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. For instance, we found that the key regulator MyoD-binding site is located at an insulator site near a gene involved in skeletal muscle differentiation and function. 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,

35 regulate the transcription of genes[1]. Chromatin interactions form loop structures
36 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 learning 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 and Note).

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, indicating 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). Furthermore, we examined the directional bias of
242 a cohesin, RAD21, which is involved in chromatin loops with CTCF, using ChIA-PET
243 chromatin interaction data for RAD21. The number of chromatin interactions and the
244 percent difference of RAD21 DNA-binding sites at a pair of chromatin interaction sites
245 more than 5-kb apart: FR 7,934 (76.3%), RF 351 (3.4%), FF 965 (9.3%), and RR 1,151
246 (11.1%), which is similar to the difference of the CTCF DNA-binding sites. The
247 analysis of the contribution scores of the RAD21 DNA-binding sites showed the same
248 difference as the RAD21 ChIA-PET analysis and CTCF (i.e. a large percent difference
249 between FR and RF and almost the same between FF and RR) (Fig. 3c and Table 1).
250 These data indicated that the deep learning analysis accurately predicted the percent
251 differences in the FR, FF, RR, and RF orientations of CTCF DNA-binding sites at
252 chromatin interaction sites on the basis of the DeepLIFT scores of the DNA-binding
253 sites in putative enhancer and open chromatin regions.

254 A total of 99 insulator-associated DBPs were predicted based on the stringent

255 criterion of the DNA-binding sites of a given DBP showing either a specific directional
256 bias or no directional bias, as determined by comparing the four possible orientations
257 (FR, RF, FF, and RR) and the absence of directional bias (X). Among 99
258 insulator-associated DBPs, EP300 and SIN3A do not directory interact with DNA, and
259 thus 97 insulator-associated DBPs directory bind to DNA. Most of the
260 insulator-associated DBPs showed a directional bias. Non-directional motifs were
261 identified for only 6 DBPs such as FOS, FOSL2, JUN, JUNB, and JUND with
262 palindromic regulatory motifs and SPI1 interacting with the DBPs. We found that 23
263 DBPs exhibited insulator and CTCF-associated functions and/or chromatin regulatory
264 functions that affected EPIs, as reported in previous studies: 13 DBPs were associated
265 with CTCF (Table S7). The 23 DBPs were multifunctional proteins, and their known
266 functions were transcriptional regulator, repressor, chromatin regulator, and interaction
267 with other proteins including enzymes that facilitate epigenetic modifications. These
268 findings suggested that CTCF regulated EPIs by interacting with other DBPs, and that
269 DBPs with other functions were also associated with the regulation of EPIs and
270 chromatin conformation.

271 To further investigate whether insulator-associated DBPs are involved in a pair of
272 chromatin interaction sites, we searched for the DNA-binding sites of
273 insulator-associated DBPs located at a pair of 400-bp and 200-bp long chromatin
274 interaction sites that were more than 5-kb apart and did not overlap with 2-kb upstream

275 and downstream of TSSs, considering the FR, RF, FF, or RR orientation of the pair of
276 DNA-binding sites (Fig. 3e), since 400-bp long chromatin interaction sites would
277 include the cluster of DNA-binding sites of insulator-associated DBPs at insulator sites
278 (explained in the next section and Fig. 4f). For the 400-bp long chromatin interaction
279 sites, the DNA-binding sites of 44 (45%) and 97 (100%) insulator-associated DBPs,
280 excluding EP300 and SIN3A, were identified at a pair of chromatin interaction sites or
281 at either site of 30,369 chromatin interactions, respectively. For the 200-bp long
282 chromatin interaction sites, the DNA-binding sites of 30 (31%) and 97 (100%)
283 insulator-associated DBPs were identified at a pair of chromatin interaction sites or at
284 either site of 10,384 chromatin interactions, respectively, although long-range chromatin
285 interactions would not be fully identified due to the limited coverage of chromatin
286 interaction data from the comparison of results based on the distance between a pair of
287 chromatin interaction sites [25]. CTCF, RAD21, and SMC3 showed directional
288 DNA-binding sites as well as CTCF and RAD21 ChIA-PET chromatin interaction data.
289

290 **Search for insulator-associated DNA-binding sites based on histone modifications**
291 **and gene expression**

292 To confirm the predicted insulator-associated DBPs via another method and select
293 their DNA-binding sites associated with insulators, we first analyzed previously
294 reported histone modification data to identify differences between the upstream and

295 downstream regions of the DNA-binding sites. CTCF bound to sites located between
296 repressed regions with H3K27me3 marks and active regions with H2AK5ac marks,
297 which showed different signal intensities between the upstream and downstream regions
298 of CTCF DNA-binding sites [26]. The transcription factor USF1 exhibits a barrier
299 function, in which histone modification factors are recruited to separate adjacent regions
300 into activating and repressive chromatin domains [27]. We selected H3K27me3 histone
301 modifications, associated with the repression of gene expression, for comparison
302 between the upstream and downstream regions of the DNA-binding sites of predicted
303 insulator-associated DBPs using H3K27me3 ChIP-seq data because H3K27me3 histone
304 marks occupy long genomic regions and are thus easily compared, between genomic
305 regions. We searched for DNA-binding sites where the mean H3K27me3 signal
306 intensities differed by more than 2-fold between the upstream and downstream regions
307 of the DNA-binding site. The higher mean signal was found to be greater than the first
308 quartile of all H3K27me3 signal intensities because some H3K27me3 signals had a low
309 intensity and appeared noisy. Thus, the criteria were sufficient to clearly distinguish
310 modification status among genomic regions.

311 We further counted the numbers of DNA-binding sites with different H3K27me3
312 signal intensities between the upstream and downstream regions of a DNA-binding site
313 and found that known insulator-associated DBPs, such as CTCF, RAD21, and SMC3,
314 generally showed the greatest numbers of DNA-binding sites (Table S4). Via statistical

315 analysis, we calculated the median and mean numbers of DNA-binding sites showing
316 different H3K27me3 signal intensities between upstream and downstream regions for
317 the DNA-binding sites of both predicted insulator-associated DBPs and other DBPs.
318 The median and mean numbers of DNA-binding sites were significantly larger among
319 predicted insulator-associated DBPs than other DBPs (Mann–Whitney U test, p value <
320 0.05) (Table 2, Table S4, Fig. 4a, b). We confirmed that a larger number of the
321 DNA-binding sites of known insulator-associated DBPs were located between activated
322 and repressed regions than non-insulator-associated DBPs.

323 To identify additional genomic regions obviously separating transcriptionally
324 activated and repressed regions, we also examined the expression levels of transcripts
325 using the signal intensity indicated by RNA-seq data for identifying transcriptionally
326 active regions based on the available HFF cell data. As in the analysis of H3K27me3
327 signal intensity, we searched for DNA-binding sites of predicted insulator-associated
328 DBPs with mean transcript signal intensities differing more than 1.5-fold between
329 upstream and downstream regions, and the higher mean signal was greater than the first
330 quartile of all transcript signal intensities. We again found that a significantly larger
331 number of the DNA-binding sites of predicted insulator-associated DBPs were located
332 between activated and repressed regions than other DBPs (Table 2, Table S4, Fig. 4a, c).

333 In the analysis of chromatin interaction data, insulation scores were calculated
334 based on the analysis of the boundary regions of chromatin interactions [28]. We also

335 examined the overlap of DNA-binding sites of predicted insulator-associated DBPs with
336 the boundary regions, and again found that a significantly larger number of the
337 DNA-binding sites of predicted insulator-associated DBPs were located in the boundary
338 regions than other DBPs (Table 2, Fig. 4d).

339 To determine whether the DNA-binding sites of predicted insulator-associated
340 DBPs potentially act as insulators or cofactors of insulators, we investigated the
341 genomic locations of the DNA-binding sites on the UCSC Genome Browser (Fig. 4a,
342 S3). We searched for DNA-binding sites located at boundary sites identified by Micro-C
343 chromatin interaction data from 4D Nucleome Data Portal. We also searched for
344 DNA-binding sites located between the H3K27me3 marks and transcribed regions of
345 genes, which were primarily located at the edges of H3K27me3 marks and were not
346 found near the binding sites of CTCF, RAD21, and SMC3. For each DBP, we looked for
347 insulator sites associated with the DNA-binding sites in the genome browser and
348 selected primarily three DNA-binding sites for each DBP (Fig. S3a for boundary sites of
349 chromatin interactions and Fig. S3f for histone marks and transcribed regions as
350 insulator sites). We discovered DNA-binding sites in the expected genomic regions of
351 all DBPs, suggesting that these DNA-binding sites are associated with insulator
352 functions. Interestingly, MyoD1, which is an insulator-associated DBP, bound to an
353 insulator and boundary site and located near SELENON (SEPN1) gene [29]. MyoD1
354 and SELENON are involved in the differentiation of human skeletal muscle, suggesting

355 that the insulator-associated DBP affects the expression of SELENON and the
356 differentiation (Fig. 4a). The same analysis was performed using H3K9me3 marks,
357 instead of H3K27me3 (Fig. S3b). We found that the distribution of H3K9me3 signal
358 was different from that of H3K27me3 in some regions, and discovered the
359 insulator-associated DNA-binding sites close to the edges of H3K9me3 regions (Fig.
360 S3b). The insulator-associated DNA-binding sites potentially regulate boundary sites
361 associated with other histone modification marks, including H3K9me3 marks. We also
362 found that the multiple DNA-binding sites of insulator-associated DNA-binding
363 proteins were located close to each other at insulator sites and were associated with
364 distinct chromatin interactions (Fig. 4f, S3c, S3d). Topologically associated domains
365 (TADs) of chromosome are structured by architectural/insulator proteins clustered
366 around their boundaries, maintaining the megabase-scale chromatin structure, while
367 intra-TAD looping is mediated by discrete motifs for insulator factor binding [30].

368 When comparing insulator sites identified by histone modification marks and
369 transcribed regions with boundary sites based on chromatin interaction data in the
370 genome browser, we observed that the insulator sites did not always overlap with 5-kb
371 long boundary sites at the same location, some insulator sites were more than 5-kb away
372 from a boundary site. Overall, among 20,837 DNA-binding sites of the 97
373 insulator-associated proteins found at insulator sites identified by H3K27me3 histone
374 modification marks (type 1 insulator sites), 1,315 (6%) overlapped with 264 of 17,126

375 5-kb long boundary sites, and 6,137 (29%) overlapped with 784 of 17,126 25-kb long
376 boundary sites in HFF cells. Among 5,205 DNA-binding sites of the 97
377 insulator-associated DNA-binding proteins found at insulator sites identified by
378 H3K27me3 histone modification marks and transcribed regions (type 2 insulator sites),
379 383 (7%) overlapped with 74 of 17,126 5-kb long boundary sites, 1,901 (37%)
380 overlapped with 306 of 17,126 25-kb long boundary sites. Although CTCF-binding sites
381 separate active and repressive domains [26], the limited number of DNA-binding sites of
382 insulator-associated proteins found at type 1 and 2 insulator sites overlapped boundary
383 sites identified by chromatin interaction data. Furthermore, by analyzing the regulatory
384 regions of genes, the DNA-binding sites of the 97 insulator-associated DNA-binding
385 proteins were found (1) at the type 1 insulator sites (based on H3K27me3 marks) in the
386 regulatory regions of 3,170 genes, (2) at the type 2 insulator sites (based on H3K27me3
387 marks and gene expression levels) in the regulatory regions of 1,044 genes, and (3) at
388 insulator sites as boundary sites identified by chromatin interaction data in the
389 regulatory regions of 6,275 genes. The boundary sites showed the highest number of
390 overlaps with the DNA-binding sites. Comparing the insulator sites identified by (1)
391 and (3), 1,212 (38%) genes have both types of insulator sites. Comparing the insulator
392 sites between (2) and (3), 389 (37%) genes have both types of insulator sites. From the
393 comparison of insulator and boundary sites, we found that (1) or (2) types of insulator
394 sites overlapped or were close to boundary sites identified by chromatin interaction

395 data.

396 While preparing Fig. 4a and S3, we noticed that several DNA-binding sites of
397 insulator-associated DBPs were found in the same genomic locus, close to each other on
398 the genome. These DBPs potentially act together to form an insulator site. In addition,
399 some transcribed regions in the genome were differentially repressed in alternative
400 transcripts and overlapped with a H3K27me3 mark that was separated with the
401 DNA-binding site of an insulator-associated DBP between the repressive and active
402 regions (Fig. 5, S3e). Furthermore, the DNA-binding sites of insulator-associated DBPs
403 were frequently located near transcription start and termination sites. Alternative
404 transcription start and termination sites have been reported to explain the majority of
405 tissue-specific exon usage rather than alternative splicing[31]. We performed the statistical
406 test to estimate the enrichment of insulator-associated DNA-binding sites compared to
407 the other DNA-binding proteins, and found that the insulator-associated DNA-binding
408 sites were significantly more abundant at splice sites than the DNA-binding sites of the
409 other proteins (Fig 4e and Table 2; Mann–Whitney U test, p value < 0.05). The
410 comparison between the splice sites of both ends of first and last introns and those of
411 other introns showed the similar statistical significance of enrichment and number of
412 splice sites with the insulator-associated DNA-binding proteins (Table 2 and Table S9).
413 The DNA-binding sites of insulator-associated DBPs could affect the form of transcripts
414 and the genomic location of transcription start and termination sites. The differences in

415 the DNA-binding sites and H3K27me3 marks between tissues could contribute to the
416 cell and tissue type specific exon usage of alternative transcripts.
417
418 **Distribution of DNA-binding sites around chromatin interaction sites**
419 In this study, we aimed to predict chromatin- and insulator-associated DBPs. To
420 estimate how many DBPs are involved in chromatin interactions and examine the
421 distribution and association of DNA-binding sites around chromatin interaction sites,
422 the DNA-binding sites of DBPs in regions within ±500 bp of the centers of chromatin
423 interaction sites were identified in HFF cells. As an improved method for chromatin
424 interaction analysis, Micro-C assays enabled us to identify chromatin interactions at a
425 higher resolution than can be achieved via Hi-C analysis. Base pair resolution was
426 reported using Micro Capture-C [32]. The distribution of CTCF, RAD21, and SMC3
427 DNA-binding sites showed a characteristic pattern around chromatin interaction sites,
428 with the highest peaks located around interaction sites (Fig. 6a, b, c). The peaks of the
429 DNA-binding sites of YY1 and ZNF143, which are known to be involved in chromatin
430 interactions, were distributed around chromatin interaction sites. The DNA-binding sites
431 of YY1 were also broadly distributed in regions within ±500 bp of interaction sites,
432 suggesting that this DBP is involved in various functions in addition to chromatin
433 interactions (Fig. 6d, e). MAZ, which exhibits insulator function independent of CTCF
434 [33], indicated the peak of distribution of MAZ DNA-binding sites around interaction sites

435 (Fig. 6f). For the insulator-associated DBPs other than CTCF, RAD21, and SMC3, the
436 DNA-binding sites that do not overlap with those of CTCF, RND21, and SMC3 were
437 used to examine their distribution around interaction sites.

438 Among the 228 DBPs, more than 1,000 of the DNA-binding sites in open
439 chromatin regions in HFF cells overlapped with the ChIP-seq peaks of the DBPs, and
440 the peaks of almost all the DBPs (approximately 90%) were located around interaction
441 sites. Furthermore, we found DBPs with sharp peaks around interaction sites, DBPs
442 with broad peaks around interaction sites, and DBPs with peaks similar to those of
443 CTCF, RAD21, and SMC3 around interaction sites, suggesting that these DBPs may be
444 associated with CTCF, RAD21, and/or SMC3 (Table S8; Fig. S4). Among the predicted
445 insulator-associated DBPs, MeCP2 and PRDM9 exhibited fewer than 1,000
446 DNA-binding sites in HFF cells, and when a DBP has fewer than 1,000 DNA-binding
447 sites, the specific distribution of DNA-binding sites around chromatin interaction sites
448 may not be evident (Fig. S4). Approximately 90% of the DBPs with more than 1,000
449 DNA-binding sites had peaks around interaction sites. This indicated that the DBPs
450 interacted with other genomic regions and that the DNA-binding sites of predicted
451 insulator-associated DBPs were involved in chromatin interactions.

452

453 **Co-localization of DNA-binding sites around chromatin interaction sites**

454 Since the DNA-binding sites of the predicted insulator-associated DBPs may be

455 located close to those of known insulator-associated DBPs, such as CTCF, RAD21,
456 and/or SMC3, and, thus, be presumed to act as insulators, we next counted the number
457 of DNA-binding sites in each 10-bp bin within ±500 bp from the center of a chromatin
458 interaction site and compared these numbers between all possible pairs of DBPs. For
459 this analysis, the DNA-binding sites that overlapped with the ChIP-seq peaks of a DBP
460 were selected to reduce the number of false-positive DNA-binding site predictions, and
461 the number of DNA-binding sites per DBP may therefore have been underestimated.
462 Among the 228 DBPs (Selection Criteria S1 in Fig. S1), 245 pairs of 135 DBPs
463 colocalized around chromatin interaction sites when at least 30% of the DNA-binding
464 sites of a pair of DBPs colocalized in a bin with the same distance from the interaction
465 site (Table S5). CTCF showed a high rate of colocalization of DNA-binding sites with
466 both RAD21 (39%) and SMC3 (81%) in bins within ±500 bp from the center of a
467 chromatin interaction site, whereas YY1 and ZNF143 showed low rates of
468 colocalization of DNA-binding site with RAD21 (1.5% and 5.5%, respectively) and
469 CTCF (2.3% and 4.8%, respectively) (Table 3). For CTCFL (CCCTC-binding factor
470 like), RXRA, PPARG, NEUROD1, and ZBTB7A, more than 10% of DNA-binding sites
471 colocalized with CTCF, RAD21, and/or SMC3 in bins within the same distance from
472 chromatin interaction sites, suggesting that the insulator function of the DBPs was
473 associated with known insulator DBPs. For the other DBPs among the 228 DBPs, less
474 than 10% of DNA-binding sites colocalized with CTCF, RAD21, and/or SMC within a

475 bin. Most of the predicted insulator-associated DBPs did not have the colocalization of
476 DNA-binding sites and were not in close proximity to known insulator-associated DBPs
477 such as CTCF, RAD21, and SMC3; that is, they were not within a 10-bp bin in the
478 regions ±500 bp from the center of an interaction site.

479 We found that the multiple DNA-binding sites of insulator-associated
480 DNA-binding proteins were located close to each other at insulator sites (Fig. S3d). If
481 we identify the co-localization of the DNA-binding sites located at boundary sites of
482 chromatin interactions compared to non-boundary sites, the co-localization of
483 DNA-binding sites is potentially involved in boundary and insulator functions. Among
484 the 201 DBPs (Selection Criteria S2 in Fig. S1), we identified 197 pairs of 101 DBPs
485 that colocalized within a 10-bp bin in the regions ±500 bp from the center of a
486 chromatin interaction site in a 5-kb long boundary site compared to non-boundary sites,
487 when at least 30% of the DNA-binding sites of a pair of DBPs colocalized in a bin with
488 the same distance from the interaction site (Table S6). This suggested that the
489 insulator-associated DNA-binding proteins interact with each other at chromatin
490 interaction sites at insulator sites like CTCF and cohesin.

491 To further examine the possibility that known insulator-associated DBPs interact
492 indirectly with the other predicted insulator-associated DBPs, we investigated the
493 overlap of DNA-binding sites with the DBP ChIP-seq peaks of CTCF, RAD21, and
494 SMC3 and found that a high percentage (> 98%) of the DNA-binding sites of the three

495 DBPs overlapped with their ChIP-seq peaks in all three examined combinations.
496 However, the DNA-binding sites of almost all other predicted insulator-associated
497 DBPs overlapped with less than 50% of the ChIP-seq peaks of the three known DBPs,
498 even though the ChIP-seq data were collected from various cell and tissue types. These
499 results suggested that the predicted insulator-associated DBPs play different functional
500 roles than the known DBPs.

501 Although a DNA-binding-site distribution pattern around chromatin interaction
502 sites similar to those of CTCF, RAD21, and SMC3 was observed for DBPs such as
503 BACH2, FOS, ATF3, NFE2, and MAFK, less than 1% of the DNA-binding sites of the
504 latter set of DBPs colocalized with CTCF, RAD21, or SMC3 in a single bin (Fig. S4c).
505 However, BACH2, FOS, ATF3, NFE2, and MAFK had DNA-binding sites that
506 colocalized with each other (Table 3). These results suggested that the two different
507 groups of colocalized DBPs were independently involved in the same chromatin-related
508 function such as chromatin remodeling [34 35 36].

509

510 **Discussion**

511 Here, we present the first, to our knowledge, computational method and
512 genome-wide analysis aimed at detecting insulator-associated DBPs. Transcriptional
513 regulation mediated by the binding of DBPs at sites located far from TSSs was
514 modelled using our deep learning method based on eQTL data. Through this approach,

515 the effects of the DNA-binding sites of insulator-associated DBPs on EPIs and gene
516 expression were estimated, and insulator-associated DBPs were predicted. We further
517 identified DBPs, including known insulator-associated DBPs that exhibit
518 insulator-associated functions and regulate transcription, thereby affecting EPIs reported
519 in previous studies. The DNA-binding sites of the predicted insulator-associated DBPs
520 showed orientation bias of their DNA-binding sites; unexpectedly, the significant
521 differences in the orientation bias of CTCF DNA-binding sites based on DeepLIFT
522 scores (contribution scores) were found to be consistent with the percent differences in
523 the orientation of CTCF DNA-binding sites at CTCF-mediated chromatin interaction
524 sites. Furthermore, statistical testing of the contribution scores revealed that the
525 directional sites of DBPs contributed more significantly to the prediction of gene
526 expression levels than the non-directional sites. The DNA-binding sites of the predicted
527 insulator-associated DBPs were located at potential insulator sites more frequently than
528 those of other DBPs, according to the differences in H3K27me3 signal intensities and
529 transcript levels, and according to insulator sites as boundary sites identified by
530 chromatin interactions. We further observed that only a small percentage of the
531 DNA-binding sites of the insulator-associated DBPs colocalized with the DNA-binding
532 sites of known insulator-associated DBPs such as CTCF, RAD21, and SMC3 around
533 chromatin interaction sites. We discovered that DBPs other than the known insulator
534 DBPs and the directionality of their DNA-binding sites are associated with insulators.

535 Our deep learning analysis accurately predicted the percent differences in the FR,
536 FF, RR, and RF orientations of CTCF DNA-binding sites at chromatin interaction sites,
537 considering the genomic locations of DNA-binding sites of putative insulator-associated
538 DBPs. The other predicted insulator-associated DBPs also showed orientation bias of
539 their DNA-binding sites. These results suggest that the directional bias of DNA-binding
540 sites of insulator-associated DBPs would be involved in insulator function and
541 chromatin regulation through structural interactions among DBPs, other proteins, DNAs,
542 and RNAs. For example, the N-terminal amino acids of CTCF have been shown to
543 interact with RAD21 in chromatin loops [8-10]. Cohesin complexes that interact with
544 convergent CTCF sites, that is, the N-terminus of CTCF, might be protected from
545 WAPL, but those that interact with divergent CTCF sites, that is, the C-terminus of
546 CTCF, might not be protected from WAPL, which could release cohesin from chromatin
547 and thus disrupt cohesin-mediated chromatin loops [37]. For another example, although
548 sequence-based predictions of EPIs underestimated actual interactions, an
549 activity-by-contact (ABC) approach, based on a simple distance model, performed well,
550 suggesting that factors in addition to sequence influence EPIs [38,39]. Furthermore,
551 CTCF-mediated chromatin looping induces the formation of transcriptional condensates
552 and phase separation between the inner and outer regions of the loops, resulting in
553 insulator function [40-42]. Therefore, structures that result in phase separation, rather than
554 functional differences in DBPs, may be responsible for insulator functions. CEBPA in

555 the identified insulator-associated DNA-binding proteins was also reported to be
556 involved in transcriptional condensates and phase separation [43]. FOXA1 pioneer factor
557 functions as an initial chromatin-binding and chromatin-remodeling factor and has been
558 reported to form biomolecular condensates [44].

559 Other types of chromatin regulation are also expected to be related to the
560 structural interactions of molecules. As the boundary-pairing (insulator-pairing) model,
561 molecules bound to insulators physically pair with their partners, either head-to-head or
562 head-to-tail, with different degrees of specificity at the termini of TADs in flies (Fig. 7).
563 Although the experiments do not reveal how partners find each other, the mechanism
564 unlikely requires loop extrusion, as presumed in CTCF and cohesin-mediated chromatin
565 looping. Homologous and heterologous insulator-insulator pairing interactions are
566 central to the architectural functions of insulators. The manner of insulator-insulator
567 interactions is orientation-dependent [45-47].

568 Regional micro-capture (Micro-C) analyses revealed that microcompartment
569 interactions are largely independent of cohesin and transcription [5]. The depletion of
570 cohesin causes a subtle reduction in longer-range enhancer-promoter interactions and
571 that CTCF depletion can cause rewiring of regulatory contacts [48]. Another group
572 reported that enhancer-promoter interactions and transcription are largely maintained
573 upon depletion of CTCF, cohesin, WAPL or YY1. Instead, cohesin depletion decreased
574 transcription factor binding to chromatin. Thus, cohesin may allow transcription factors

575 to find and bind their targets more efficiently [49]. Furthermore, the loop extrusion is not
576 essential for enhancer-promoter interactions, but contributes to their robustness and
577 specificity and to precise regulation of gene expression [48]. The 'loop extrusion'
578 hypothesis is motivated by in vitro observations. The experiment in yeast, in which
579 cohesin variants that are unable to extrude DNA loops but retain the ability to
580 topologically entrap DNA, suggested that in vivo chromatin loops are formed
581 independently of loop extrusion. Instead, transcription promotes loop formation and acts
582 as an extrinsic motor that extends these loops and defines their final positions [50]. Maz
583 and MyoD1 among the identified insulator-associated DNA-binding proteins form loops
584 without CTCF [4,33,51]. Furthermore, protein post-transcriptional modifications (PTMs)
585 facilitate control over the molecular valency and strength of protein-protein interactions
586 [52]. O-GlcNAcylation as a PTM inhibits CTCF binding to chromatin [53]. We found that
587 the identified insulator-associated DNA-binding proteins tend to form a cluster at
588 potential insulator sites (Fig. 4f and Supplementary Fig. 3c). These proteins may
589 interact and actively regulate chromatin interactions, transcriptional condensation, and
590 phase separation through PTMs. Our analyses will contribute to reveal novel functional
591 and structural roles of DBPs for insulator functions, chromatin regulation, and phase
592 separation and to advance their experimental analysis.

593 High-resolution chromatin interaction data from a Micro-C assay indicated that
594 most of the predicted insulator-associated DBPs showed the peak of the

595 DNA-binding-site distribution around chromatin interaction sites, suggesting that these
596 DBPs are involved in chromatin interactions and that the chromatin interaction data has
597 a high degree of resolution. Base pair resolution was reported using Micro Capture-C [32].
598 However, due to the high resolution of a Micro-C assay, it is difficult to obtain all
599 chromatin interaction data at a reasonable and available cost, while Micro Capture-C
600 and Region Capture Micro-C assays focus on a specific genomic region to increase the
601 coverage of chromatin interactions. The coverage of chromatin interactions will change
602 according to the experimental data of a Micro-C assay. In addition, chromatin
603 interactions include other types of interactions that are not associated with EPIs and
604 insulator functions [54-56]. Therefore, we analyzed and predicted insulator-associated
605 DBPs without using chromatin interaction data. Our deep learning analysis successfully
606 learned EPIs that affect gene expression levels and accurately predicted the percent
607 differences in the FR, FF, RR, and RF orientations of CTCF DNA-binding sites at
608 chromatin interaction sites.

609 When comparing insulator sites separated by active and repressive domains with
610 boundary sites identified by chromatin interaction data, we identified insulator sites
611 based on the total signal levels of H3K27me3 histone modification marks and the
612 expression level of transcripts in a region. Some insulator sites may not have smooth
613 continuous distributions of histone marks and expression levels, which may cause the
614 difference between the insulator and boundary sites. Another reason for the difference

615 may be that boundary sites are more closely associated with topologically associated
616 domains (TADs) of chromosome than are insulator sites [30]. Boundary sites are regions
617 identified based on the separation of numerous chromatin interactions. On the other
618 hand, we found that the multiple DNA-binding sites of insulator-associated
619 DNA-binding proteins were located close to each other at insulator sites and were
620 associated with distinct nested and focal chromatin interactions, as reported by Micro-C
621 assay [5]. These interactions may be transient and relatively weak, such as tissue/cell type,
622 conditional or lineage-specific interactions [57].

623 AI (Artificial Intelligence) is considered as a black box, since the reason and
624 cause of prediction are difficult to know. To solve this issue, tools and methods have
625 been developed to know the reason and cause. These technologies are called
626 Explainable AI. DeepLIFT is considered to be a tool for Explainable AI. However,
627 DeepLIFT does not answer the reason and cause for a prediction. It calculates scores
628 representing the contribution of the input data to the prediction. The analysis of many
629 contribution scores (i.e., DeepLIFT scores in this study) for directional DNA-binding
630 sites of putative insulator-associated DBPs in the prediction of transcript expression
631 levels successfully predicted the percent differences in the FR, FF, RR, and RF
632 orientations of CTCF DNA-binding sites at chromatin interaction sites based on the
633 contribution scores of DNA-binding sites in putative enhancer and open chromatin
634 regions. Some contribution scores may not be accurate and thus the prediction of

635 insulator-associated DBPs or DNA-binding sites based on a small number of
636 contribution scores may be erroneous. To compare contribution scores computed based
637 on different pairs of input and output data with our deep learning model, we normalized
638 contribution scores for each pair of input and output data as previously reported [58]. The
639 analysis of many contribution scores will be used to estimate the effects of other
640 molecules and functions involved in gene regulation.

641 One of the next computational challenges will be to predict and narrow the
642 potential set of insulator-associated DNA-binding sites of multifunctional DBPs. To
643 confirm the machine learning prediction and to identify insulator-associated
644 DNA-binding sites in the genome for further experimental study, information on histone
645 modifications such as H3K27me3 and transcript levels will be beneficial, and other
646 epigenomic data can be used for this purpose. As a reference for an application of this
647 study, Liu et al. focused on CTCF-binding sites and performed various assays in human
648 and mouse cell cultures and in mice with the goal of applying these sites in gene therapy
649 [59]. The analysis of insulator-associated DBPs may contribute to the development of
650 technologies to regulate gene expression. In conclusion, our novel computational
651 method discovered insulator-associated DBPs, including known DBPs, and revealed
652 that the directional bias of insulator-associated DBPs is potentially involved in
653 chromatin looping, phase separation, transcriptional condensates and insulator functions
654 through the structural interactions of the DBPs. This will lead to a general and global

655 understanding of transcriptional regulation and gene expression.

656

657 # Methods

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

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

660 To predict DNA-binding sites that also act as insulator sites in the human genome,
661 we modified a previously developed deep learning method designed to predict the
662 transcript expression levels of genes in tissues and cell types based on the DNA-binding
663 sites of DBPs in promoter regions and other DNA regions (Fig. 2a; Fig. S4) [14]. To
664 reduce the influence of cellular heterogeneity observed when analyzing tissues, we used
665 the gene expression data from four distinct cell types where various experimental data
666 are available in public databases: HFFs (GSM941744), monocytes (Blueprint RNA-seq
667 FPKM [fragments per kilobase of exon per million reads mapped] data
668 'C0010KB1.transcript_quantification.rsem_grape2_crg.GRCh38.20150622.results' in
669 the Blueprint DCC portal at http://dcc.blueprint-epigenome.eu/#/files), human
670 mammary epithelial cells (HMECs) (Encyclopedia of DNA Elements at UCSC
671 http://genome.ucsc.edu/encode/,
672 'wgEncodeCshlLongRnaSeqHmecCellPapAlnRep1.bam'), and neural progenitor cells
673 (NPCs) (GSM915326, ENCFF529SIO). For gene expression data, FPKM values
674 representing the expression levels of genes were retrieved or calculated based on

675 GENCODE V19 gene annotations in the ENCODE portal database [60]
676 (https://www.encodeproject.org/) using the RSeQC tool [61] (http://rseqc.sourceforge.net/).
677 We calculated log2-fold-transformed FPKM values (adding 0.25 to prevent results of
678 negative infinity), median log2-FPKM values across the four cell types, and log2-fold
679 changes relative to the median for all cell types according to the original deep learning
680 method (transcriptome data processing as described by Tasaki et al.) [14].
681
682 **DNA-binding sites as input data for machine learning analysis**
683 To investigate the effect of insulator sites on EPIs and gene expression, the
684 DNA-binding sites in putative enhancer and open chromatin regions located away from
685 promoter regions were added as input data in the deep learning method. To identify the
686 DNA-binding sites of DBPs in promoter and enhancer regions, ChIP-seq data of the
687 DBPs were used and were updated to GTRD database release v20.06
688 (https://gtrd.biouml.org) [62]. The binding site of a DBP regulating the expression level of
689 each transcript was selected using eQTL data (GTEx_Analysis_v8_eQTL.tar) from
690 GTEx portal release v8 (https://gtexportal.org). If a DNA-binding site was located
691 within a 100-bp region around a single-nucleotide polymorphism (SNP) of an eQTL,
692 we assumed that the DBP regulated the expression of the transcript corresponding to the
693 eQTL. If the SNP of an eQTL was not located on the same chromosome as the
694 transcribed region of the eQTL (i.e., *trans*-eQTL), the eQTL data were not used. The

695 genomic positions of the eQTLs were converted to the corresponding positions in the
696 hg19 version of the human genome using the LiftOver tool
697 (http://hgdownload.soe.ucsc.edu/downloads.html#liftover).
698 The original model of the input data of DNA-binding sites in promoter regions
699 used for the deep learning analysis consisted of the lengths of a DBP ChIP-seq peak in
700 each 100-base bin in a region between 2 kbp upstream and 1 kbp downstream of the
701 transcription start site (TSS) of each gene or transcript [14]. To add the input data of
702 DNA-binding sites of DBPs in more upstream and downstream regions of a TSS and to
703 reduce the number of bins, we modified the model and first computed the length of a
704 DBP ChIP-seq peak in a 100-base bin and then selected the maximum length in each
705 interval of 100 bins in a region between 1 Mbp upstream and downstream of a TSS.
706 This step outputted the 200 longest DBP ChIP-seq peaks in the examined regions.
707
708 **Training step for machine learning analysis**
709 In the training of the modified deep learning model, we used the same test,
710 training, and validation data as in the original method: all 2,705 genes encoded on
711 chromosome 1 as test data, training data (22,251 genes), and validation data (2,472
712 genes) randomly selected from the other genes [14]. The input data for promoters and
713 putative enhancer regions were concatenated into a total of 230 bins of the lengths of
714 1,310 DBP ChIP-seq peaks, consisting of 30 bins for promoters and 200 bins for

715 putative enhancers because the DeepLIFT tool 'RevealCancel rule on dense layers and
716 the Rescale rule on conv layers (genomics default)', with high scoring accuracy, [24]
717 cannot be applied to a neural network with multiple input layers (Fig. S5). Ubuntu
718 (version 20.04), Python (version 3.8.10), TensorFlow (version 2.6.0), NVIDIA CUDA
719 toolkit (version 11.4.2), GeForce RTX 3080 GPU with 10 GB memory, and Intel Core
720 i7-10700 CPU with 128 GB memory were used for training and prediction, and the
721 original scripts were modified to run on newer versions of the software. The training
722 method was the same as the original, and the training time was two to four hours. In this
723 study, we focused on the analysis of HFF cells and assessed the training process using
724 all the input data as the test data, which slightly improved the prediction of gene
725 expression. Therefore, we used the trained model for the analysis of insulator-associated
726 DBPs.
727

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

729 **DNA-binding sites of putative insulator-associated DBPs**

730 To identify the direction of DBP-binding motifs in the DNA-binding sites of
731 predicted insulator-associated DBPs, we searched for DBP-binding motifs in open
732 chromatin regions in HFF cells (GSE170964) using the PIQ tool [17] and the position
733 frequency matrices of the DBP-binding motifs. The motif sequences were collected
734 from databases and previous studies, specifically, TRANSFAC (2019.2), JASPAR

735 (2018), UniPROBE (2018), DNA motifs from high-throughput SELEX, DNA motifs
736 from ENCODE DBP ChIP-seq data, and HOCOMOCO versions 9 and 11 [18-23]. To
737 reduce the number of false-positive prediction of DNA-binding sites, only
738 DNA-binding sites that overlapped the ChIP-seq peaks of the same DBP (GTRD
739 database release v20.06 [https://gtrd.biouml.org]) were used for further analyses. The
740 genomic positions of ChIP-seq and DNase-seq data in public databases were converted
741 to those provided in the hg19 version of the human genome using the LiftOver tool.

742 To confirm that CTCF DNA-binding sites were located at chromatin interaction
743 sites, we examined the overlap between CTCF ChIP-seq peaks and the sites of *in situ*
744 ChIA-PET chromatin interactions mediated by CTCF in human foreskin fibroblast
745 (HFF) cells (4D Nucleome data portal,
746 https://data.4dnucleome.org/files-processed/4DNFI73AFN12/). The directional
747 difference between RAD21 DNA-binding sites located at chromatin interaction sites
748 was also examined using ChIA-PET chromatin interactions mediated by RAD21 in
749 human fibroblast from skin (ENCLB147AJZ). RAD21 DNA-binding sites within 100
750 bp of chromatin interaction sites were associated with chromatin interactions. To
751 eliminate DNA-binding sites that overlapped EPI, we obtained ChIP-seq peaks of
752 histone modifications, such as H3K4me3, H3K27ac, and H3K4me1, associated with
753 promoter and enhancer regions in HFF cells (4D Nucleome data portal,
754 4DNFIJOK1BWR, 4DNFI442BQLO, and 4DNFIAGV3X99).

755

756 **Preparation of input data considering the DNA-binding sites of an**
757 **insulator-associated DBP**

758 To prepare the input data of the DNA-binding sites of DBPs in enhancers and
759 promoters for the deep learning analysis, first, a transcript regulated by a DBP bound to
760 enhancers and promoters was predicted based on the single nearest gene association rule
761 (Fig. 2c; McLean et al. [15] Fig. S2c), as we found that the association rule predicted EPIs
762 more accurately than the other rules in our previous study [15]. The maximum length of
763 the regulatory domain for each transcript was set to 1 Mbp upstream and downstream of
764 the TSS. Then, considering the genomic locations and effects of the insulator sites of a
765 DBP on EPIs, the length of a DBP ChIP-seq peak in a 100-base bin was calculated from
766 a TSS to the DNA-binding site of a predicted insulator-associated DBP. To reduce the
767 computational time of the analysis, we first calculated the length of a DBP ChIP-seq
768 peak in the regulatory domains of all transcripts in a 100-base bin. Then, for each
769 interval of 100 bins between a TSS and a DNA-binding site of a predicted
770 insulator-associated DBP in the regulatory domain, we computed the maximum length
771 of each DBP ChIP-seq peak in 100 bins on the basis of the precomputed length of the
772 DBP ChIP-seq peak in each bin. In the case of a set of 100 bins containing the
773 DNA-binding site of a predicted insulator-associated DBP, we computed the maximum
774 lengths from the bin closest to a TSS to the first bin containing the DNA-binding site of

775 the insulator-associated DBP, taking into account the orientation of the DNA-binding
776 site.
777
778 **Prediction of insulator-associated DBPs using DeepLIFT tool**
779 The DeepLIFT tool (version 0.6.13.0) was used to calculate the contribution
780 (DeepLIFT score) of each DNA-binding site to the prediction of the expression level of
781 a transcript [24]. The contribution scores of DNA-binding sites of DBPs in putative
782 enhancers and promoters upstream and downstream of each transcript were normalized
783 by the sum of the absolute values of all contribution scores for the transcript, as the
784 range of contribution scores changes according to the range of the predicted output (i.e.,
785 the expression level of a transcript), as described in a previous study [58]. To estimate the
786 effect of the DNA-binding sites of a predicted insulator-associated DBP on the
787 prediction of transcript levels, positive normalized contribution scores for transcripts
788 with DNA-binding sites in upstream or downstream regions were subjected to
789 comparisons between all combinations of two different orientations (i.e., all
790 combinations of two orientations among FR, RF, forward-forward [FF], and
791 reverse-reverse [RR] orientations, and no orientation [X]: FR vs. RF, FR vs. X, RF vs.
792 X, FF vs. RR, FF vs. X, RR vs. X, FF vs. FR, FF vs. RF, RR vs. FR, and RR vs. RF)
793 among the DNA-binding sites of a predicted insulator-associated DBP by using a
794 statistical test (Mann–Whitney *U* test, FDR-adjusted p value < 0.05 and p value < 10^{-3}).

795 For EPIs that were common to the predictions based on two different orientations of
796 DNA-binding sites of a predicted insulator-associated DBP, the DNA-binding sites of
797 DBPs in putative enhancers and promoters among the common EPIs were omitted in
798 the comparison of positive normalized contribution scores between the two different
799 orientations. Based on the results of the statistical analysis and the mean differences in
800 contribution scores, we predicted an insulator-associated DBP on the basis of the
801 directional bias or lack of orientation bias of DNA-binding sites; for example, to predict
802 an insulator-associated DBP with a DNA-binding site in the FR orientation, we selected
803 a DBP for which the contribution scores among all combinations of FR vs. RF, FR vs.
804 FF, FR vs. RR, and FR vs. X were significant (Mann–Whitney *U* test, FDR-adjusted *p*
805 value < 0.05 and *p* value $< 10^{-3}$) and showed positive (high) mean differences among
806 the contribution scores of all the combinations.
807

808 **Search for insulator-associated DNA-binding sites based on histone modifications**
809 **and gene expression levels**

810 The genomic positions in the BAM files of H3K27me3 and H3K9me3 ChIP-seq
811 data (ENCFF267JBO, ENCFF528UZJ) from HFF cells were converted to the
812 corresponding positions in the hg19 version of the human genome using the CrossMap
813 tool [63] and then converted to WIG files using bam2wig.pl from Bio::ToolBox
814 (https://github.com/tjparnell/biotoolbox) to visualize the data on the UCSC Genome

815 Browser through custom tracks (http://www.genome.ucsc.edu/cgi-bin/hgCustom) and a
816 track hub (https://genome.ucsc.edu/cgi-bin/hgHubConnect). The BAM file of RNA-seq
817 data (GSM941744) of HFF cells was converted to a WIG file using bam2wig.pl in
818 Bio::ToolBox. The genomic positions in the BED file of the boundaries and the
819 distribution of insulation scores identified through a Micro-C assay (4D Nucleome data
820 portal, https://data.4dnucleome.org/files-processed/4DNFI5OZ7TBE/ and
821 4DNFIPMHBA41) were converted to the corresponding positions in the hg19 version
822 of the human genome.
823
824 **Distribution of DNA-binding sites around chromatin interaction sites**
825 We examined the distribution of the DNA-binding sites of predicted
826 insulator-associated DBPs in regions within ±500 bp of the center of a chromatin
827 interaction site using chromatin interaction data from HFF cells obtained via a Micro-C
828 assay (4D Nucleome data portal,
829 https://data.4dnucleome.org/files-processed/4DNFI1MCF9GZ/). For the
830 insulator-associated DBPs other than CTCF, RAD21, and SMC3, the DNA-binding
831 sites that do not overlap with those of CTCF, RND21, and SMC3 were used to examine
832 their distribution around interaction sites. The number of DNA-binding sites of
833 predicted insulator-associated DBPs that cooccurred with DNA-binding sites of known
834 insulator-associated DBPs, such as CTCF, RAD21, and SMC3, was counted in each

835 10-bp bin in the regions within ±500 bp of the center of a chromatin interaction site.

836

837 **Insulator-associated DNA-binding sites near splice sites**

838 Insulator-associated DNA-binding sites were searched within 200 bp of the center
839 of splice sites using the file of the genomic positions of alternative transcripts in the
840 UCSC Genome Browser
841 (https://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/all_mrna.txt.gz ;
842 https://genome.ucsc.edu/cgi-bin/hgTables?db=hg19&hgta_group=rna&hgta_track=mrn
843 a&hgta_table=all_mrna&hgta_doSchema=describe+table+schema). We performed the
844 statistical test to estimate the enrichment of insulator-associated DNA-binding proteins
845 at the splicing sites compared with the other DNA-binding proteins (Mann–Whitney *U*
846 test, p value < 0.05).

847

848 **Statistical analysis**

849 Comparison of the distribution of DeepLIFT scores between different orientations
850 of DNA-binding sites was performed using the Mann-Whitney *U* test with the
851 scipy.stats.mannwhitneyu library. Two-tailed p values were calculated. For multiple
852 comparisons, adjusted p values were calculated by false discovery rate (FDR) using the
853 Benjamini-Hochberg procedure. Z scores for the Mann-Whitney U test were not output
854 by the library. Therefore, we calculated z scores by z-standardizing the U values. The

855 code is made available on GitHub at https://github.com/reposit2/insulator/tree/main.

856

857 **Data availability**

858 Human foreskin fibroblast (HFF) cell RNA-seq data were obtained from GSM941744.
859 Human monocyte RNA-seq data were obtained from the Blueprint DCC portal
860 http://dcc.blueprint-epigenome.eu/#/files at
861 'C0010KB1.transcript_quantification.rsem_grape2_crg.GRCh38.20150622.results'. The
862 RNA-seq data of human mammary epithelial cells (HMEC) were found on ENCODE at
863 UCSC http://genome.ucsc.edu/encode/ at
864 'wgEncodeCshlLongRnaSeqHmecCellPapAlnRep1.bam'. Human neural progenitor cell
865 (NPC) RNA-seq data were obtained from GSM915326. The DNase-seq data of HFF
866 cells were obtained from GSE170964. The *in situ* ChIA-PET chromatin interaction data
867 mediated by CTCF in HFF cells were available at the 4D Nucleome data portal,
868 https://data.4dnucleome.org/files-processed/4DNFI73AFN12/. The ChIP-seq data of
869 H3K4me3, H3K27ac, and H3K4me1 of HFF cells were found at the 4D Nucleome data
870 portal, 4DNFIJOK1BWR, 4DNFI442BQLO, and 4DNFIAGV3X99. The ChIP-seq data
871 of H3K27me3 and H3K9me3 in HFF cells were obtained from the ENCODE portal,
872 ENCFF267JBO and ENCFF528UZJ. The chromatin interaction data and boundary data
873 of HFF cells by the Micro-C were obtained from the 4D Nucleome data portal,
874 https://data.4dnucleome.org/files-processed/4DNFI1MCF9GZ/ and

875 https://data.4dnucleome.org/files-processed/4DNFI5OZ7TBE/. ChIP-seq data of TFs
876 were downloaded from the v20.06 release of the GTRD database
877 (https://gtrd.biouml.org). eQTL data (GTEx_Analysis_v8_eQTL.tar) were downloaded
878 from the v8 release of the GTEx portal (https://gtexportal.org). DNA binding motifs of
879 TFs were collected from TRANSFAC (2019.2), JASPAR (2018)
880 (https://jaspar.genereg.net/), UniPROBE (2018)
881 (http://the_brain.bwh.harvard.edu/uniprobe/), high-throughput SELEX
882 (https://doi.org/10.1016/j.cell.2012.12.009), transcription factor binding sequences of
883 ENCODE TF ChIP-seq data (http://compbio.mit.edu/encode-motifs/), and
884 HOCOMOCO version 9 and 11 (https://autosome.org/HOCOMOCOS/)
885 (https://hocomoco11.autosome.org/). The genomic positions of alternative transcripts in
886 the UCSC Genome Browser
887 (https://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/all_mrna.txt.gz).
888

889 **Code availability**

890 Demonstration codes are made available at
891 https://github.com/reposit2/insulator/tree/main and
892 https://doi.org/10.5281/zenodo.8216164.
893

894 **Figures**

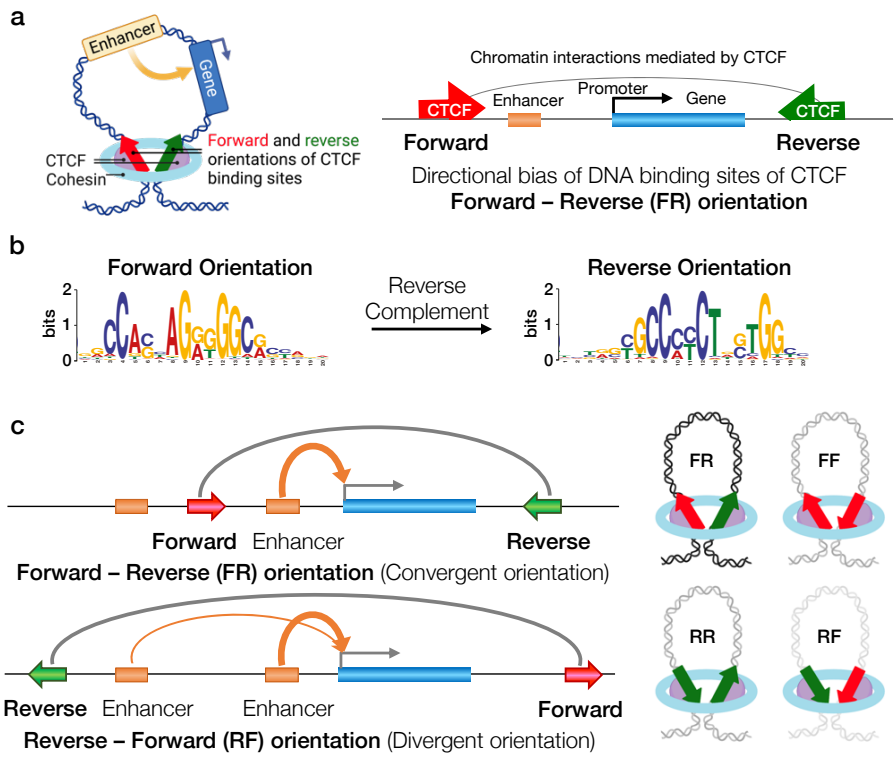
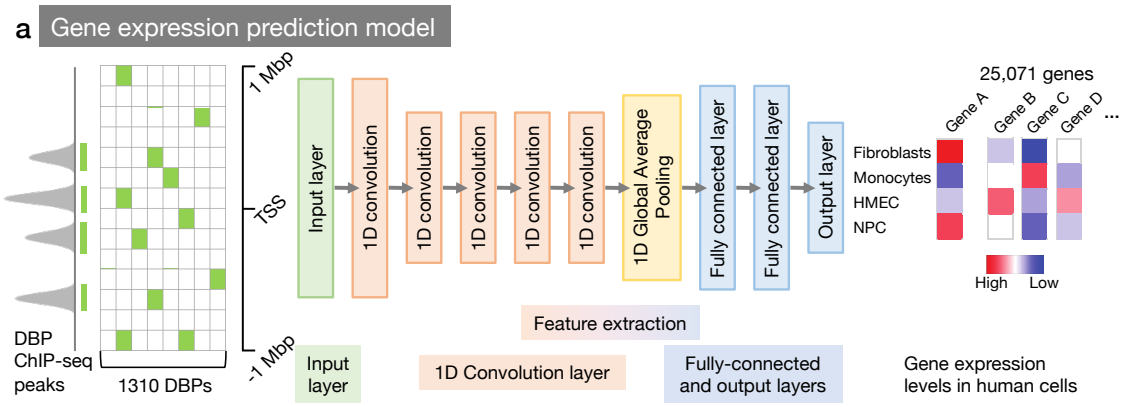


Fig. 1. CTCF insulator function is involved in enhancer–promoter interactions. a CTCF DNA-binding sites with a forward–reverse orientation are frequently found at chromatin interaction sites. The forward and reverse orientations of DNA-binding sites have complementary sequences and are relatively distinct 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.



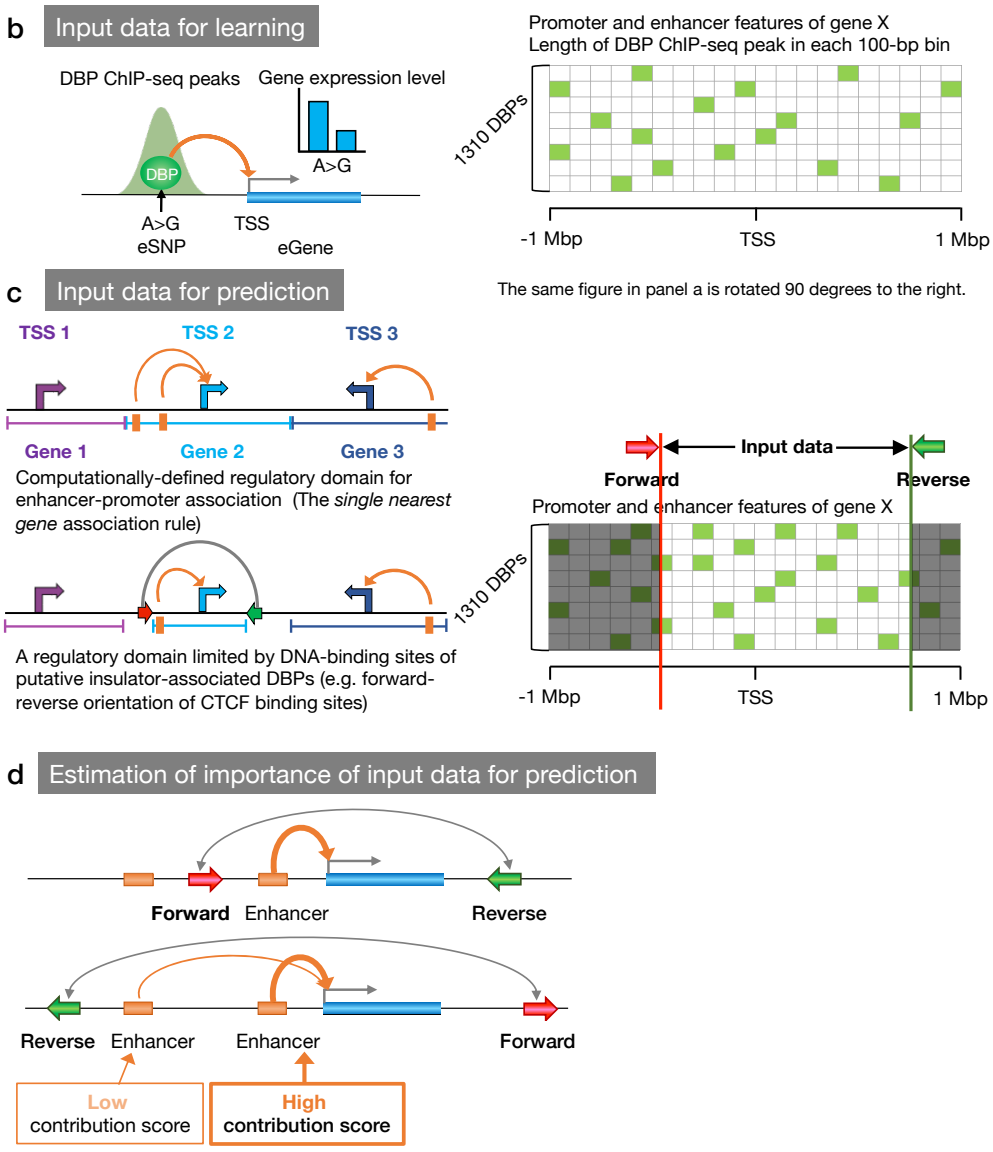
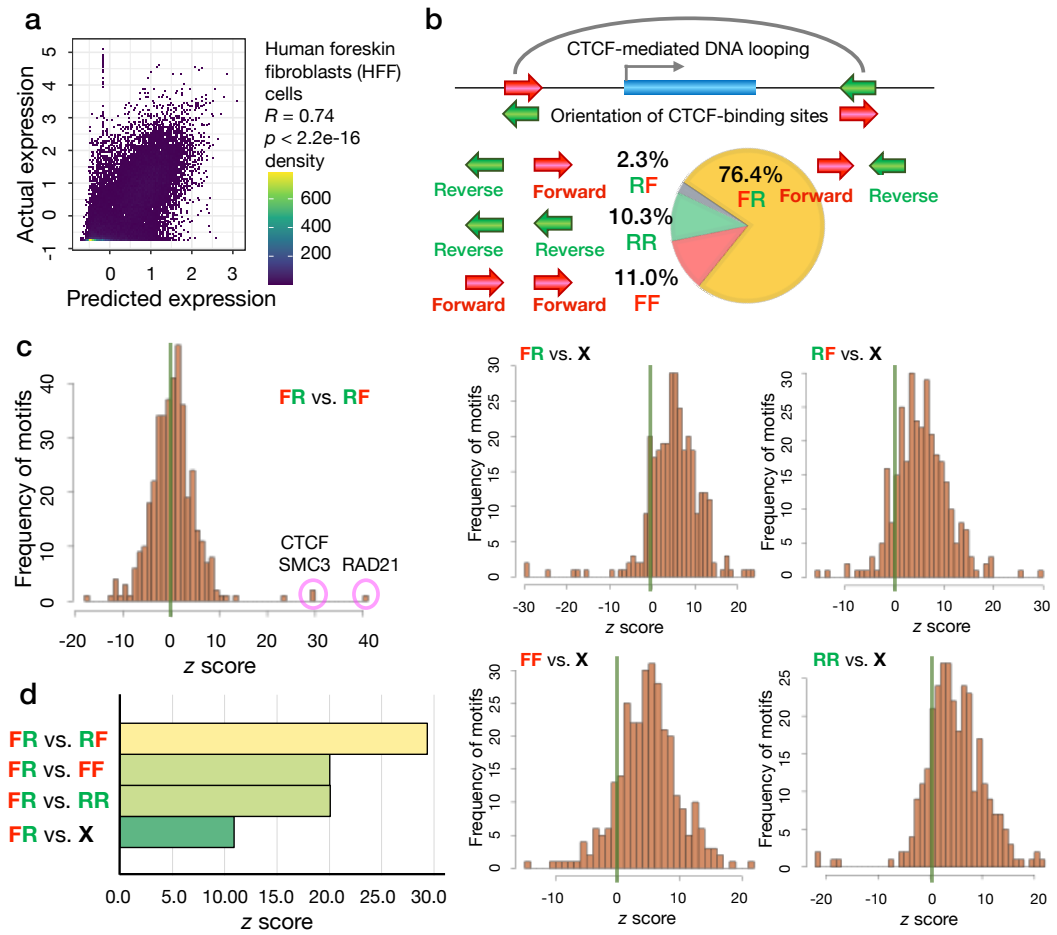


Fig. 2. Overview of the prediction of insulator-associated DBPs. a We used a machine learning method named DEcode, with some modifications. The output data of gene expression levels in human cells were predicted from the input data of genomic locations of DBP binding at promoters and putative enhancers. **b** In the machine learning phase, a DBP binding to a putative enhancer overlapping an eSNP (single

925 nucleotide polymorphism) was associated with a gene (eGene) that was potentially
926 regulated by the DBP based on eSNP and eGene of eQTL data. Based on the association
927 of each gene, the locations of DBP ChIP-seq peaks were binned (green box) and
928 summarized in a matrix as input data: Promoter consists of 30 bins of 100 bp between
929 -2 kbp and 1 kbp region from TSS. Enhancer-gene regulatory interaction regions consist
930 of 200 bins of 10 kbp between -1 Mbp and 1 Mbp region from TSS, not including
931 promoter. **c** In the phase of putative insulator-associated DBP prediction, a DBP binding
932 to a putative enhancer was associated with a gene potentially regulated by the DBP
933 based on a computationally defined regulatory domain without using eQTL data. The
934 *single nearest gene* association rule defines a regulatory domain connecting the
935 midpoints between the TSS of a gene and the TSS of the nearest gene in the upstream
936 and downstream regions. The TSS and regulatory domain of each gene are indicated
937 with a purple, light blue, or blue arrow and a bracketed line, respectively. The regulatory
938 domain of enhancer–promoter association was restricted by the DNA-binding sites (red
939 and green arrows) of a putative insulator-associated DBP in this study. Based on the
940 association of each gene, the locations of DBP ChIP-seq peaks were binned (green box
941 in non-gray regions) and summarized in a matrix and used as input data. **d** Estimation of
942 the directional effect of the DNA-binding sites (red and green arrows) of a putative
943 insulator-associated DBP on the prediction of gene expression levels. A DBP whose

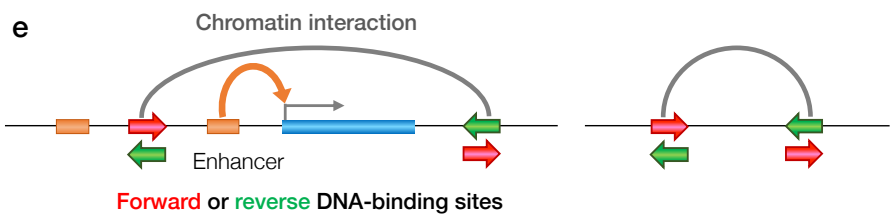
944 binding (orange box) is important for predicting a gene expression level has a high

945 contribution score.



946

947



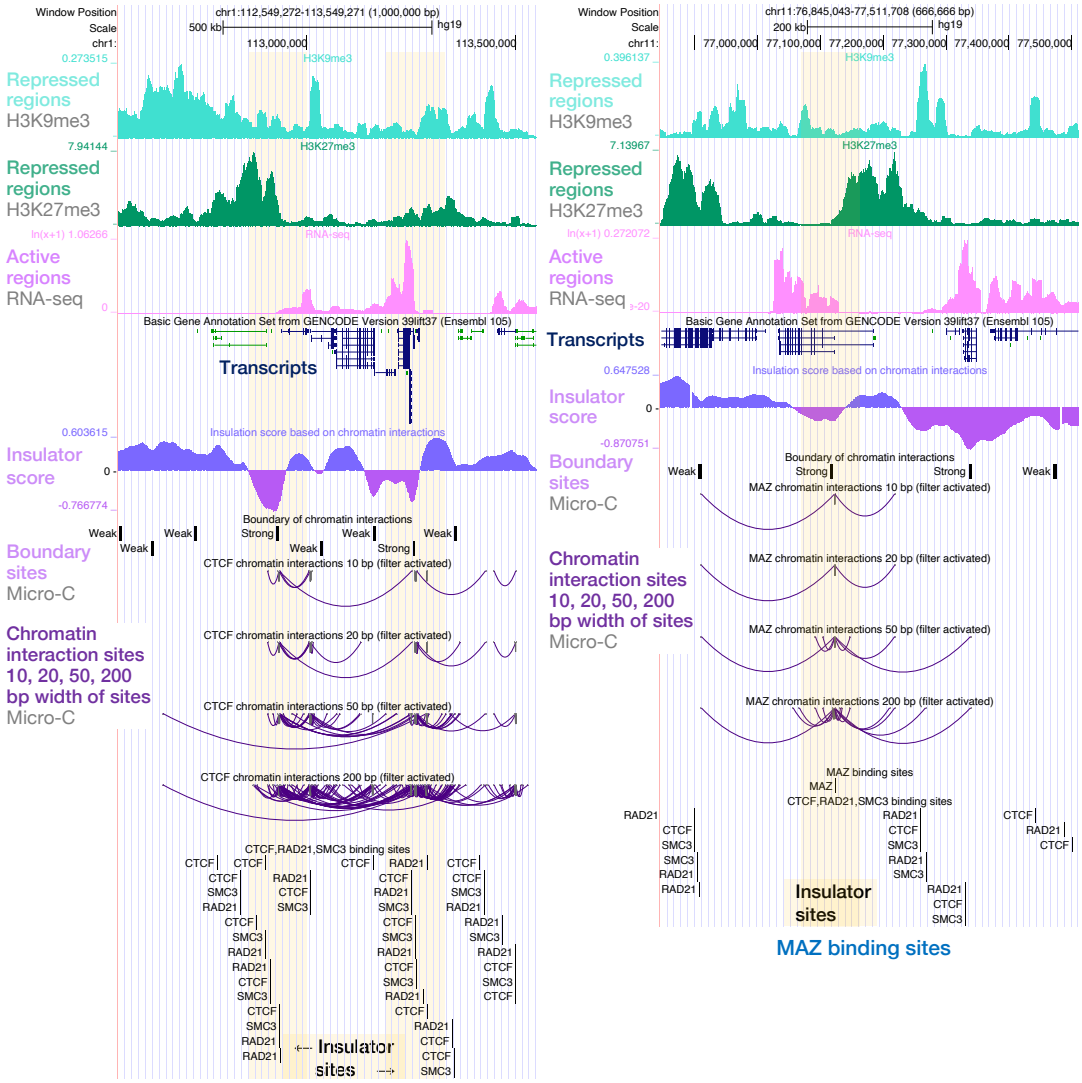
948

949 **Fig. 3. Estimation of the effect of directional regulatory motifs on gene expression.**

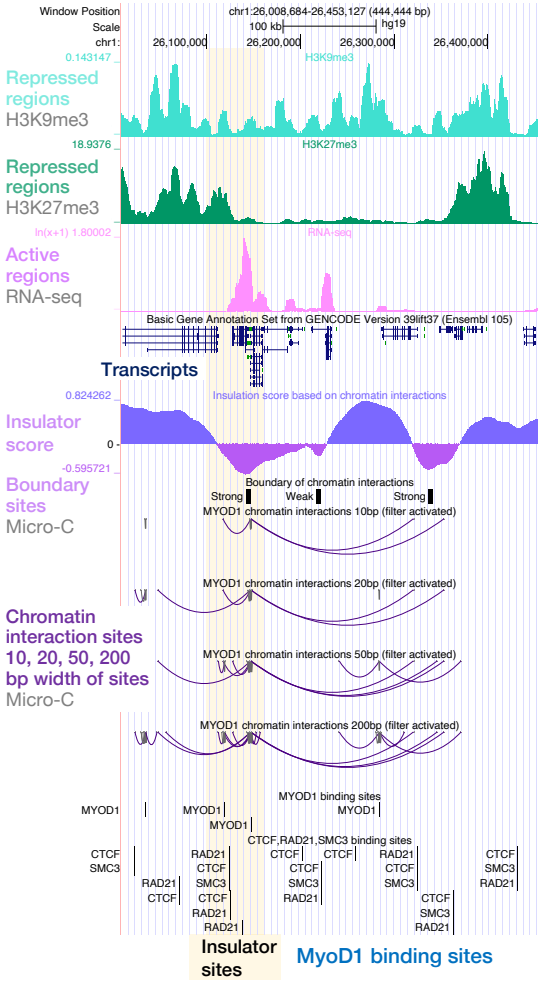
950 **a** The predicted log2-TPM (transcripts per million) values of genes were compared with

951 the actual expression levels in human foreskin fibroblast (HFF) cells using Spearman's
952 rank correlation. **b** An FR orientation of CTCF-binding sites is common (76.4%) at
953 chromatin interaction sites. Forward-forward (FF), reverse-reverse (RR), and RF
954 orientations occur at low frequencies (11.0%, 10.3%, and 2.3%, respectively) in
955 chromatin interaction sites (Guo Y et al. *Cell* 2015) [3]. **c** Distribution of *z* scores from the
956 Mann–Whitney *U* test for DeepLIFT scores between forward–reverse (FR) and
957 reverse–forward (RF) orientations of DNA-binding sites of proteins. Distribution of *z*
958 scores between no orientation (X) and another orientation (FR, RF, FF, or RR) is shown
959 in the right panel. The graphs show the results based on Selection Criteria S2 in
960 Supplementary Fig. 1 and Note. **d** *Z* scores from the Mann–Whitney *U* test for
961 DeepLIFT scores in pairwise comparisons of CTCF-binding sites with four different
962 orientations (FR, RF, FF, and RR) and no orientation (X, i.e., orientation is not
963 distinguished). The differences in *z* scores followed the same trend as the percentages of
964 CTCF-binding site orientations in chromatin interaction sites. **e** The DNA-binding sites
965 of insulator-associated proteins were found at a pair of long-range chromatin interaction
966 sites or at either site.
967
968

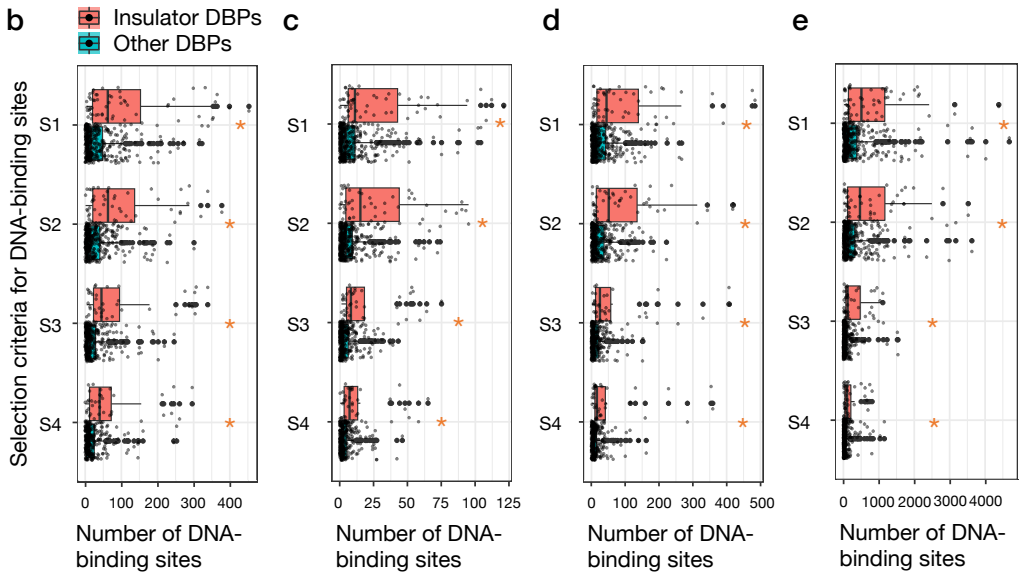
a


Window Position
chr1:112,549,272-113,549,271 (1,000,000 bp)
Scale
500 kb
hg19
chr1:
113,000,000
113,500,000
0.273515
H3K9me3
Repressed regions H3K9me3
7.94144
H3K27me3
Repressed regions H3K27me3
ln(x+1) 1.06266
RNA-seq
Active regions RNA-seq
0
Basic Gene Annotation Set from GENCODE Version 39lift37 (Ensembl 105)
Transcripts
0.603615
Insulation score based on chromatin interactions
Insulator score
0
-0.766774
Boundary of chromatin interactions
Weak
Strong
Boundary sites Micro-C
CTCF chromatin interactions 10 bp (filter activated)
CTCF chromatin interactions 20 bp (filter activated)
CTCF chromatin interactions 50 bp (filter activated)
CTCF chromatin interactions 200 bp (filter activated)
Chromatin interaction sites 10, 20, 50, 200 bp width of sites Micro-C
CTCF,RAD21,SMC3 binding sites
CTCF
RAD21
SMC3
Insulator sites
CTCF, RAD21 and SMC3 binding sites
Window Position
chr11:76,845,043-77,511,708 (666,666 bp)
Scale
200 kb
hg19
chr11:
77,000,000
77,100,000
77,200,000
77,300,000
77,400,000
77,500,000
0.396137
H3K9me3
Repressed regions H3K9me3
7.13967
H3K27me3
Repressed regions H3K27me3
ln(x+1) 0.272072
RNA-seq
Active regions RNA-seq
-20
Basic Gene Annotation Set from GENCODE Version 39lift37 (Ensembl 105)
Transcripts
0.647528
Insulation score based on chromatin interactions
Insulator score
0
-0.870751
Boundary of chromatin interactions
Weak
Strong
Boundary sites Micro-C
MAZ chromatin interactions 10 bp (filter activated)
MAZ chromatin interactions 20 bp (filter activated)
MAZ chromatin interactions 50 bp (filter activated)
MAZ chromatin interactions 200 bp (filter activated)
Chromatin interaction sites 10, 20, 50, 200 bp width of sites Micro-C
MAZ binding sites
MAZ
CTCF,RAD21,SMC3 binding sites
RAD21
CTCF
SMC3
Insulator sites
MAZ binding sites

969


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

970



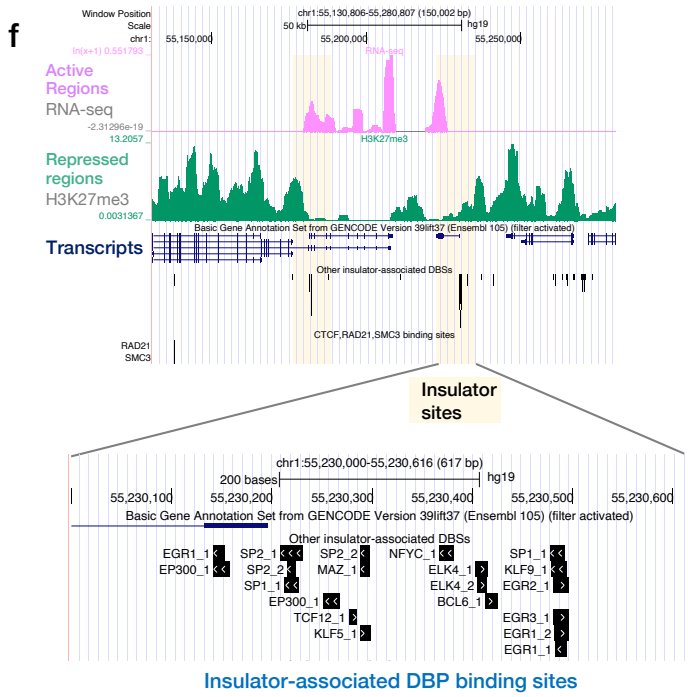


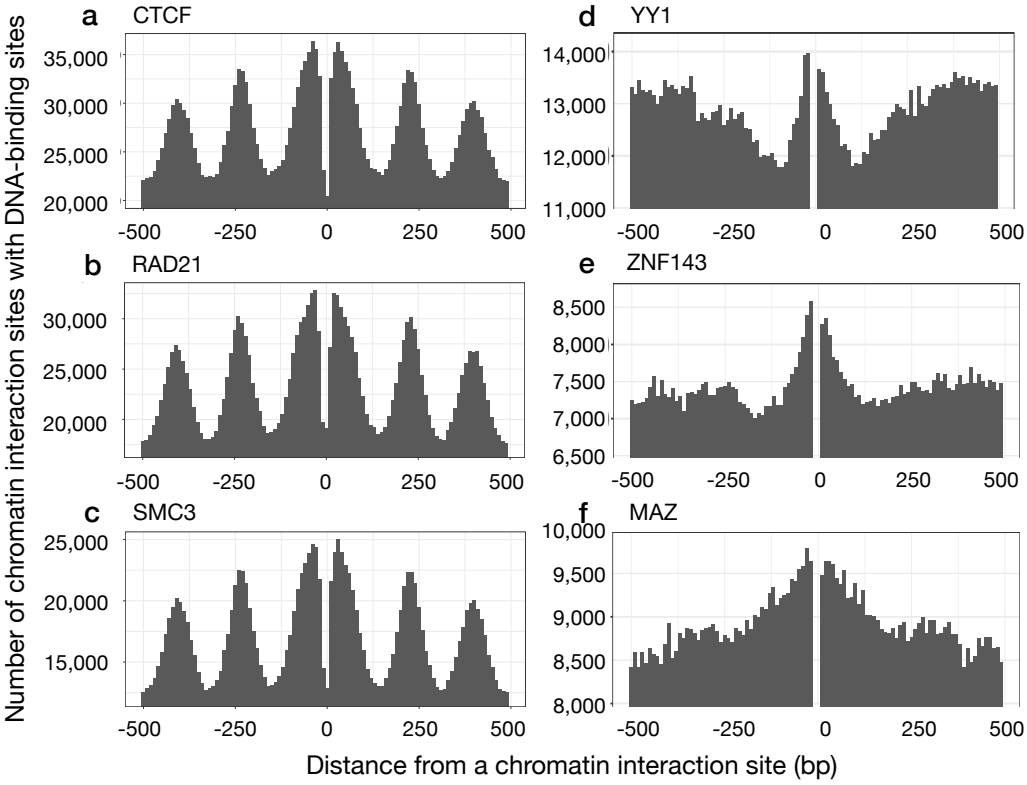
Fig. 4. DNA-binding sites of predicted insulator-associated DBPs located at potential insulator sites. a Distribution of histone modification marks H3K27me3 (green color) and H3K9me3 (turquoise color) and transcript levels (pink color) in the

976 upstream and downstream regions of a potential insulator site (light orange color).
977 These histone modification marks are observed in transcriptionally repressed regions.
978 Insulator sites would be located at the boundary between transcriptionally active and
979 repressed regions (light orange vertical bars). Blue and green lines in the GENCODE
980 track represent transcripts of genes. Insulator score and boundary sites were estimated
981 from chromatin interactions, which are available in 4D Nucleome Data Portal. Dark
982 blue arcs show chromatin interactions within a 10 bp, 20 bp, 50 bp, and 200 bp of the
983 binding sites of CTCF (upper left panel), MAZ (upper right panel), and MyoD1 (lower
984 left panel), respectively. The binding sites of CTCF, RAD21, SMC3 (upper left panel),
985 MAZ (upper right panel), and MyoD1 (lower left panel) are shown as black bars. Homo
986 sapiens (human) genome assembly GRCh37 (hg19) from Genome Reference
987 Consortium was used in panels **a**. **b** Number of DNA-binding sites of predicted
988 insulator-associated DBPs, where H3K27me3 signal intensities differed between the
989 upstream and downstream regions of the DNA-binding site. * p value < 0.01 (Table 2).
990 Boxplots show the median as the centerline, the boxes define the interquartile range
991 (IQR), the whiskers extend up to 1.5 times the IQR, and all points represent outliers
992 from that range. All data points are displayed. **c** Number of DNA-binding sites of
993 predicted insulator-associated DBPs, where H3K27me3 signal intensities and transcript
994 levels differed between the upstream and downstream regions of the DNA-binding site.
995 **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

1007 repressed regions. Insulator sites would be located at the boundary between
1008 transcriptionally active and repressed regions (light orange vertical bars). Blue lines in
1009 the GENCODE track represent transcripts of genes. Chromatin interactions overlapped
1010 with a 200-bp region centered on the binding sites of RBPJ (left panel) and SP1 (right
1011 panel) are shown as dark blue arcs, respectively. The binding sites of RBPJ (left panel),
1012 SP1 (right panel), and CTCF, RAD21, SMC3 (both panels) are shown as black bars.
1013 The DNA-binding sites of insulator-associated DBPs were frequently located near
1014 transcription start and termination sites. Insulator-associated DBPs potentially regulate
1015 transcribed regions differentially repressed in alternative transcripts.
1016



1017

1018 **Fig. 6. DNA-binding sites distributed around a chromatin interaction site. a-f** The

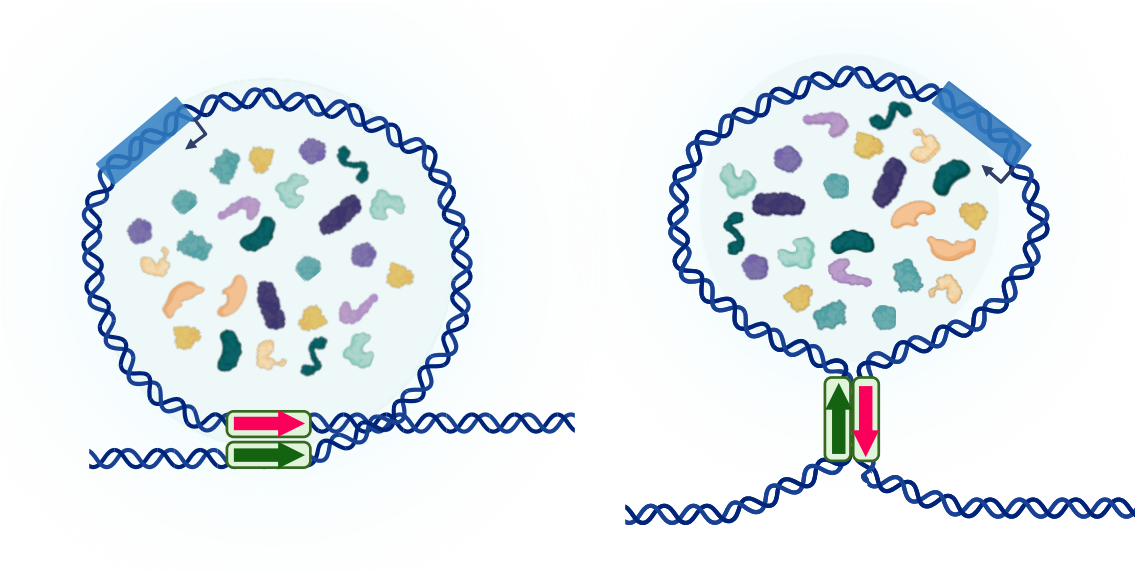
1019 X-axis shows the distance of a DNA-binding site from a chromatin interaction site. The

1020 Y-axis shows the number of chromatin interaction sites with DNA-binding sites within a

1021 ±500-bp region from the center of the interaction site. For YY1, ZNF143, and MAZ,

1022 DNA-binding sites overlapping with those of CTCF, RAD21, and SMC3 were removed.

1023



1024

1025 **Fig. 7. Boundary-pairing (insulator-pairing) model for the regulation of**

1026 **enhancer-promoter interactions.** Molecules bound to insulators physically pair with

1027 their partners, either head-to-head or head-to-tail, with different degrees of specificity at

1028 the termini of TADs in flies [45-47]. Although the experiments do not reveal how partners

1029 find each other, the mechanism unlikely requires loop extrusion. Homologous and

1030 heterologous insulator-insulator pairing interactions are central to the architectural

1031 functions of insulators. The manner of insulator-insulator interactions is

1032 orientation-dependent. Homologous and heterologous interactions between the

1033 insulator-associated DNA-binding proteins potentially occur with the same mechanism

1034 in humans. Chromatin looping could induce the formation of transcriptional

1035 condensates and phase separation as well as cohesin-CTCF-anchored chromatin looping

1036 [40-42]. The figure was created with BioRender.com.

1037

1038 # Tables

1039 **Table 1. Predicted insulator-associated DBPs with directional or nondirectional**

1040 **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

1042 *Z* scores from the Mann–Whitney *U* test for DeepLIFT scores between forward-reverse
1043 (FR) and other orientations such as reverse-forward (RF), forward-forward (FF),
1044 reverse-reverse (RR), and non-orientation (X) of DNA-binding sites of proteins. Mean
1045 differences of DeepLIFT scores. Color scales are shown for each DNA-binding protein
1046 (DBP). Top: Selection Criteria S1 in Supplementary Fig. 1; Middle: Selection Criteria
1047 S2 in Supplementary Fig. 1; Bottom: List of predicted insulator-associated DBPs. DBPs
1048 in a yellow cell are associated with insulators, CTCF, and/or other transcriptional
1049 regulations that have been shown to affect enhancer-promoter interactions (EPIs) in
1050 previous studies.
1051
1052 **Table 2. Number of DNA-binding sites of predicted insulator-associated DBPs at**
1053 **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

1060 DNA-binding sites do not overlap with the ChIP-seq peaks of H3K4me3, H3K4me1, or
1061 H3K27ac; Selection Criteria S4: DNA-binding sites of predicted insulator-associated
1062 DBPs, where the DNA-binding sites overlap a chromatin interaction site in a region
1063 between more than 5-kb upstream and 5-kb downstream of a transcript and the
1064 DNA-binding sites do not overlap the ChIP-seq peaks of H3K4me3, H3K4me1, or
1065 H3K27ac. *P* values were calculated from the Mann–Whitney *U* test.
1066
1067 **Table 3. Pairs of DNA-binding sites colocalize at the same distance from chromatin**
1068 **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

1076 bp of the interaction site. Ratio (DBP1 and 2)/DBP2 is the ratio of the number of
1077 chromatin interaction sites with both DBP1 and DBP2 binding sites to the number of
1078 interaction sites with DBP2 binding sites.
1079
1080

1081 **References**

1082 1 Weintraub, A. S. *et al.* YY1 Is a Structural Regulator of Enhancer-Promoter
1083 Loops. *Cell* **171**, 1573-1588 e1528, doi:10.1016/j.cell.2017.11.008 (2017).
1084 2 Rao, S. S. *et al.* A 3D map of the human genome at kilobase resolution reveals
1085 principles of chromatin looping. *Cell* **159**, 1665-1680,
1086 doi:10.1016/j.cell.2014.11.021 (2014).
1087 3 Guo, Y. *et al.* CRISPR Inversion of CTCF Sites Alters Genome Topology and
1088 Enhancer/Promoter Function. *Cell* **162**, 900-910, doi:10.1016/j.cell.2015.07.038
1089 (2015).
1090 4 Ortabozkoyun, H. *et al.* CRISPR and biochemical screens identify MAZ as a
1091 cofactor in CTCF-mediated insulation at Hox clusters. *Nature genetics* **54**,
1092 202-212, doi:10.1038/s41588-021-01008-5 (2022).
1093 5 Goel, V. Y., Huseyin, M. K. & Hansen, A. S. Region Capture Micro-C reveals
1094 coalescence of enhancers and promoters into nested microcompartments. *Nature*
1095 *genetics*, doi:10.1038/s41588-023-01391-1 (2023).
1096 6 Bailey, S. D. *et al.* ZNF143 provides sequence specificity to secure chromatin
1097 interactions at gene promoters. *Nat Commun* **2**, 6186, doi:10.1038/ncomms7186
1098 (2015).
1099 7 Lis, M. & Walther, D. The orientation of transcription factor binding site motifs
1100 in gene promoter regions: does it matter? *BMC Genomics* **17**, 185,
1101 doi:10.1186/s12864-016-2549-x (2016).
1102 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

1133 binding microarray data on protein-DNA interactions. *Nucleic acids research* **37**,
1134 D77-82, doi:10.1093/nar/gkn660 (2009).
1135 20 Portales-Casamar, E. *et al.* JASPAR 2010: the greatly expanded open-access
1136 database of transcription factor binding profiles. *Nucleic acids research* **38**,
1137 D105-110, doi:10.1093/nar/gkp950 (2010).
1138 21 Jolma, A. *et al.* DNA-binding specificities of human transcription factors. *Cell*
1139 **152**, 327-339, doi:10.1016/j.cell.2012.12.009 (2013).
1140 22 Kheradpour, P. & Kellis, M. Systematic discovery and characterization of
1141 regulatory motifs in ENCODE TF binding experiments. *Nucleic acids research*
1142 **42**, 2976-2987, doi:10.1093/nar/gkt1249 (2014).
1143 23 Kulakovskiy, I. V. *et al.* HOCOMOCO: towards a complete collection of
1144 transcription factor binding models for human and mouse via large-scale
1145 ChIP-Seq analysis. *Nucleic acids research* **46**, D252-D259,
1146 doi:10.1093/nar/gkx1106 (2018).
1147 24 Shrikumar, A., Greenside, P. & Kundaje, A. in *Proceedings of the 34th*
1148 *International Conference on Machine Learning* Vol. 70 (eds Precup Doina &
1149 Teh Yee Whye) 3145--3153 (PMLR, Proceedings of Machine Learning Research,
1150 2017).
1151 25 Lee, B. H., Wu, Z. & Rhie, S. K. Characterizing chromatin interactions of
1152 regulatory elements and nucleosome positions, using Hi-C, Micro-C, and
1153 promoter capture Micro-C. *Epigenetics Chromatin* **15**, 41,
1154 doi:10.1186/s13072-022-00473-4 (2022).
1155 26 Cuddapah, S. *et al.* Global analysis of the insulator binding protein CTCF in
1156 chromatin barrier regions reveals demarcation of active and repressive domains.
1157 *Genome research* **19**, 24-32, doi:10.1101/gr.082800.108 (2009).
1158 27 Huang, S., Li, X., Yusufzai, T. M., Qiu, Y. & Felsenfeld, G. USF1 recruits
1159 histone modification complexes and is critical for maintenance of a chromatin
1160 barrier. *Mol Cell Biol* **27**, 7991-8002, doi:10.1128/MCB.01326-07 (2007).
1161 28 Crane, E. *et al.* Condensin-driven remodelling of X chromosome topology
1162 during dosage compensation. *Nature* **523**, 240-244, doi:10.1038/nature14450

1163 (2015).
1164 29 Noda, Y., Okada, S. & Suzuki, T. Regulation of A-to-I RNA editing and stop
1165 codon recoding to control selenoprotein expression during skeletal myogenesis.
1166 *Nat Commun* **13**, 2503, doi:10.1038/s41467-022-30181-2 (2022).
1167 30 Matharu, N. K. & Ahanger, S. H. Chromatin Insulators and Topological
1168 Domains: Adding New Dimensions to 3D Genome Architecture. *Genes (Basel)*
1169 **6**, 790-811, doi:10.3390/genes6030790 (2015).
1170 31 Reyes, A. & Huber, W. Alternative start and termination sites of transcription
1171 drive most transcript isoform differences across human tissues. *Nucleic acids*
1172 *research* **46**, 582-592, doi:10.1093/nar/gkx1165 (2018).
1173 32 Hua, P. *et al.* Defining genome architecture at base-pair resolution. *Nature* **595**,
1174 125-129, doi:10.1038/s41586-021-03639-4 (2021).
1175 33 Xiao, T., Li, X. & Felsenfeld, G. The Myc-associated zinc finger protein (MAZ)
1176 works together with CTCF to control cohesin positioning and genome
1177 organization. *Proc Natl Acad Sci U S A* **118**, doi:10.1073/pnas.2023127118
1178 (2021).
1179 34 Ishihara, K., Oshimura, M. & Nakao, M. CTCF-dependent chromatin insulator
1180 is linked to epigenetic remodeling. *Molecular cell* **23**, 733-742,
1181 doi:10.1016/j.molcel.2006.08.008 (2006).
1182 35 Vierbuchen, T. *et al.* AP-1 Transcription Factors and the BAF Complex Mediate
1183 Signal-Dependent Enhancer Selection. *Molecular cell* **68**, 1067-1082 e1012,
1184 doi:10.1016/j.molcel.2017.11.026 (2017).
1185 36 Gong, Q. H., McDowell, J. C. & Dean, A. Essential role of NF-E2 in remodeling
1186 of chromatin structure and transcriptional activation of the epsilon-globin gene
1187 in vivo by 5' hypersensitive site 2 of the beta-globin locus control region. *Mol*
1188 *Cell Biol* **16**, 6055-6064, doi:10.1128/MCB.16.11.6055 (1996).
1189 37 Davidson, I. F. & Peters, J. M. Genome folding through loop extrusion by SMC
1190 complexes. *Nat Rev Mol Cell Biol* **22**, 445-464,
1191 doi:10.1038/s41580-021-00349-7 (2021).
1192 38 Karollus, A., Mauermeier, T. & Gagneur, J. Current sequence-based models

capture gene expression determinants in promoters but mostly ignore distal enhancers. *Genome Biol* **24**, 56, doi:10.1186/s13059-023-02899-9 (2023).

39 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).

40 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).

41 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).

42 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).

43 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).

44 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).

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

46 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).

47 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).

48 Aljahani, A. *et al.* Analysis of sub-kilobase chromatin topology reveals

1223 nano-scale regulatory interactions with variable dependence on cohesin and
1224 CTCF. *Nat Commun* **13**, 2139, doi:10.1038/s41467-022-29696-5 (2022).
1225 49 Hsieh, T. S. *et al.* Enhancer-promoter interactions and transcription are largely
1226 maintained upon acute loss of CTCF, cohesin, WAPL or YY1. *Nature genetics*
1227 **54**, 1919-1932, doi:10.1038/s41588-022-01223-8 (2022).
1228 50 Guerin, T. M., Barrington, C., Pobegalov, G., Molodtsov, M. I. & Uhlmann, F.
1229 An extrinsic motor directs chromatin loop formation by cohesin. *EMBO J* **43**,
1230 4173-4196, doi:10.1038/s44318-024-00202-5 (2024).
1231 51 Wang, R. *et al.* MyoD is a 3D genome structure organizer for muscle cell
1232 identity. *Nat Commun* **13**, 205, doi:10.1038/s41467-021-27865-6 (2022).
1233 52 Snead, W. T. & Gladfelter, A. S. The Control Centers of Biomolecular Phase
1234 Separation: How Membrane Surfaces, PTMs, and Active Processes Regulate
1235 Condensation. *Molecular cell* **76**, 295-305, doi:10.1016/j.molcel.2019.09.016
1236 (2019).
1237 53 Tang, X. *et al.* The PTM profiling of CTCF reveals the regulation of 3D
1238 chromatin structure by O-GlcNAcylation. *Nat Commun* **15**, 2813,
1239 doi:10.1038/s41467-024-47048-3 (2024).
1240 54 Phillips, J. E. & Corces, V. G. CTCF: master weaver of the genome. *Cell* **137**,
1241 1194-1211, doi:10.1016/j.cell.2009.06.001 (2009).
1242 55 Kumar, S., Kaur, S., Seem, K., Kumar, S. & Mohapatra, T. Understanding 3D
1243 Genome Organization and Its Effect on Transcriptional Gene Regulation Under
1244 Environmental Stress in Plant: A Chromatin Perspective. *Front Cell Dev Biol* **9**,
1245 774719, doi:10.3389/fcell.2021.774719 (2021).
1246 56 Dekker, J. & Misteli, T. Long-Range Chromatin Interactions. *Cold Spring Harb*
1247 *Perspect Biol* **7**, a019356, doi:10.1101/cshperspect.a019356 (2015).
1248 57 Spurrell, C. H., Dickel, D. E. & Visel, A. The Ties That Bind: Mapping the
1249 Dynamic Enhancer-Promoter Interactome. *Cell* **167**, 1163-1166,
1250 doi:10.1016/j.cell.2016.10.054 (2016).
1251 58 Kim, D. S. *et al.* The dynamic, combinatorial cis-regulatory lexicon of
1252 epidermal differentiation. *Nature genetics* **53**, 1564-1576,

doi:10.1038/s41588-021-00947-3 (2021).

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

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

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

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

63 Zhao, H. *et al.* CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics (Oxford, England)* **30**, 1006-1007, doi:10.1093/bioinformatics/btt730 (2014).

Acknowledgements

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

Author contributions

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.