance centered on a chromatin interaction site to the 10-bp bin with the highest number of DBP1 and DBP2 binding sites among 100 bins (1 kbp) around the site. The positive value indicates the distance to the downstream side of an insulator site and the negative value indicates the distance to the upstream side. # chromatin interactions with DBP1 and DBP2 sites shows the number of chromatin interaction sites with DBP1 and DBP2 binding sites within ±500

1064 bp of the interaction site. Ratio (DBP1 and 2)/DBP2 is the ratio of the number of
1065 chromatin interaction sites with both DBP1 and DBP2 binding sites to the number of
1066 interaction sites with DBP2 binding sites.
1067
1068

1069 **References**


1070 1 Weintraub, A. S. *et al.* YY1 Is a Structural Regulator of Enhancer-Promoter
1071 Loops. *Cell* **171**, 1573-1588 e1528, doi:10.1016/j.cell.2017.11.008 (2017).
1072 2 Rao, S. S. *et al.* A 3D map of the human genome at kilobase resolution reveals
1073 principles of chromatin looping. *Cell* **159**, 1665-1680,
1074 doi:10.1016/j.cell.2014.11.021 (2014).
1075 3 Guo, Y. *et al.* CRISPR Inversion of CTCF Sites Alters Genome Topology and
1076 Enhancer/Promoter Function. *Cell* **162**, 900-910, doi:10.1016/j.cell.2015.07.038
1077 (2015).
1078 4 Ortabozkoyun, H. *et al.* CRISPR and biochemical screens identify MAZ as a
1079 cofactor in CTCF-mediated insulation at Hox clusters. *Nature genetics* **54**,
1080 202-212, doi:10.1038/s41588-021-01008-5 (2022).
1081 5 Goel, V. Y., Huseyin, M. K. & Hansen, A. S. Region Capture Micro-C reveals
1082 coalescence of enhancers and promoters into nested microcompartments. *Nature*
1083 *genetics*, doi:10.1038/s41588-023-01391-1 (2023).
1084 6 Bailey, S. D. *et al.* ZNF143 provides sequence specificity to secure chromatin
1085 interactions at gene promoters. *Nat Commun* **2**, 6186, doi:10.1038/ncomms7186
1086 (2015).
1087 7 Lis, M. & Walther, D. The orientation of transcription factor binding site motifs
1088 in gene promoter regions: does it matter? *BMC Genomics* **17**, 185,
1089 doi:10.1186/s12864-016-2549-x (2016).
1090 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 Consortium, G. T. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science (New York, N.Y.)* **369**, 1318-1330, doi:10.1126/science.aaz1776 (2020).

17 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).

18 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).

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

1121 binding microarray data on protein-DNA interactions. *Nucleic acids research* **37**,
1122 D77-82, doi:10.1093/nar/gkn660 (2009).
1123 20 Portales-Casamar, E. *et al.* JASPAR 2010: the greatly expanded open-access
1124 database of transcription factor binding profiles. *Nucleic acids research* **38**,
1125 D105-110, doi:10.1093/nar/gkp950 (2010).
1126 21 Jolma, A. *et al.* DNA-binding specificities of human transcription factors. *Cell*
1127 **152**, 327-339, doi:10.1016/j.cell.2012.12.009 (2013).
1128 22 Kheradpour, P. & Kellis, M. Systematic discovery and characterization of
1129 regulatory motifs in ENCODE TF binding experiments. *Nucleic acids research*
1130 **42**, 2976-2987, doi:10.1093/nar/gkt1249 (2014).
1131 23 Kulakovskiy, I. V. *et al.* HOCOMOCO: towards a complete collection of
1132 transcription factor binding models for human and mouse via large-scale
1133 ChIP-Seq analysis. *Nucleic acids research* **46**, D252-D259,
1134 doi:10.1093/nar/gkx1106 (2018).
1135 24 Shrikumar, A., Greenside, P. & Kundaje, A. in *Proceedings of the 34th*
1136 *International Conference on Machine Learning* Vol. 70 (eds Precup Doina &
1137 Teh Yee Whye) 3145--3153 (PMLR, Proceedings of Machine Learning Research,
1138 2017).
1139 25 Lee, B. H., Wu, Z. & Rhie, S. K. Characterizing chromatin interactions of
1140 regulatory elements and nucleosome positions, using Hi-C, Micro-C, and
1141 promoter capture Micro-C. *Epigenetics Chromatin* **15**, 41,
1142 doi:10.1186/s13072-022-00473-4 (2022).
1143 26 Cuddapah, S. *et al.* Global analysis of the insulator binding protein CTCF in
1144 chromatin barrier regions reveals demarcation of active and repressive domains.
1145 *Genome research* **19**, 24-32, doi:10.1101/gr.082800.108 (2009).
1146 27 Huang, S., Li, X., Yusufzai, T. M., Qiu, Y. & Felsenfeld, G. USF1 recruits
1147 histone modification complexes and is critical for maintenance of a chromatin
1148 barrier. *Mol Cell Biol* **27**, 7991-8002, doi:10.1128/MCB.01326-07 (2007).
1149 28 Crane, E. *et al.* Condensin-driven remodelling of X chromosome topology
1150 during dosage compensation. *Nature* **523**, 240-244, doi:10.1038/nature14450

1151 (2015).
1152 29 Matharu, N. K. & Ahanger, S. H. Chromatin Insulators and Topological
1153 Domains: Adding New Dimensions to 3D Genome Architecture. *Genes (Basel)*
1154 **6**, 790-811, doi:10.3390/genes6030790 (2015).
1155 30 Reyes, A. & Huber, W. Alternative start and termination sites of transcription
1156 drive most transcript isoform differences across human tissues. *Nucleic acids*
1157 *research* **46**, 582-592, doi:10.1093/nar/gkx1165 (2018).
1158 31 Hua, P. *et al.* Defining genome architecture at base-pair resolution. *Nature* **595**,
1159 125-129, doi:10.1038/s41586-021-03639-4 (2021).
1160 32 Xiao, T., Li, X. & Felsenfeld, G. The Myc-associated zinc finger protein (MAZ)
1161 works together with CTCF to control cohesin positioning and genome
1162 organization. *Proc Natl Acad Sci U S A* **118**, doi:10.1073/pnas.2023127118
1163 (2021).
1164 33 Ishihara, K., Oshimura, M. & Nakao, M. CTCF-dependent chromatin insulator
1165 is linked to epigenetic remodeling. *Molecular cell* **23**, 733-742,
1166 doi:10.1016/j.molcel.2006.08.008 (2006).
1167 34 Vierbuchen, T. *et al.* AP-1 Transcription Factors and the BAF Complex Mediate
1168 Signal-Dependent Enhancer Selection. *Molecular cell* **68**, 1067-1082 e1012,
1169 doi:10.1016/j.molcel.2017.11.026 (2017).
1170 35 Gong, Q. H., McDowell, J. C. & Dean, A. Essential role of NF-E2 in remodeling
1171 of chromatin structure and transcriptional activation of the epsilon-globin gene
1172 in vivo by 5' hypersensitive site 2 of the beta-globin locus control region. *Mol*
1173 *Cell Biol* **16**, 6055-6064, doi:10.1128/MCB.16.11.6055 (1996).
1174 36 Davidson, I. F. & Peters, J. M. Genome folding through loop extrusion by SMC
1175 complexes. *Nat Rev Mol Cell Biol* **22**, 445-464,
1176 doi:10.1038/s41580-021-00349-7 (2021).
1177 37 Karollus, A., Mauermeier, T. & Gagneur, J. Current sequence-based models
1178 capture gene expression determinants in promoters but mostly ignore distal
1179 enhancers. *Genome Biol* **24**, 56, doi:10.1186/s13059-023-02899-9 (2023).
1180 38 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).

39 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).

40 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).

41 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).

42 Christou-Kent, M. *et al.* CEBPA phase separation links transcriptional activity and 3D chromatin hubs. *Cell Rep* **42**, 112897, doi:10.1016/j.celrep.2023.112897 (2023).

43 Ji, D. *et al.* FOXA1 forms biomolecular condensates that unpack condensed chromatin to function as a pioneer factor. *Molecular cell* **84**, 244-260 e247, doi:10.1016/j.molcel.2023.11.020 (2024).

44 Bing, X. *et al.* Chromosome structure in Drosophila is determined by boundary pairing not loop extrusion. *eLife* **13**, doi:10.7554/eLife.94070 (2024).

45 Ke, W., Fujioka, M., Schedl, P. & Jaynes, J. B. Stem-loop and circle-loop TADs generated by directional pairing of boundary elements have distinct physical and regulatory properties. *eLife* **13**, doi:10.7554/eLife.94114 (2024).

46 Fujioka, M., Mistry, H., Schedl, P. & Jaynes, J. B. Determinants of Chromosome Architecture: Insulator Pairing in cis and in trans. *PLoS Genet* **12**, e1005889, doi:10.1371/journal.pgen.1005889 (2016).

47 Aljahani, A. *et al.* Analysis of sub-kilobase chromatin topology reveals nano-scale regulatory interactions with variable dependence on cohesin and CTCF. *Nat Commun* **13**, 2139, doi:10.1038/s41467-022-29696-5 (2022).

48 Hsieh, T. S. *et al.* Enhancer-promoter interactions and transcription are largely

1211 maintained upon acute loss of CTCF, cohesin, WAPL or YY1. *Nature genetics*
1212 **54**, 1919-1932, doi:10.1038/s41588-022-01223-8 (2022).
1213 49 Guerin, T. M., Barrington, C., Pobegalov, G., Molodtsov, M. I. & Uhlmann, F.
1214 An extrinsic motor directs chromatin loop formation by cohesin. *EMBO J* **43**,
1215 4173-4196, doi:10.1038/s44318-024-00202-5 (2024).
1216 50 Wang, R. *et al.* MyoD is a 3D genome structure organizer for muscle cell
1217 identity. *Nat Commun* **13**, 205, doi:10.1038/s41467-021-27865-6 (2022).
1218 51 Snead, W. T. & Gladfelter, A. S. The Control Centers of Biomolecular Phase
1219 Separation: How Membrane Surfaces, PTMs, and Active Processes Regulate
1220 Condensation. *Molecular cell* **76**, 295-305, doi:10.1016/j.molcel.2019.09.016
1221 (2019).
1222 52 Tang, X. *et al.* The PTM profiling of CTCF reveals the regulation of 3D
1223 chromatin structure by O-GlcNAcylation. *Nat Commun* **15**, 2813,
1224 doi:10.1038/s41467-024-47048-3 (2024).
1225 53 Phillips, J. E. & Corces, V. G. CTCF: master weaver of the genome. *Cell* **137**,
1226 1194-1211, doi:10.1016/j.cell.2009.06.001 (2009).
1227 54 Kumar, S., Kaur, S., Seem, K., Kumar, S. & Mohapatra, T. Understanding 3D
1228 Genome Organization and Its Effect on Transcriptional Gene Regulation Under
1229 Environmental Stress in Plant: A Chromatin Perspective. *Front Cell Dev Biol* **9**,
1230 774719, doi:10.3389/fcell.2021.774719 (2021).
1231 55 Dekker, J. & Misteli, T. Long-Range Chromatin Interactions. *Cold Spring Harb*
1232 *Perspect Biol* **7**, a019356, doi:10.1101/cshperspect.a019356 (2015).
1233 56 Spurrell, C. H., Dickel, D. E. & Visel, A. The Ties That Bind: Mapping the
1234 Dynamic Enhancer-Promoter Interactome. *Cell* **167**, 1163-1166,
1235 doi:10.1016/j.cell.2016.10.054 (2016).
1236 57 Kim, D. S. *et al.* The dynamic, combinatorial cis-regulatory lexicon of
1237 epidermal differentiation. *Nature genetics* **53**, 1564-1576,
1238 doi:10.1038/s41588-021-00947-3 (2021).
1239 58 Liu, M. *et al.* Genomic discovery of potent chromatin insulators for human gene
1240 therapy. *Nature biotechnology* **33**, 198-203, doi:10.1038/nbt.3062 (2015).

1241 59 Sloan, C. A. *et al.* ENCODE data at the ENCODE portal. *Nucleic acids research*
1242 **44**, D726-732, doi:10.1093/nar/gkv1160 (2016).
1243 60 Wang, L., Wang, S. & Li, W. RSeQC: quality control of RNA-seq experiments.
1244 *Bioinformatics (Oxford, England)* **28**, 2184-2185,
1245 doi:10.1093/bioinformatics/bts356 (2012).
1246 61 Kolmykov, S. *et al.* GTRD: an integrated view of transcription regulation.
1247 *Nucleic acids research* **49**, D104-D111, doi:10.1093/nar/gkaa1057 (2021).
1248 62 Zhao, H. *et al.* CrossMap: a versatile tool for coordinate conversion between
1249 genome assemblies. *Bioinformatics (Oxford, England)* **30**, 1006-1007,
1250 doi:10.1093/bioinformatics/btt730 (2014).

1251

1252

1253 **Acknowledgements**

1254 We would like to thank Prof. Martin Frith for reviewing the manuscript.
1255 Supercomputing resources were provided by the Human Genome Center of the Institute
1256 of Medical Science at the University of Tokyo and ROIS National Institute of Genetics.
1257 This work was partially supported by JSPS KAKENHI (Grant Number 16K00387), JST
1258 CREST (Grant Number JPMJCR15G1), Osaka University start-up research grant,
1259 KIOXIA young scientist research grant at Waseda University, Sumitomo basic science
1260 research projects to NO, AMED [grant nos. JP21ae0121049, 21479280], and JSPS
1261 KAKENHI [grant nos. 22H04925, 20H00624, 17K20032] to MH.

1262

1263 **Author contributions**

N.O. designed and performed the research, and wrote and revised the manuscript. M.H. revised the manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

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

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