Title

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

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

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

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

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

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

[*] Corresponding author.

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

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

Abstract

Insulator proteins act as a barrier of enhancer–promoter interactions (EPIs). 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 99 directional and minor nondirectional motifs in human fibroblast cells that corresponded to 23 DBPs related to insulator function, CTCF, and/or other types of chromosomal transcriptional regulation reported in previous studies. The estimated CTCF orientation bias was consistently proportional to CTCF orientation bias observed in chromatin interactions. Furthermore, we observed that the motifs could regulate transcribed regions differentially repressed in alternative transcripts. These analyses reveal that the directional bias of motifs is associated with insulators, chromosome conformation and alternative transcription, potentially mediated by structural interactions. These findings contribute to elucidate novel transcriptional regulations.

Introduction

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

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

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

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

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

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

76 To predict chromatin interactions and associated DBPs, computational methods
77 have been developed [11] [12]. However, these methods were not designed to identify the
78 binding motifs of DBPs associated with insulators. As for regulatory motifs, Xie et al. [13]

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

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

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

105 ## Results

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

107 Insulator proteins act as a barrier of EPIs (Fig. 1a, b, c). To predict the EPIs
108 between DNA-binding sites distal to transcription start sites (TSSs) and the
109 transcription initiation complex at TSSs, we identified the interactions between the
110 DNA-binding sites in putative enhancer regions and genes based on expression
111 quantitative trait locus (eQTL) data showing the association between a SNP (single
112 nucleotide polymorphism) and a gene expression level (Fig. 2a, b). To predict gene
113 expression levels based on DNA-binding sites of DBPs around the TSS of a gene, we
114 modified a machine learning tool reported in a previous study to train with
115 DNA-binding sites distal as well as proximal to the TSS [14]. To avoid the cellular
116 heterogeneity found in tissues, we used gene expression data from human cell lines,
117 including human foreskin fibroblasts (HFFs), monocytes, human mammary epithelial
118 cells (HMECs), and neural progenitor cells. The machine learning tool was trained to

predict gene expression levels in the cell lines based on DNA-binding sites of DBPs around TSSs of genes. The predicted expression levels were highly consistent with the actual expression levels (Spearman's rho = 0.74, 0.71, 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 test data in the four cell types) without using the data on the RNA-binding proteins and miRNAs obtained via the original machine learning method (Fig. 3a; Fig. S1a) [14].

Then, to investigate the DNA-binding sites of putative insulator-associated DBPs, we obtained ChIP-seq data from the Gene Transcription Regulation Database (GTRD) (v20.06 release), and searched the ChIP-seq peaks of candidate DBPs for associated binding motifs. First, the DNA-binding sites of a given DBP in open chromatin regions were predicted using the Protein Interaction Quantitation (PIQ) tool with DNase-seq data and DBP-binding motif sequences, which were obtained from databases and publications, specifically, TRANSFAC (2019.2), JASPAR (2018), UniPROBE (2018), DNA motifs from high-throughput SELEX, DNA motifs from ENCODE DBP ChIP-seq data, and HOCOMOCO versions 9 and 11 [16] [17-22]. Then, to reduce the number of false-positive predictions of DNA-binding sites, we selected the DNA-binding sites that overlapped with the ChIP-seq peaks of the same DBP. Among the 1,356 DBPs with available ChIP-seq data, the DBP-binding motif sequences were available for 715 DBPs.

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

140 To investigate the effect of directional differences in the DNA-binding sites of
141 putative insulator-associated DBPs on EPIs and gene expression levels, we computed
142 DeepLIFT scores for use as inputs for deep learning analyses. The DeepLIFT method
143 calculates a score representing the contribution of the input data to the prediction. By
144 comparing the DeepLIFT scores of the DNA-binding sites to the ChIP-seq data, we
145 compared the effects of putative insulator sites, including insulator-associated
146 DNA-binding sites, on the prediction of gene expression levels according to the
147 directional and positional differences among the DNA-binding sites (Fig. 2d).

148 At the sites of distal CTCF-mediated chromatin interactions identified *in situ* via
149 ChIA-pet, 76.4% of the CTCF DNA-binding sites showed an FR orientation, also
150 known as a convergent orientation (Fig. 3b) [3]. To investigate the effect of orientation of
151 CTCF DNA-binding sites on the DeepLIFT scores of the DNA-binding sites in putative
152 enhancer and open chromatin regions, we searched CTCF DNA-binding sites with an
153 FR or RF orientation in the upstream and downstream regions of TSSs and predicted
154 EPIs based on the enhancer–promoter association rule (Fig. 2c). We then compared the
155 distribution of the DeepLIFT scores and mean DeepLIFT scores between the CTCF
156 DNA-binding sites with FR and RF orientations and observed a significant difference in
157 the distribution of DeepLIFT scores (Mann–Whitney U test, FDR-adjusted p value <
158 0.05 and p value < 10^{-3}; Fig. 3c, d and Table S1). Our method was able to compare the

159 effect of directional difference of CTCF DNA-binding site on the prediction of gene
160 expression level.

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

179 and thus reduce the number of false-positive DBP predictions, we set the p value
180 threshold to $< 10^{-3}$. We confirmed that all the DBPs with p values $< 10^{-3}$ constituted the
181 subset of DBPs with false discovery rate (FDR)-adjusted p values < 0.05 (included in
182 Table S1).
183

184 **Discovery of directional and nondirectional DNA-binding sites of**
185 **insulator-associated DBPs**

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

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

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

219 regions.

220 A total of 99 insulator-associated DBPs were predicted based on the stringent
221 criterion of the DNA-binding sites of a given DBP showing either a specific directional
222 bias or no directional bias, as determined by comparing the four possible orientations
223 (FR, RF, FF, and RR) and the absence of directional bias (X). We found that 23 DBPs
224 exhibited insulator and CTCF-associated functions and/or chromatin regulatory
225 functions that affected EPIs, as reported in previous studies: 13 DBPs were associated
226 with CTCF (Table S6). The 23 DBPs were multifunctional proteins, and their known
227 functions were transcriptional regulator, repressor, chromatin regulator, and interaction
228 with other proteins including enzymes that facilitate epigenetic modifications. These
229 findings suggested that CTCF regulated EPIs by interacting with other DBPs, and that
230 DBPs with other functions were also associated with the regulation of EPIs and
231 chromatin conformation.

232

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

235 To confirm the predicted insulator-associated DBPs via another method and select
236 their DNA-binding sites associated with insulators, we first analyzed previously
237 reported histone modification data to identify differences between the upstream and
238 downstream regions of the DNA-binding sites. CTCF bound to sites located between

239 repressed regions with H3K27me3 marks and active regions with H2AK5ac marks,
240 which showed different signal intensities between the upstream and downstream regions
241 of CTCF DNA-binding sites [23]. The transcription factor USF1 exhibits a barrier
242 function, in which histone modification factors are recruited to separate adjacent regions
243 into activating and repressive chromatin domains [24]. We selected H3K27me3 histone
244 modifications, associated with the repression of gene expression, for comparison
245 between the upstream and downstream regions of the DNA-binding sites of predicted
246 insulator-associated DBPs using H3K27me3 ChIP-seq data because H3K27me3 histone
247 marks occupy long genomic regions and are thus easily compared between genomic
248 regions. We searched for DNA-binding sites where the mean H3K27me3 signal
249 intensities differed by more than 2-fold between the upstream and downstream regions
250 of the DNA-binding site. The higher mean signal was found to be greater than the first
251 quartile of all H3K27me3 signal intensities because some H3K27me3 signals exhibited
252 a low intensity and appeared noisy. Thus, the criteria were sufficient to clearly
253 distinguish modification status among genomic regions.

254 We further counted the numbers of DNA-binding sites with different H3K27me3
255 signal intensities between the upstream and downstream regions of a DNA-binding site
256 and found that known insulator-associated DBPs, such as CTCF, RAD21, and SMC3,
257 generally showed the greatest numbers of DNA-binding sites (Table S4). Via statistical
258 analysis, we calculated the median and mean numbers of DNA-binding sites showing

259 different H3K27me3 signal intensities between upstream and downstream regions for
260 the DNA-binding sites of both predicted insulator-associated DBPs and other DBPs.
261 The median and mean numbers of DNA-binding sites were significantly larger among
262 predicted insulator-associated DBPs than other DBPs (Mann–Whitney U test, p value <
263 0.05) (Table 2, Table S4, Fig. 4a, b). We confirmed that a larger number of the
264 DNA-binding sites of known insulator-associated DBPs were located between activated
265 and repressed regions than non-insulator-associated DBPs.

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

276 In the analysis of chromatin interaction data, insulation scores were calculated
277 based on the analysis of the boundary regions of chromatin interactions [25]. We also
278 examined the overlap of DNA-binding sites of predicted insulator-associated DBPs with

279 the boundary regions, and again found that a significantly larger number of the
280 DNA-binding sites of predicted insulator-associated DBPs were located in the boundary
281 regions than other DBPs (Table 2, Fig. 4d).

282 To determine whether the DNA-binding sites of predicted insulator-associated
283 DBPs potentially act as insulators or cofactors of insulators, we investigated the
284 genomic locations of the DNA-binding sites on the UCSC Genome Browser (Fig. S2).
285 We searched for DNA-binding sites located between the H3K27me3 marks and
286 transcribed regions of genes, which were primarily located at the edges of H3K27me3
287 marks and were not found near the binding sites of CTCF, RAD21, and SMC3. For each
288 DBP, we looked for insulator sites associated with the DNA-binding sites in the genome
289 browser and selected primarily three DNA-binding sites for each DBP (Fig. S2). We
290 discovered DNA-binding sites in the expected genomic regions of all DBPs, suggesting
291 that these DNA-binding sites are associated with insulator functions. Genome browser
292 analysis is useful for selecting DNA-binding sites for experimental validation and
293 further study.

294 While preparing Fig. 4a and S2, we noticed that several DNA-binding sites of
295 insulator-associated DBPs were found in the same genomic locus, close to each other on
296 the genome. These DBPs potentially act together to form an insulator site. In addition,
297 some transcribed regions in the genome were differentially repressed in alternative
298 transcripts and overlapped with a H3K27me3 mark that was separated with the

299 DNA-binding site of an insulator-associated DBP between the repressed and active
300 regions (Fig. 5, S2a, b). Furthermore, the DNA-binding sites of insulator-associated
301 DBPs were frequently located near transcription start and termination sites. Alternative
302 transcription start and termination sites have been reported to explain the majority of
303 tissue-specific exon usage rather than alternative splicing[26]. The DNA-binding sites of
304 insulator-associated DBPs could affect the form of transcripts and the genomic location
305 of transcription start and termination sites. The differences in the DNA-binding sites
306 and H3K27me3 marks between tissues could contribute to the tissue-specific exon
307 usage of alternative transcripts.
308
309 **Distribution of DNA-binding sites around chromatin interaction sites**
310 In this study, we aimed to predict chromatin- and insulator-associated DBPs. To
311 estimate how many DBPs are involved in chromatin interactions and examine the
312 distribution and association of DNA-binding sites around chromatin interaction sites,
313 the DNA-binding sites of DBPs in regions within ±500 bp of the centers of chromatin
314 interaction sites were identified in HFF cells. As an improved method for chromatin
315 interaction analysis, Micro-C assays enabled us to identify chromatin interactions at a
316 higher resolution than can be achieved via Hi-C analysis. The distribution of CTCF,
317 RAD21, and SMC3 DNA-binding sites showed a characteristic pattern around
318 chromatin interaction sites, with the highest peaks located around interaction sites (Fig.

319 6a, b, c). The peaks of the DNA-binding sites of YY1 and ZNF143, which are known to
320 be involved in chromatin interactions, were distributed around chromatin interaction
321 sites. The DNA-binding sites of YY1 were also broadly distributed in regions within
322 ±500 bp of interaction sites, suggesting that this DBP is involved in various functions in
323 addition to chromatin interactions (Fig. 6d, e). MAZ, which exhibits insulator function
324 independent of CTCF[27], indicated the peak of distribution of MAZ DNA-binding sites
325 around interaction sites (Fig. 6f).

326 Among the 228 DNA-binding sites, more than 1,000 of the DNA-binding sites in
327 open chromatin regions in HFF cells overlapped with the ChIP-seq peaks of the DBPs,
328 and the peaks of almost all the DBPs (approximately 90%) were located around
329 interaction sites. Furthermore, we found DBPs with sharp peaks around interaction sites,
330 DBPs with broad peaks around interaction sites, and DBPs with peaks similar to those
331 of CTCF, RAD21, and SMC3 around interaction sites, suggesting that these DBPs may
332 be associated with CTCF, RAD21, and/or SMC3 (Table S7; Fig. S3). Among the
333 predicted insulator-associated DBPs, MeCP2 and PRDM9 exhibited fewer than 1,000
334 DNA-binding sites in HFF cells, and when a DBP has fewer than 1,000 DNA-binding
335 sites, the specific distribution of DNA-binding sites around chromatin interaction sites
336 may not be evident (Fig. S3). Approximately 90% of the DBPs with more than 1,000
337 DNA-binding sites had peaks around interaction sites. This indicated that the DBPs
338 interacted with other genomic regions and that the DNA-binding sites of predicted

339 insulator-associated DBPs were involved in chromatin interactions.

340

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

342 Since the DNA-binding sites of the predicted insulator-associated DBPs may be
343 located close to those of known insulator-associated DBPs, such as CTCF, RAD21,
344 and/or SMC3, and, thus, be presumed to act as insulators, we next counted the number
345 of DNA-binding sites in each 10-bp bin within ±500 bp from the center of a chromatin
346 interaction site and compared these numbers between all possible pairs of DBPs. We
347 found that the number of DNA-binding sites of pairs of DBPs located within the same
348 chromatin interaction site tended to be the highest within the same bin distance from the
349 interaction site, suggesting that the resolution of the chromatin interaction data was high
350 enough to determine the distance of DNA-binding sites of DBPs from interaction sites
351 and distinguish two DNA-bound DBPs located close to each other (Table S5). For this
352 analysis, the DNA-binding sites that overlapped with the ChIP-seq peaks of a DBP were
353 selected to reduce the number of false-positive DNA-binding site predictions, and the
354 number of DNA-binding sites per DBP may therefore have been underestimated.
355 Among the 228 DBPs, 91 pairs of 70 DBPs colocalized around chromatin interaction
356 sites when at least 50% of the DNA-binding sites of a pair of DBPs colocalized in a bin
357 with the same distance from the interaction site (Table S5). CTCF showed a high rate of
358 colocalization of DNA-binding sites with both RAD21 (39%) and SMC3 (81%) in bins

359 within ±500 bp from the center of a chromatin interaction site, whereas YY1 and
360 ZNF143 showed low rates of colocalization of DNA-binding site with RAD21 (1.5%
361 and 5.5%, respectively) and CTCF (2.3% and 4.8%, respectively) (Table 3). For CTCFL
362 (CCCTC-binding factor like), RXRA, PPARG, NEUROD1, and ZBTB7A, more than
363 10% of DNA-binding sites colocalized with CTCF, RAD21, and/or SMC3 in bins
364 within the same distance from chromatin interaction sites, suggesting that the insulator
365 function of the DBPs was associated with known insulator DBPs. For the other DBPs
366 among the 228 DBPs, less than 10% of DNA-binding sites colocalized with CTCF,
367 RAD21, and/or SMC within a bin. Most of the predicted insulator-associated DBPs did
368 not have the colocalization of DNA-binding sites and were not in close proximity to
369 known insulator-associated DBPs such as CTCF, RAD21, and SMC3; that is, they were
370 not within a 10-bp bin in the regions ±500 bp from the center of an interaction site.

371 To further examine the possibility that known insulator-associated DBPs interact
372 indirectly with the other predicted insulator-associated DBPs, we investigated the
373 overlap of DNA-binding sites with the DBP ChIP-seq peaks of CTCF, RAD21, and
374 SMC3 and found that a high percentage (> 98%) of the DNA-binding sites of the three
375 DBPs overlapped with their ChIP-seq peaks in all three examined combinations.
376 However, the DNA-binding sites of almost all other predicted insulator-associated
377 DBPs overlapped with less than 50% of the ChIP-seq peaks of the three known DBPs,
378 even though the ChIP-seq data were collected from various cell and tissue types. These

379 results suggested that the predicted insulator-associated DBPs play different functional
380 roles than the known DBPs.

381 Although a DNA-binding-site distribution pattern around chromatin interaction
382 sites similar to those of CTCF, RAD21, and SMC3 was observed for DBPs such as
383 BACH2, FOS, ATF3, NFE2, and MAFK, less than 1% of the DNA-binding sites of the
384 latter set of DBPs colocalized with CTCF, RAD21, or SMC3 in a single bin. However,
385 BACH2, FOS, ATF3, NFE2, and MAFK had DNA-binding sites that colocalized with
386 each other (Table 3). These results suggested that the two different groups of
387 colocalized DBPs were independently involved in the same chromatin-related function
388 such as chromatin remodeling [28 29 30].

389

390 **Discussion**

391 Here, we present the first, to our knowledge, computational method and
392 genome-wide analysis aimed at detecting insulator-associated DBPs. Transcriptional
393 regulation mediated by the binding of DBPs at sites located far from TSSs was
394 modelled using our deep learning method based on eQTL data. Through this approach,
395 the effects of the DNA-binding sites of insulator-associated DBPs on EPIs and gene
396 expression were estimated, and insulator-associated DBPs were predicted. We further
397 identified DBPs, including known insulator-associated DBPs that exhibit
398 insulator-associated functions and regulate transcription, thereby affecting EPIs reported

399 in previous studies. The DNA-binding sites of the predicted insulator-associated DBPs
400 showed orientation bias of their DNA-binding sites; unexpectedly, the significant
401 differences in the orientation bias of CTCF DNA-binding sites based on DeepLIFT
402 scores (contribution scores) were found to be consistent with the percent differences in
403 the orientation of CTCF DNA-binding sites at CTCF-mediated chromatin interaction
404 sites. The DNA-binding sites of the predicted insulator-associated DBPs were located at
405 potential insulator sites more frequently than those of other DBPs, according to the
406 differences in H3K27me3 signal intensities and transcript levels. We further observed
407 that only a small percentage of the DNA-binding sites of the insulator-associated DBPs
408 colocalized with the DNA-binding sites of known insulator-associated DBPs such as
409 CTCF, RAD21, and SMC3 around chromatin interaction sites. We discovered that
410 DBPs other than the known insulator DBPs and the directionality of their DNA-binding
411 sites are associated with insulators. Our novel computational method enabled this study,
412 which will contribute to further promote the functional and structural analyses of
413 insulator-associated DBPs effectively through experimental analyses.

414 Our deep learning analysis accurately predicted the percent differences in the FR,
415 FF, RR, and RF orientations of CTCF DNA-binding sites at chromatin interaction sites,
416 considering the genomic locations of DNA-binding sites of putative insulator-associated
417 DBPs. The other predicted insulator-associated DBPs also showed orientation bias of
418 their DNA-binding sites. These results suggest that the directional bias of DNA-binding

419 sites of insulator-associated DBPs may be involved in insulator function and chromatin
420 regulation through structural interactions among DBPs, other proteins, DNAs, and
421 RNAs. For example, the N-terminal amino acids of CTCF have been shown to interact
422 with RAD21 in chromatin loops [8-10]. For another example, although sequence-based
423 predictions of EPIs underestimated actual interactions, an activity-by-contact (ABC)
424 approach, based on a simple distance model, performed well, suggesting that factors in
425 addition to sequence influence EPIs [31,32]. Furthermore, CTCF-mediated chromatin
426 looping induces phase separation between the inner and outer regions of the loops,
427 resulting in the formation of transcriptional condensates and insulator function [33-35].
428 Therefore, structures that result in phase separation, rather than functional differences in
429 DBPs, may be responsible for insulator functions. Other types of chromatin regulation
430 are also expected to be related to the structural interactions of molecules. Our analyses
431 will allow us to reveal novel structural roles of DBPs for insulator functions, chromatin
432 regulation, and phase separation and to advance their experimental analysis.

433 High-resolution chromatin interaction data from a Micro-C assay indicated that
434 most of the predicted insulator-associated DBPs showed DNA-binding-site distribution
435 peaks around chromatin interaction sites, suggesting that these DBPs are involved in
436 chromatin interactions. However, due to the high resolution of a Micro-C assay, it is
437 difficult to obtain all chromatin interaction data at a reasonable and available cost, while
438 a Region Capture Micro-C assay focuses on a specific genomic region to increase the

coverage of chromatin interactions. The coverage of chromatin interactions will change according to the experimental data of a Micro-C assay. In addition, chromatin interactions include other types of interactions that are not associated with EPIs and insulator functions [36-38]. Therefore, we analyzed and predicted insulator-associated DBPs without using chromatin interaction data. Our deep learning analysis successfully learned EPIs and accurately predicted the percent differences in the FR, FF, RR, and RF orientations of CTCF DNA-binding sites at chromatin interaction sites.

The analysis of many contribution scores (i.e., DeepLIFT scores in this study) for directional DNA-binding sites of putative insulator-associated DBPs in the prediction of transcript expression levels successfully predicted the percent differences in the FR, FF, RR, and RF orientations of CTCF DNA-binding sites at chromatin interaction sites based on the contribution scores of DNA-binding sites in putative enhancer and open chromatin regions. Some contribution scores may not be accurate and thus the prediction of insulator-associated DBPs or DNA-binding sites based on a small number of contribution scores may be erroneous. To compare contribution scores computed based on different pairs of input and output data with our deep learning model, we normalized contribution scores for each pair of input and output data as previously reported [39]. The analysis of many contribution scores will be used to estimate the effects of other molecules and functions involved in gene regulation.

One of the next computational challenges will be to predict and narrow the

459 potential set of insulator-associated DNA-binding sites of multifunctional DBPs. To
460 confirm the machine learning prediction and to identify insulator-associated
461 DNA-binding sites in the genome for further experimental study, information on histone
462 modifications such as H3K27me3 and transcript levels will be beneficial, and other
463 epigenomic data can be used for this purpose. As a reference for an application of this
464 study, Liu et al. focused on CTCF-binding sites and performed various assays in human
465 and mouse cell cultures and in mice with the goal of applying these sites in gene therapy
466 [40]. The analysis of insulator-associated DBPs may contribute to the development of
467 technologies to regulate gene expression. In conclusion, our novel computational
468 method discovered insulator-associated DBPs, including known DBPs, and revealed
469 that the directional bias of insulator-associated DBPs is potentially involved in
470 chromatin looping, phase separation, transcriptional condensates and insulator functions
471 through the structural interactions of the DBPs. This will lead to a general and global
472 understanding of transcriptional regulation and gene expression.
473

474 ## Methods

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

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

477 To predict DNA-binding sites that also act as insulator sites in the human genome,
478 we modified a previously developed deep learning method designed to predict the

transcript expression level of genes in tissues and cell types based on the DNA-binding sites of DBPs in promoter regions and other DNA regions (Fig. 2a; Fig. S4) [14]. To reduce the influence of cellular heterogeneity observed when analyzing tissues, we used the gene expression data from four distinct cell types where various experimental data are available in public databases: HFFs (GSM941744), monocytes (Blueprint RNA-seq FPKM [fragments per kilobase of exon per million reads mapped] data 'C0010KB1.transcript_quantification.rsem_grape2_crg.GRCh38.20150622.results' in the Blueprint DCC portal at http://dcc.blueprint-epigenome.eu/#/files), human mammary epithelial cells (HMECs) (Encyclopedia of DNA Elements at UCSC http://genome.ucsc.edu/encode/, 'wgEncodeCshlLongRnaSeqHmecCellPapAlnRep1.bam'), and neural progenitor cells (NPCs) (GSM915326, ENCFF529SIO). For gene expression data, FPKM values representing the expression levels of genes were retrieved or calculated based on GENCODE V19 gene annotations in the ENCODE portal database [41] (https://www.encodeproject.org/) using the RSeQC tool [42] (http://rseqc.sourceforge.net/). We calculated log2-fold-transformed FPKM values (adding 0.25 to prevent results of negative infinity), median log2-FPKM values across the four cell types, and log2-fold changes relative to the median for all cell types according to the original deep learning method (transcriptome data processing as described by Tasaki et al.) [14].

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

500 To investigate the effect of insulator sites on EPIs and gene expression, the
501 DNA-binding sites in putative enhancer and open chromatin regions located away from
502 promoter regions were added as input data in the deep learning method. To identify the
503 DNA-binding sites of DBPs in promoter and enhancer regions, ChIP-seq data of the
504 DBPs were used and were updated to GTRD database release v20.06
505 (https://gtrd.biouml.org) [43]. The binding site of a DBP regulating the expression level of
506 each transcript was selected using eQTL data from GTEx portal release v8
507 (https://gtexportal.org). If a DNA-binding site was located within a 100-bp region
508 around a single-nucleotide polymorphism (SNP) of an eQTL, we assumed that the DBP
509 regulated the expression of the transcript corresponding to the eQTL. If the SNP of an
510 eQTL was not located on the same chromosome as the transcribed region of the eQTL
511 (i.e., *trans*-eQTL), the eQTL data were not used. The genomic positions of the eQTLs
512 were converted to the corresponding positions in the hg19 version of the human genome
513 using the LiftOver tool (http://hgdownload.soe.ucsc.edu/downloads.html#liftover).

514 The original input data of DNA-binding sites in promoter regions used for the
515 deep learning analysis consisted of the lengths of a DBP ChIP-seq peaks in each
516 100-base bin in a region between 2 kbp upstream and 1 kbp downstream of the
517 transcription start site (TSS) of each gene or transcript [14]. To add the input data of
518 DNA-binding sites of DBPs in more upstream and downstream regions of a TSS and

reduce the number of bins, we first computed the length of a DBP ChIP-seq peak in a 100-base bin and then selected the maximum length in each interval of 100 bins in a region between 1 megabase upstream and downstream of a TSS. This step outputted the 200 longest DBP ChIP-seq peaks in the examined regions.

Training step for machine learning analysis

In the training of the modified deep learning model, we used the same test, training, and validation data as in the original method: all 2,705 genes encoded on chromosome 1 as test data, training data (22,251 genes), and validation data (2,472 genes) randomly selected from the other genes [14]. The input data for promoters and putative enhancer regions were concatenated into a total of 230 bins of the lengths of 1,311 DBP ChIP-seq peaks, consisting of 30 bins for promoters and 200 bins for putative enhancers, because the DeepLIFT tool 'RevealCancel rule on dense layers and the Rescale rule on conv layers (genomics default)', with high scoring accuracy, [44] cannot be applied to a neural network with multiple input layers (Fig. S4). Ubuntu (version 20.04), Python (version 3.8.10), TensorFlow (version 2.6.0), NVIDIA CUDA toolkit (version 11.4.2), GeForce RTX 3080 GPU with 10 GB memory, and Intel Core i7-10700 CPU with 128 GB memory were used for training and prediction, and the original scripts were modified to run on newer versions of the software. The training method was the same as the original, and the training time was two to four hours. In this

539 study, we focused on the analysis of HFF cells and assessed the training process using
540 all the input data as the test data, which slightly improved the prediction of gene
541 expression. Therefore, we used the trained model for the analysis of insulator-associated
542 DBPs.
543

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

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

546 To identify the direction of DBP-binding motifs in the DNA-binding sites of
547 predicted insulator-associated DBPs, we searched for DBP-binding motifs in open
548 chromatin regions in HFF cells (GSE170964) using the PIQ tool [16] and the position
549 frequency matrices of the DBP-binding motifs. The motif sequences were collected
550 from databases and previous studies, specifically, TRANSFAC (2019.2), JASPAR
551 (2018), UniPROBE (2018), DNA motifs from high-throughput SELEX, DNA motifs
552 from ENCODE DBP ChIP-seq data, and HOCOMOCO versions 9 and 11 [17-22]. To
553 reduce the number of false-positive prediction of DNA-binding sites, only
554 DNA-binding sites that overlapped the ChIP-seq peaks of the same DBP (GTRD
555 database release v20.06 [https://gtrd.biouml.org]) were used for further analyses. The
556 genomic positions of ChIP-seq and DNase-seq data in public databases were converted
557 to those provided in the hg19 version of the human genome using the LiftOver tool.

558 To confirm that CTCF DNA-binding sites were located at chromatin interaction

559 sites, we examined the overlap between CTCF ChIP-seq peaks and the sites of *in situ*
560 ChIA-pet chromatin interactions mediated by CTCF in human foreskin fibroblast (HFF)
561 cells (4D Nucleome data portal,
562 https://data.4dnucleome.org/files-processed/4DNFI73AFN12/). To eliminate
563 DNA-binding sites that overlapped EPI, we obtained ChIP-seq peaks of histone
564 modifications, such as H3K4me3, H3K27ac, and H3K4me1, associated with promoter
565 and enhancer regions in HFF cells (4D Nucleome data portal, 4DNFIJOK1BWR,
566 4DNFI442BQLO, and 4DNFIAGV3X99).
567

568 **Preparation of input data considering the DNA-binding sites of an**
569 **insulator-associated DBP**

570 To prepare the input data of the DNA-binding sites of DBPs in enhancers and
571 promoters for the deep learning analysis, first, a transcript regulated by a DBP bound to
572 enhancers and promoters was predicted based on the single nearest gene association rule
573 (Fig. 2c; McLean et al. [15] Fig. S2c), as we found that the association rule predicted EPIs
574 more accurately than the other rules in our previous study [15]. The maximum length of
575 the regulatory domain for each transcript was set to 1 Mbp upstream and downstream of
576 the TSS. Then, considering the genomic locations and effects of the insulator sites of a
577 DBP on EPIs, the length of a DBP ChIP-seq peak in a 100-base bin was calculated from
578 a TSS to the DNA-binding site of a predicted insulator-associated DBP. To reduce the

579 computational time of the analysis, we first calculated the length of a DBP ChIP-seq
580 peak in the regulatory domains of all transcripts in a 100-base bin. Then, for each
581 interval of 100 bins between a TSS and a DNA-binding site of a predicted
582 insulator-associated DBP in the regulatory domain, we computed the maximum length
583 of each DBP ChIP-seq peak in 100 bins on the basis of the precomputed length of the
584 DBP ChIP-seq peak in each bin. In the case of a set of 100 bins containing the
585 DNA-binding site of a predicted insulator-associated DBP, we computed the maximum
586 lengths from the bin closest to a TSS to the first bin containing the DNA-binding site of
587 the insulator-associated DBP, taking into account the orientation of the DNA-binding
588 site.
589

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

591 The DeepLIFT tool (version 0.6.13.0) was used to calculate the contribution
592 (DeepLIFT score) of each DNA-binding site to the prediction of the expression level of
593 a transcript [44]. The contribution scores of DNA-binding sites of DBPs in putative
594 enhancers and promoters upstream and downstream of each transcript were normalized
595 by the sum of the absolute values of all contribution scores for the transcript, as the
596 range of contribution scores changes according to the range of the predicted output (i.e.,
597 the expression level of a transcript), as described in a previous study [39]. To estimate the
598 effect of the DNA-binding sites of a predicted insulator-associated DBP on the

599 prediction of transcript levels, positive normalized contribution scores for transcripts
600 with DNA-binding sites in upstream or downstream regions were subjected to
601 comparisons between all combinations of two different orientations (i.e., all
602 combinations of two orientations among FR, RF, forward-forward [FF], and
603 reverse-reverse [RR] orientations, and no orientation [X]: FR vs. RF, FR vs. X, RF vs.
604 X, FF vs. RR, FF vs. X, RR vs. X, FF vs. FR, FF vs. RF, RR vs. FR, and RR vs. RF)
605 among the DNA-binding sites of a predicted insulator-associated DBP by using a
606 statistical test (Mann–Whitney *U* test, FDR-adjusted p value < 0.05 and p value < 10^{-3}).
607 For EPIs that were common to the predictions based on two different orientations of
608 DNA-binding sites of a predicted insulator-associated DBP, the DNA-binding sites of
609 DBPs in putative enhancers and promoters among the common EPIs were omitted in
610 the comparison of positive normalized contribution scores between the two different
611 orientations. Based on the results of the statistical analysis and the mean differences in
612 contribution scores, we predicted an insulator-associated DBP on the basis of the
613 directional bias or lack of orientation bias of DNA-binding sites; for example, to predict
614 an insulator-associated DBP with a DNA-binding site in the FR orientation, we selected
615 a DBP for which the contribution scores among all combinations of FR vs. RF, FR vs.
616 FF, FR vs. RR, and FR vs. X were significant (Mann–Whitney *U* test, FDR-adjusted p
617 value < 0.05 and p value < 10^{-3}) and showed positive (high) mean differences among
618 the contribution scores of all the combinations.

619

620 **Search for insulator-associated DNA-binding sites based on histone modifications**

621 **and gene expression levels**

622 The genomic positions in the bigWig file of H3K27me3 ChIP-seq data

623 (ENCFF854VTY) from HFF cells were converted to the corresponding positions in the

624 hg19 version of the human genome using the LiftOver tool and then converted to a

625 bedGraph file using the UCSC bigWigToBedGraph tool to visualize the data on the

626 UCSC Genome Browser through custom tracks

627 (http://www.genome.ucsc.edu/cgi-bin/hgCustom). The BAM file of RNA-seq data

628 (GSM941744) from HFF cells was converted to a WIG file using bam2wig.pl from

629 Bio::ToolBox. The genomic positions in the BED file of the boundaries identified

630 through a Micro-C assay (4D Nucleome data portal,

631 https://data.4dnucleome.org/files-processed/4DNFI5OZ7TBE/) were converted to the

632 corresponding positions in the hg19 version of the human genome.

633

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

635 We examined the distribution of the DNA-binding sites of predicted

636 insulator-associated DBPs in regions within ±500 bp of the center of a chromatin

637 interaction site using chromatin interaction data from HFF cells obtained via a Micro-C

638 assay (4D Nucleome data portal,

639 https://data.4dnucleome.org/files-processed/4DNFI1MCF9GZ/). The number of
640 DNA-binding sites of predicted insulator-associated DBPs that cooccurred with
641 DNA-binding sites of known insulator-associated DBPs, such as CTCF, RAD21, and
642 SMC3, was counted in each 10-bp bin in the regions within ±500 bp of the center of a
643 chromatin interaction site.
644
645 **Statistical analysis**
646 Comparison of the distribution of DeepLIFT scores between different orientations
647 of DNA-binding sites was performed using the Mann-Whitney *U* test with the
648 scipy.stats.mannwhitneyu library. Two-tailed *p* values were calculated. For multiple
649 comparisons, adjusted *p* values were calculated by false discovery rate (FDR) using the
650 Benjamini-Hochberg procedure. *Z* scores for the Mann-Whitney *U* test were not output
651 by the library. Therefore, we calculated *Z* scores by z-standardizing the *U* values. The
652 code is made available on GitHub at https://github.com/reposit2/insulator/tree/main.
653
654 **Data availability**
655 Human foreskin fibroblast (HFF) cell RNA-seq data were obtained from GSM941744.
656 Human monocyte RNA-seq data were obtained from the Blueprint DCC portal
657 http://dcc.blueprint-epigenome.eu/#/files at
658 'C0010KB1.transcript_quantification.rsem_grape2_crg.GRCh38.20150622.results'. The

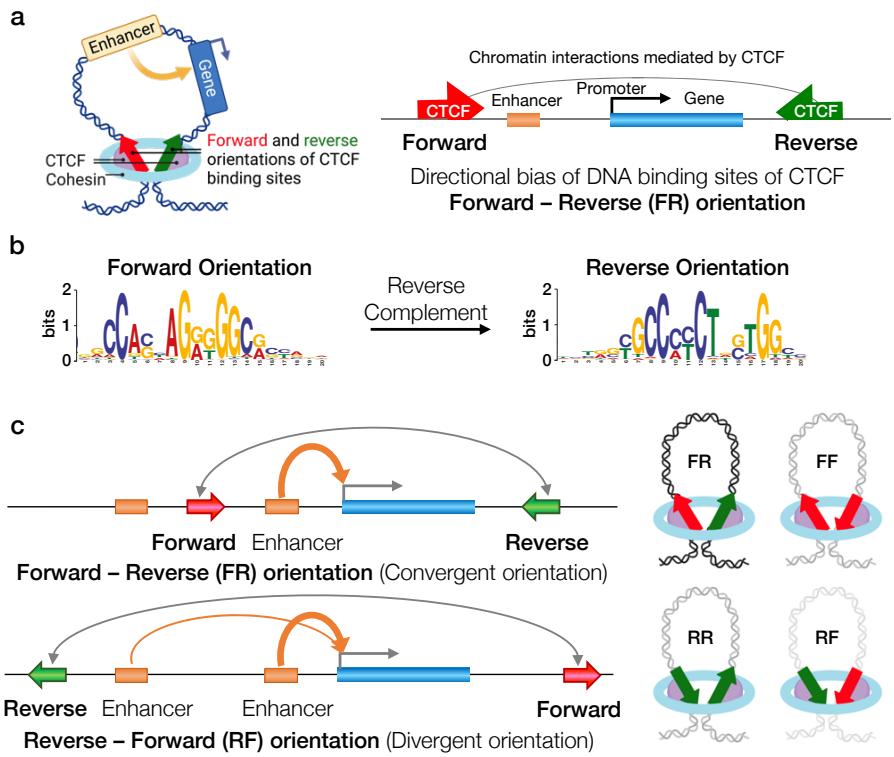
659 RNA-seq data of human mammary epithelial cells (HMEC) were found on ENCODE at
660 UCSC http://genome.ucsc.edu/encode/ at
661 'wgEncodeCshlLongRnaSeqHmecCellPapAlnRep1.bam'. Human neural progenitor cell
662 (NPC) RNA-seq data were obtained from GSM915326. The DNase-seq data of HFF
663 cells were obtained from GSE170964. The *in situ* ChIA-pet chromatin interaction data
664 mediated by CTCF in HFF cells were available at the 4D Nucleome data portal,
665 https://data.4dnucleome.org/files-processed/4DNFI73AFN12/. The ChIP-seq data of
666 H3K4me3, H3K27ac, and H3K4me1 of HFF cells were found at the 4D Nucleome data
667 portal, 4DNFIJOK1BWR, 4DNFI442BQLO, and 4DNFIAGV3X99. The ChIP-seq data
668 of H3K27me3 in HFF cells were obtained from the ENCODE portal, ENCFF854VTY.
669 The chromatin interaction data and boundary data of HFF cells by the Micro-C were
670 obtained from the 4D Nucleome data portal,
671 https://data.4dnucleome.org/files-processed/4DNFI1MCF9GZ/ and
672 https://data.4dnucleome.org/files-processed/4DNFI5OZ7TBE/. ChIP-seq data of TFs
673 were downloaded from the v20.06 release of the GTRD database
674 (https://gtrd.biouml.org). eQTL data were downloaded from the v8 release of the GTEx
675 portal (https://gtexportal.org). DNA binding motifs of TFs were collected from
676 TRANSFAC (2019.2), JASPAR (2018) (https://jaspar.genereg.net/), UniPROBE (2018)
677 (http://the_brain.bwh.harvard.edu/uniprobe/), high-throughput SELEX
678 (https://doi.org/10.1016/j.cell.2012.12.009), transcription factor binding sequences of

ENCODE TF ChIP-seq data (http://compbio.mit.edu/encode-motifs/), and HOCOMOCO version 9 and 11 (https://autosome.org/HOCOMOCOS/) (https://hocomoco11.autosome.org/).

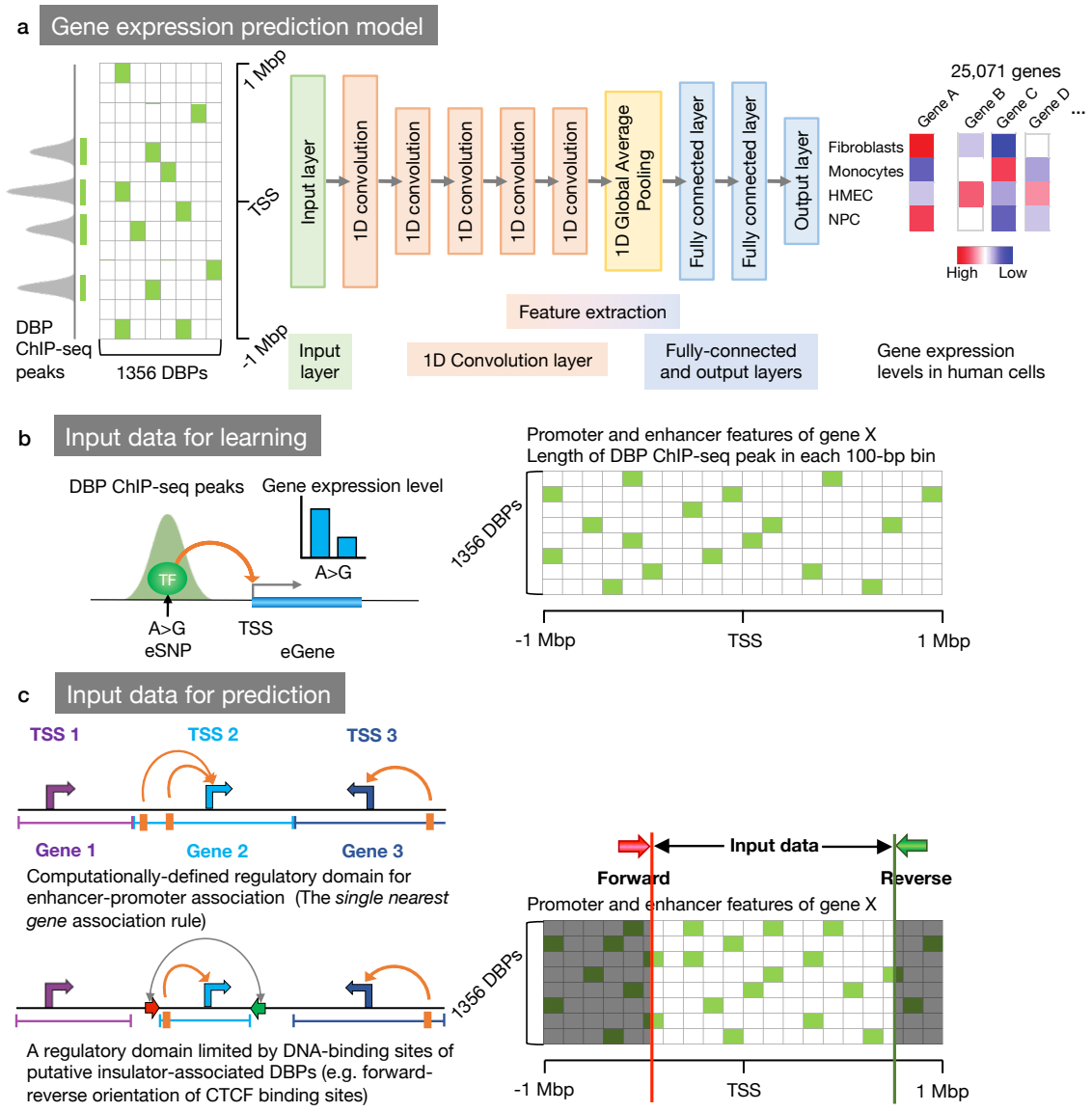
Code availability

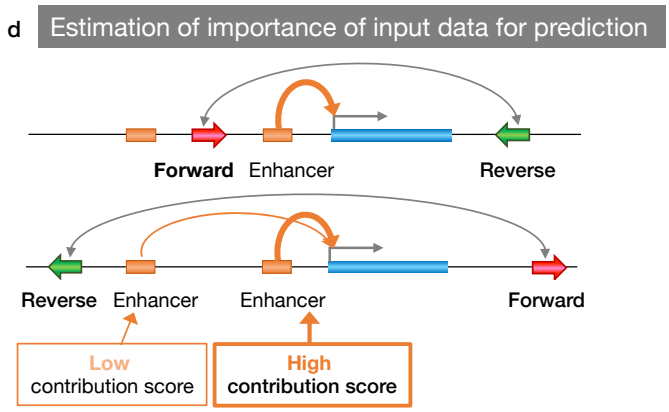
Demonstration codes are made available at https://github.com/reposit2/insulator/tree/main and https://doi.org/10.5281/zenodo.8216164.

Figures



690 **Fig. 1. CTCF insulator function is involved in enhancer–promoter interactions. a**
691 CTCF DNA-binding sites with a forward–reverse orientation are frequently found at
692 chromatin interaction sites. The forward and reverse orientations of DNA-binding sites
693 have complementary sequences and are relatively distinct in orientation according to
694 reference genome sequences. **b** Position weight matrices (PWMs) of CTCF binding
695 motifs. The forward and reverse orientations of the binding motifs are complementary
696 on a DNA strand. A sequence logo is a graphical representation of the sequence
697 conservation of nucleotides, and consists of a stack of letters at each position. The
698 relative sizes of the letters indicate their frequency in the sequences (Wikipedia
699 'Sequence logo'). **c** CTCF can block the interaction between an enhancer and a
700 promoter (or the transcriptional start site [TSS] of a gene), thereby restricting the
701 activity of an enhancer to a certain functional domain (region between red and green
702 arrows). The red and green arrows represent the DNA-binding sites of a putative
703 insulator-associated DBP (e.g., forward–reverse orientation of the DNA-binding sites of
704 CTCF in the figure). The numbers of DNA-binding sites with forward–reverse [FR],
705 reverse–forward [RF], forward-forward [FF], and reverse-reverse [RR] orientations
706 differ, as indicated by the intensity of the chromatin loop (right panel). Images of
707 double-stranded DNA loops were adapted from "Transcriptional Regulation by CTCF
708 and Cohesin" at BioRender.com (2023). Retrieved from
709 https://app.biorender.com/biorender-templates.


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

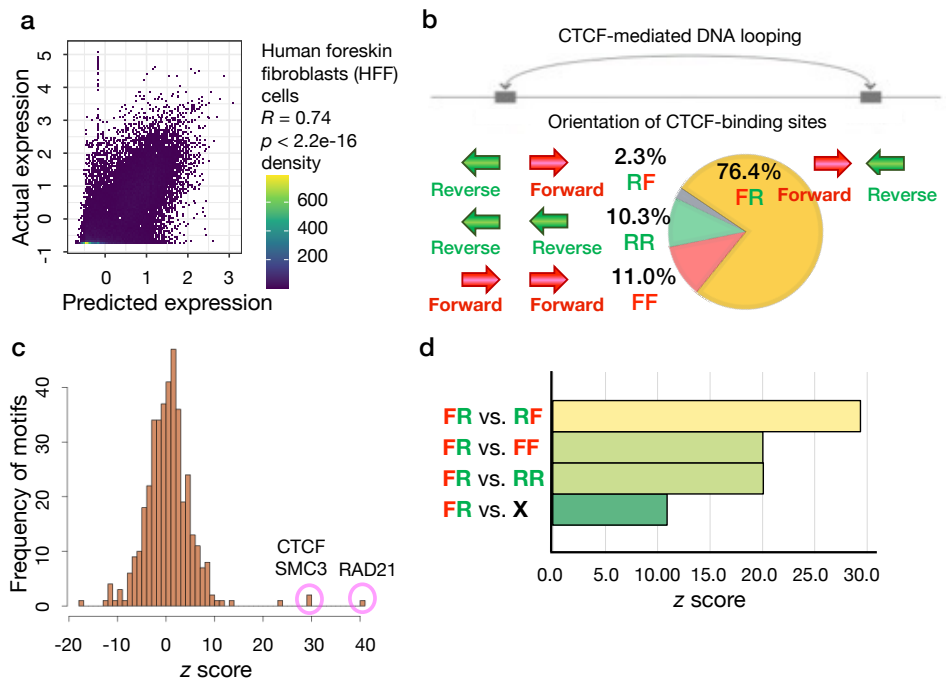


713

714 **Fig. 2. Overview of the prediction of insulator-associated DBPs. a** We used a
715 machine learning method named DEcode, with some modifications. The output data of
716 gene expression levels in human cells were predicted from the input data of genomic
717 locations of DBP binding at promoters and putative enhancers. **b** In the machine
718 learning phase, a DBP binding to a putative enhancer overlapping an eSNP (single
719 nucleotide polymorphism) was associated with a gene (eGene) that was potentially
720 regulated by the DBP based on eSNP and eGene of eQTL data. Based on the association
721 of each gene, the locations of DBP ChIP-seq peaks were binned (green box) and
722 summarized in a matrix as input data: Promoter consists of 30 bins of 100 bp between
723 -2 kbp and 1 kbp region from TSS. Enhancer regions consist of 200 bins of 10,000 bp
724 between -1 Mbp and 1 Mbp region from TSS, not including promoter. **c** In the phase of
725 putative insulator-associated DBP prediction, a DBP binding to a putative enhancer was
726 associated with a gene potentially regulated by the DBP based on a computationally
727 defined regulatory domain without using eQTL data. The *single nearest gene*
728 association rule defines a regulatory domain connecting the midpoints between the TSS

729 of a gene and the TSS of the nearest gene in the upstream and downstream regions. The
730 TSS and regulatory domain of each gene are indicated with a purple, light blue, or blue
731 arrow and a bracketed line, respectively. The regulatory domain of enhancer–promoter
732 association was restricted by the DNA-binding sites (red and green arrows) of a putative
733 insulator-associated DBP in this study. Based on the association of each gene, the
734 locations of DBP ChIP-seq peaks were binned (green box in non-gray regions) and
735 summarized in a matrix and used as input data. **d** Estimation of the directional effect of
736 the DNA-binding sites (red and green arrows) of a putative insulator-associated DBP on
737 the prediction of gene expression levels. A DBP whose binding (orange box) is
738 important for predicting a gene expression level has a high contribution score.

739



740

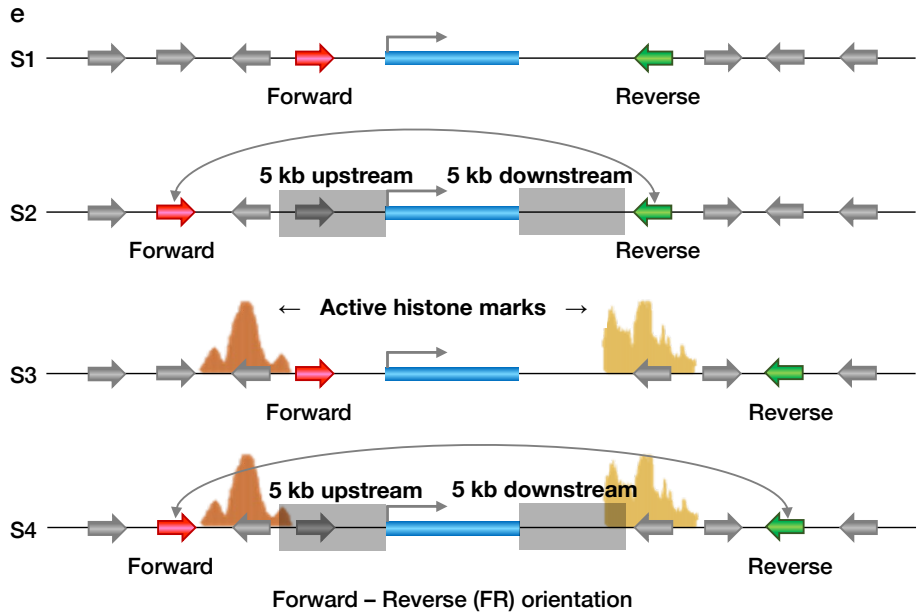
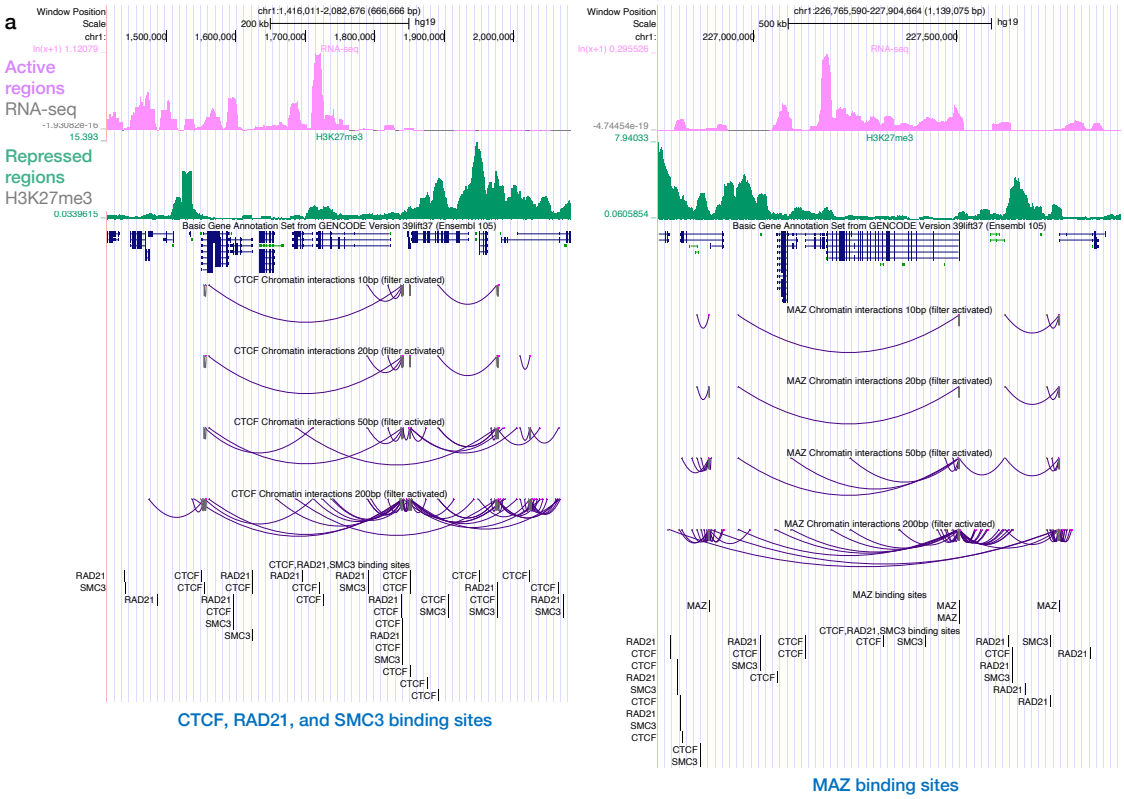
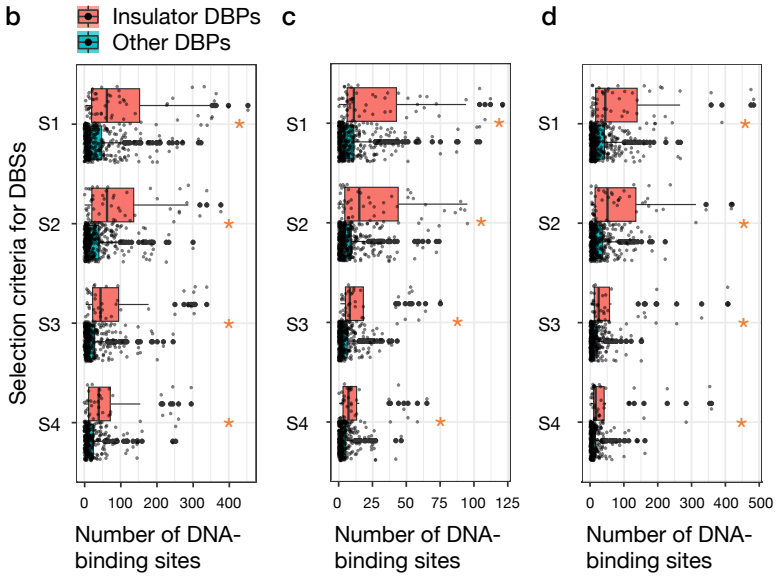


Fig. 3. Estimation of the effect of directional regulatory motifs on gene expression. **a** The predicted log2-TPM (transcripts per million) values of genes were compared with the actual expression levels in human foreskin fibroblast (HFF) cells using Spearman's rank correlation. **b** An FR orientation of CTCF-binding sites is common (76.4%) at chromatin interaction sites. Forward-forward (FF), reverse-reverse (RR), and RF orientations occur at low frequencies (11.0%, 10.3%, and 2.3%, respectively) in chromatin interaction sites. **c** Distribution of *z* scores from the Mann–Whitney *U* test for DeepLIFT scores between forward–reverse (FR) and reverse–forward (RF) orientations of DNA-binding sites of proteins. The graph shows the results based on Selection Criteria S2 in Fig. 3e and the Extended Data. **d** *Z* scores from the Mann–Whitney *U* test for DeepLIFT scores in pairwise comparisons of CTCF-binding sites with four different orientations (FR, RF, FF, and RR) and no orientation (X, i.e., orientation is not

754 distinguished). The differences in *z* scores followed the same trend as the percentages of
755 CTCF-binding site orientations in chromatin interaction sites. **e** Selection Criteria S1:
756 DNA-binding sites of predicted insulator-associated DBPs; S2: DNA-binding sites of
757 predicted insulator-associated DBPs, where the DNA-binding sites overlap with a
758 chromatin interaction site in a region between more than 5 kb upstream and 5 kb
759 downstream of a transcribed region; S3: DNA-binding sites of predicted
760 insulator-associated DBPs, where the DNA-binding sites do not overlap with the
761 ChIP-seq peaks of H3K4me3, H3K4me1, or H3K27ac; S4: DNA-binding sites of
762 predicted insulator-associated DBPs, where the DNA-binding sites overlap a chromatin
763 interaction site in a region between more than 5 kb upstream and 5 kb downstream of a
764 transcript and the DNA-binding sites do not overlap the ChIP-seq peaks of H3K4me3,
765 H3K4me1, or H3K27ac.
766


a
Active regions
RNA-seq
Repressed regions
H3K27me3
CTCF, RAD21, and SMC3 binding sites
MAZ binding sites

767


b
Insulator DBPs
Other DBPs
c
d
Selection criteria for DBSs
S1
S2
S3
S4
Number of DNA-binding sites

768

769 **Fig. 4. DNA-binding sites of predicted insulator-associated DBPs located at**
770 **potential insulator sites. a** Distribution of H3K27me3 marks (green color) and
771 transcript levels (pink color) in upstream and downstream regions of a potential
772 insulator site. Blue and green lines below H3K27me3 marks represent transcripts of
773 genes. The binding sites of CTCF, RAD21, SMC3, and MAZ are shown as black bars.
774 Chromatin interactions overlapped with a 10bp, 20bp, 50bp, 200bp region centered on
775 the binding sites of CTCF (left panel) and MAZ (right panel) are shown as dark blue
776 arcs, respectively. **b** Number of DNA-binding sites of predicted insulator-associated
777 DBPs, where H3K27me3 signal intensities differed between the upstream and
778 downstream regions of the DNA-binding site. * p value < 0.01 (Table 2). Boxplots show
779 the median as the centerline, the boxes define the interquartile range (IQR), the
780 whiskers extend up to 1.5 times the IQR, and all points represent outliers from that
781 range. All data points are displayed. **c** Number of DNA-binding sites of predicted
782 insulator-associated DBPs, where H3K27me3 signal intensities and transcript levels
783 differed between the upstream and downstream regions of the DNA-binding site. **d**
784 Number of DNA-binding sites of predicted insulator-associated DBPs in boundary
785 regions assigned based on chromatin interaction data from human foreskin fibroblast
786 (HFF) cells. Homo sapiens (human) genome assembly GRCh37 (hg19) from Genome
787 Reference Consortium was used in panels **a**.
788

789

790 **Fig. 5. Insulator-associated DBPs potentially regulate alternative transcription.**

791 The DNA-binding sites of insulator-associated DBPs were frequently located near

792 transcription start and termination sites. Insulator-associated DBPs potentially regulate

793 transcribed regions differentially repressed in alternative transcripts.

794

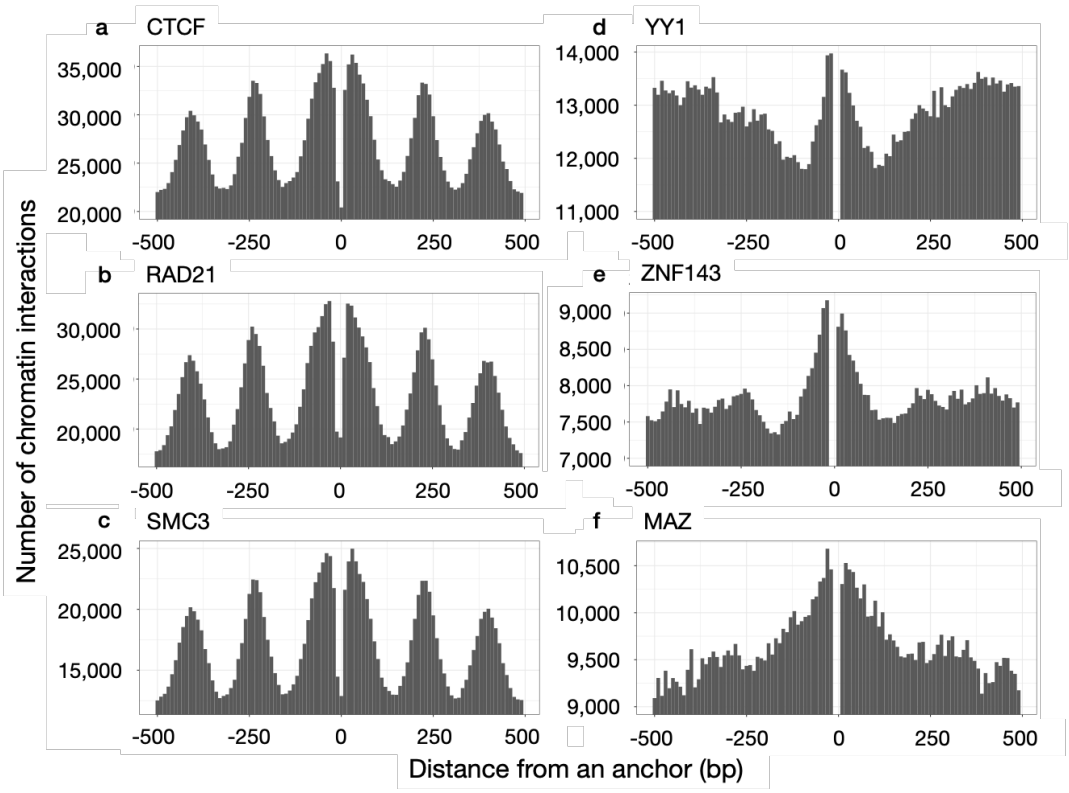


Fig. 6. DNA-binding sites distributed around a chromatin interaction site. a-f The X-axis shows the distance of a DNA-binding site from a chromatin interaction site. The Y-axis shows the number of chromatin interaction sites with DNA-binding sites within a ±500-bp region from the center of the interaction site.

Tables

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

806 *Z* scores from the Mann–Whitney *U* test for DeepLIFT scores between forward-reverse
807 (FR) and other orientations such as reverse-forward (RF), forward-forward (FF),
808 reverse-reverse (RR), and non-orientation (X) of DNA-binding sites of proteins. Mean
809 differences of DeepLIFT scores. Color scales are shown for each DNA-binding protein
810 (DBP). Top: Selection Criteria S1 in Fig. 3e; Middle: Selection Criteria S2 in Fig. 3e;
811 Bottom: List of predicted insulator-associated DBPs. DBPs in a yellow cell are
812 associated with insulators, CTCF, and/or other transcriptional regulations that have been
813 shown to affect enhancer-promoter interactions (EPIs) in previous studies.
814
815 **Table 2. Number of DNA-binding sites of predicted insulator-associated DBPs**
816 **where DNA-binding sites are located at insulator sites.**

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

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 DNA-binding sites do not overlap with the ChIP-seq peaks of H3K4me3, H3K4me1, or H3K27ac; Selection Criteria S4: DNA-binding sites of predicted insulator-associated DBPs, where the DNA-binding sites overlap a chromatin interaction site in a region between more than 5 kb upstream and 5 kb downstream of a transcript and the

827 DNA-binding sites do not overlap the ChIP-seq peaks of H3K4me3, H3K4me1, or

828 H3K27ac. *P* values were calculated from the Mann–Whitney *U* test.

829

830 **Table 3. Pairs of DNA-binding sites colocalize at the same distance from chromatin**

831 **interaction sites.**

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

832

833 # chromatin interactions with DBP1 and DBP2 sites shows the number of chromatin
834 interaction sites with DBP1 and DBP2 binding sites within ±500 bp of the interaction
835 site. Ratio (DBP1 and 2)/DBP2 is the ratio of the number of chromatin interaction sites
836 with both DBP1 and DBP2 binding sites to the number of interaction sites with DBP2
837 binding sites.
838
839

840 **References**

841 1 Weintraub, A. S. *et al.* YY1 Is a Structural Regulator of Enhancer-Promoter
842 Loops. *Cell* **171**, 1573-1588 e1528, doi:10.1016/j.cell.2017.11.008 (2017).
843 2 Rao, S. S. *et al.* A 3D map of the human genome at kilobase resolution reveals
844 principles of chromatin looping. *Cell* **159**, 1665-1680,
845 doi:10.1016/j.cell.2014.11.021 (2014).
846 3 Guo, Y. *et al.* CRISPR Inversion of CTCF Sites Alters Genome Topology and
847 Enhancer/Promoter Function. *Cell* **162**, 900-910, doi:10.1016/j.cell.2015.07.038
848 (2015).
849 4 Ortabozkoyun, H. *et al.* CRISPR and biochemical screens identify MAZ as a
850 cofactor in CTCF-mediated insulation at Hox clusters. *Nature genetics* **54**,
851 202-212, doi:10.1038/s41588-021-01008-5 (2022).
852 5 Goel, V. Y., Huseyin, M. K. & Hansen, A. S. Region Capture Micro-C reveals
853 coalescence of enhancers and promoters into nested microcompartments. *Nature*
854 *genetics*, doi:10.1038/s41588-023-01391-1 (2023).
855 6 Bailey, S. D. *et al.* ZNF143 provides sequence specificity to secure chromatin
856 interactions at gene promoters. *Nat Commun* **2**, 6186, doi:10.1038/ncomms7186
857 (2015).
858 7 Lis, M. & Walther, D. The orientation of transcription factor binding site motifs

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

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

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

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

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

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

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

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

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

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

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

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

889 binding microarray data on protein-DNA interactions. *Nucleic acids research* **37**,
890 D77-82, doi:10.1093/nar/gkn660 (2009).
891 19 Portales-Casamar, E. *et al.* JASPAR 2010: the greatly expanded open-access
892 database of transcription factor binding profiles. *Nucleic acids research* **38**,
893 D105-110, doi:10.1093/nar/gkp950 (2010).
894 20 Jolma, A. *et al.* DNA-binding specificities of human transcription factors. *Cell*
895 **152**, 327-339, doi:10.1016/j.cell.2012.12.009 (2013).
896 21 Kheradpour, P. & Kellis, M. Systematic discovery and characterization of
897 regulatory motifs in ENCODE TF binding experiments. *Nucleic acids research*
898 **42**, 2976-2987, doi:10.1093/nar/gkt1249 (2014).
899 22 Kulakovskiy, I. V. *et al.* HOCOMOCO: towards a complete collection of
900 transcription factor binding models for human and mouse via large-scale
901 ChIP-Seq analysis. *Nucleic acids research* **46**, D252-D259,
902 doi:10.1093/nar/gkx1106 (2018).
903 23 Cuddapah, S. *et al.* Global analysis of the insulator binding protein CTCF in
904 chromatin barrier regions reveals demarcation of active and repressive domains.
905 *Genome research* **19**, 24-32, doi:10.1101/gr.082800.108 (2009).
906 24 Huang, S., Li, X., Yusufzai, T. M., Qiu, Y. & Felsenfeld, G. USF1 recruits
907 histone modification complexes and is critical for maintenance of a chromatin
908 barrier. *Mol Cell Biol* **27**, 7991-8002, doi:10.1128/MCB.01326-07 (2007).
909 25 Crane, E. *et al.* Condensin-driven remodelling of X chromosome topology
910 during dosage compensation. *Nature* **523**, 240-244, doi:10.1038/nature14450
911 (2015).
912 26 Reyes, A. & Huber, W. Alternative start and termination sites of transcription
913 drive most transcript isoform differences across human tissues. *Nucleic acids*
914 *research* **46**, 582-592, doi:10.1093/nar/gkx1165 (2018).
915 27 Xiao, T., Li, X. & Felsenfeld, G. The Myc-associated zinc finger protein (MAZ)
916 works together with CTCF to control cohesin positioning and genome
917 organization. *Proc Natl Acad Sci U S A* **118**, doi:10.1073/pnas.2023127118
918 (2021).

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Acknowledgements

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

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

Author contributions

N.O. designed and performed the research, and wrote the manuscript. M.H. contributed to part of the interpretation of the results, and 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.