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 function as barriers to enhancer–promoter interactions (EPIs), thereby regulating gene expression. The primary insulator protein in vertebrates is CTCF, a DNA-binding protein (DBP); however, the roles of other DBPs in EPI insulation are not fully understood. To address this, we developed a systematic and comprehensive deep learning–based approach to identify DNA motifs of DBPs associated with insulator function. Applying this method to human fibroblast cells, we identified 97 directional motifs and a smaller number of non-directional motifs. These motifs were mapped to 23 DBPs previously linked to insulator activity, CTCF, and/or other forms of chromosomal transcriptional regulation. We found that the estimated orientation bias of CTCF was consistently proportional to the orientation bias observed in chromatin interaction data. Furthermore, these motifs showed significant enrichment at insulator sites that separate repressive and active chromatin regions, at chromatin interaction–defined boundaries, and at splice sites, compared to motifs of other DBPs. For instance, we observed that the key regulator MyoD-binding site is located at an insulator site near a gene involved in skeletal muscle differentiation and function. Importantly, our findings support the previously proposed insulator-pairing model, which suggests that insulator–insulator interactions are orientation-dependent, and highlight the involvement of multiple DNA-binding proteins beyond CTCF. Together, these results provide new insights into transcriptional regulatory mechanisms mediated by insulator-associated DBPs.

38 ## Introduction

39 Enhancer−promoter interactions (EPIs), which are organized by chromatin
40 architecture, regulate gene transcription [1]. Chromatin interactions establish loop
41 structures stabilized by transcription factors (TFs) and DNA-binding proteins (DBPs),
42 such as CTCF and components of the cohesin complex (e.g., RAD21 and SMC3).
43 CTCF typically acts as an insulator by anchoring chromatin loops and restricting
44 enhancer activity within these loops [2]. Pairs of CTCF-binding sites display a
45 characteristic directional bias, favoring a forward−reverse (FR) orientation over forward
46 −forward (FF), reverse−reverse (RR), or reverse−forward (RF) orientations [3]. The
47 forward and reverse orientations of DNA-binding sites are complementary in sequence
48 and can be readily distinguished based on reference genome sequences. Moreover,
49 although FR and RF orientations are sequence complements, their frequencies at
50 chromatin loop anchors differ significantly [3], suggesting that the directional bias of
51 DNA-binding sites is associated with chromatin loop formation.

52 While CTCF and cohesin are recognized as the primary insulator proteins in
53 vertebrates [4], regional micro-capture (Micro-C) analyses have demonstrated that
54 microcompartment interactions are largely independent of cohesin and transcription [5].
55 These findings imply that other DBPs may function as insulators or be associated with
56 insulator activity. However, experimentally identifying insulator-associated DBPs
57 among the numerous DBPs present in the genome remains labor-intensive and
58 technically challenging when relying on in situ assays.

59 One potential approach to identifying insulator-associated DBPs is to analyze the

60 enrichment of DNA-binding sites for chromatin- and insulator-associated proteins at
61 chromatin interaction sites. For example, DBPs such as ZNF143 and YY1 have been
62 shown to participate in chromatin interactions and enhancer−promoter interactions
63 (EPIs) [1,6]. However, while the binding sites of ZNF143 are indeed enriched at chromatin
64 interaction sites, those of YY1 are not significantly enriched [6]. Moreover, CTCF,
65 ZNF143, and YY1 participate in a variety of biological functions beyond chromatin
66 organization. This multifunctionality likely reduces motif enrichment at chromatin
67 interaction sites, complicating the identification of DBPs that are specifically associated
68 with insulator functions.

69 Another potential approach for predicting insulator-associated DBPs is to focus on
70 the directional bias of CTCF DNA-binding sites. The orientation of a DBP's binding
71 site may not be relevant when the protein binds to a single genomic location [7]. However,
72 when a DBP or a complex of DBPs binds to multiple genomic loci through chromatin
73 interactions, their binding is often structurally organized, contributing to the observed
74 directional bias of DNA-binding sites [8-10]. Some DBPs involved in chromatin
75 interactions, such as YY1, are multifunctional proteins, with certain binding sites
76 participating in both chromatin interactions and insulator functions. As a result, the
77 directional bias of DNA-binding sites for multifunctional DBPs may be diminished.
78 This multifunctionality complicates the identification of their specific
79 insulator-associated roles.

80 Computational methods have been developed to predict chromatin interactions
81 and their associated DBPs [1112]. However, these methods are not designed to identify the

82 binding motifs of DBPs specifically associated with insulator function. Regarding DBP
83 binding motifs, Xie et al. noted that a key challenge is developing systematic
84 approaches for determining their specific functions on a genome-wide scale [13]. In terms
85 of gene regulation, insulators influence EPIs and gene expression; therefore,
86 incorporating gene expression data may improve the identification of
87 insulator-associated DBPs compared to approaches that rely solely on enrichment
88 analysis or the directional bias of DNA-binding sites. Recently, a novel deep
89 learning–based method was introduced to estimate gene expression levels in human
90 tissues based on DNA-binding sites within promoter regions [14]. However, this method
91 does not account for DNA-binding sites associated with EPIs as input features and was
92 not intended to identify DBPs linked to insulator function.

93 Therefore, we aimed to identify insulator-associated DBPs using a computational
94 approach that integrates and analyzes genome-wide experimental data. In this study, we
95 refined an existing deep learning method by incorporating EPI information based on
96 eQTL data and enhancer–promoter association rules (McLean et al., Supplementary Fig.
97 2) [15]. We analyzed the directional bias of DNA-binding sites for CTCF and
98 insulator-associated DBPs, including known factors such as RAD21 and SMC3, using
99 contribution scores generated by the deep learning model. Our analysis demonstrated
100 that the directional bias of CTCF binding sites closely matches the orientation bias
101 observed in chromatin interaction data. Furthermore, statistical testing of contribution
102 scores between directional and non-directional DNA-binding sites of
103 insulator-associated DBPs showed that directional sites contribute more significantly to

gene expression prediction than non-directional sites. Overall, we identified 23 insulator-associated DBPs related to insulator function, CTCF-associated activity, and/or other transcriptional regulatory mechanisms affecting EPIs, as previously reported. To validate these findings, we showed that the DNA-binding sites of the identified DBPs are located within putative insulator regions, and that some of these sites exist independently of nearby CTCF or cohesin binding sites. Finally, consistent with the insulator-pairing model proposed from experimental studies in flies, we confirmed that both homologous and heterologous insulator–insulator pairing interactions are orientation-dependent. Our approach facilitates the identification of insulator- and chromatin-associated DNA-binding sites that modulate EPIs and uncovers novel functional roles and molecular mechanisms of DBPs involved in transcriptional condensation, phase separation, and transcriptional regulation.

Results

Predictive modeling of gene expression from directional DBP binding

Accurate gene expression prediction can be achieved by integrating both distal and proximal DNA-binding protein (DBP) binding sites as features within a machine learning framework. By combining ChIP-seq data from the Gene Transcription Regulation Database (GTRD v20.06) with expression quantitative trait locus (eQTL) data from GTEx v8 [16], our model was able to account for enhancer–promoter interactions (EPIs) by linking single nucleotide polymorphisms (SNPs) to gene

126 expression outcomes (**Fig. 1a-c**, **Fig. 2a**, **b**). Specifically, DBP binding sites within ±50
127 bp of an eQTL SNP were assumed to influence the associated transcript. Focusing on
128 four human cell types—HFFs, monocytes, HMECs, and neural progenitor cells—the
129 model achieved strong concordance with observed expression levels: Spearman's ρ
130 values reached 0.74, 0.71, 0.74, and 0.73 for all 27,428 genes, and 0.70, 0.67, 0.69, and
131 0.70 for 2,705 test genes (**Fig. 3a**, **Fig. S2a**). Importantly, this performance was
132 obtained without incorporating RNA-binding protein or miRNA data as required in
133 previous approaches [14].

134 To build a robust foundation for regulatory modeling, we comprehensively
135 identified binding motif sequences for 715 out of 1,310 DBPs using ChIP-seq peaks
136 from the Gene Transcription Regulation Database (GTRD, v20.06 release) and the
137 Protein Interaction Quantitation (PIQ) tool, which integrates DNase-seq data and a
138 curated library of motifs from TRANSFAC (2019.2), JASPAR (2018), UniPROBE
139 (2018), high-throughput SELEX, ENCODE ChIP-seq–derived motifs, and
140 HOCOMOCO (versions 9 and 11) [17,18-23]. Only motif occurrences overlapping ChIP-seq
141 peaks for each DBP were retained, thereby minimizing false positives and enabling the
142 high-resolution prediction of DBP binding in open chromatin regions.

143 We next examined how the directionality and genomic position of
144 insulator-associated DBP binding sites affect gene expression prediction. Using
145 DeepLIFT, which quantifies each feature's contribution to the model by measuring its
146 deviation from a reference, we demonstrated that both the orientation and the position
147 of DBP binding sites have substantial regulatory effects [24]. Directional differences in

148 putative insulator-associated DBP sites emerged as critical determinants of EPI strength
149 and gene expression outcomes (**Fig. 2d**). This systematic approach allowed us to dissect
150 how the regulatory effects of insulator-associated DBPs depend on both their chromatin
151 context and strand orientation.

152 In particular, we found that CTCF binding site orientation is a key determinant of
153 its regulatory impact on gene expression. CTCF sites in the forward–reverse (FR,
154 convergent) orientation—accounting for 76.4% of distal CTCF-mediated chromatin
155 interaction sites identified by ChIA-PET (**Fig. 3b**)—exhibited significantly different
156 DeepLIFT score distributions compared to the reverse–forward (RF) orientation
157 (Mann–Whitney U test, FDR-adjusted $p < 0.05$; $p < 10^{-3}$; **Fig. 3c**, **d**; **Table S1**) [3]. This
158 difference was robustly detected when comparing FR and RF CTCF-binding sites
159 located in predicted enhancer and open chromatin regions both upstream and
160 downstream of TSSs (**Fig. 2c**). Further analyses using alternative DNA-binding site
161 selection criteria supported these findings (**Supplementary Fig. 1** and **Note**),
162 demonstrating that our approach captures the regulatory significance of CTCF binding
163 site orientation.

164 Extending this analysis, we observed that directional orientation is a widespread
165 and biologically meaningful predictor of gene expression among putative
166 insulator-associated DBPs. DBPs such as RAD21, SMC3, and CTCF exhibited strong
167 differences in DeepLIFT score distributions between FR and RF orientations (**Fig. 3c**,
168 **Table S1**), and this effect was not limited to these proteins. Among 228 DBPs with at
169 least 1,000 binding sites in HFF cells, 92—including ZNF143 and YY1—showed

170 significant differences between sites with different orientations (e.g., FR vs. RF), while
171 79—including YY1—also showed differences between same-orientation sites (e.g., FF
172 vs. RR) (**Fig. 3c**, **Fig. S2b**, **Table S1**, sheet **"**DBP List 1A"). Comparisons between
173 directional (FR, RF, FF, RR) and non-directional (X) binding sites consistently
174 produced higher *z*-scores for the former, highlighting the greater contribution of
175 directional binding sites to gene expression prediction. The most pronounced effects
176 were observed for established insulator- and chromatin-associated proteins, underlining
177 the biological importance and sensitivity of our model. For increased stringency, we
178 restricted the analysis to DBPs with $p < 10^{-3}$, confirming all such cases overlapped with
179 those defined by the broader FDR-adjusted $p < 0.05$ threshold (**Table S1**). These
180 findings prompted us to systematically classify insulator-associated DBPs by their
181 directional or non-directional binding patterns for further analysis.
182

183 **Identification and classification of insulator-associated DBPs**

184 Insulator-associated DBPs demonstrate widespread and significant
185 orientation-dependent as well as orientation-independent DNA-binding preferences,
186 reflecting their diverse roles in chromatin regulation. Our analysis revealed that
187 established insulator DBPs—including CTCF, RAD21, and SMC3—exhibit a
188 predominant forward–reverse (FR) orientation in their binding site distributions.
189 Pairwise comparisons of DeepLIFT scores across four orientations (FR, RF, FF, RR)
190 and non-directional sites showed significant differences for 23 DBPs with different
191 orientations and for 17 DBPs with the same orientation. Additionally, non-directional

192 binding sites showed significant score differences for six DBPs ($p < 10^{-3}$; **Table 1** and
193 **Table S2**, sheet "DBP List 1A"; FDR-adjusted $p < 0.05$; **Table S3**, "DBP List 1A").
194 Beyond these established factors, we identified several novel DBPs that displayed both
195 directional and non-directional binding preferences, expanding the pool of proteins
196 potentially involved in insulator function.

197 Importantly, these widespread orientation biases were supported by chromatin
198 interaction data. Our deep learning–based predictions accurately recapitulated the
199 characteristic orientation bias of CTCF and RAD21 DNA-binding sites at chromatin
200 interaction sites, as validated by ChIA-PET data. Specifically, among CTCF binding
201 sites overlapping distal interaction anchors, 76.4% displayed an FR orientation, with
202 substantially lower frequencies of FF (11.0%), RR (10.3%), and RF (2.3%) (**Fig. 3b**;
203 Guo Y et al., *Cell* 2015 [3], Fig. 4). DeepLIFT scores revealed that the difference between
204 FR and RF orientations was significantly greater than that between FR and FF or RR
205 (Mann–Whitney U test), consistent with experimental distributions (**Fig. 3c**, **d**).
206 Similarly, RAD21 ChIA-PET data revealed 76.3% FR, 3.4% RF, 9.3% FF, and 11.1%
207 RR orientations, and DeepLIFT scores for RAD21 binding sites closely mirrored these
208 results (**Fig. 3c**; **Table 1**). These findings confirm that our framework robustly predicts
209 the experimentally observed orientation biases of insulator-associated DBPs at
210 chromatin interaction sites.

211 Applying stringent orientation criteria, we identified a comprehensive set of 97
212 insulator-associated DBPs that directly interact with DNA and exhibit either strong
213 directional or non-directional binding preferences. Most of these DBPs exhibited clear

214 directional biases, while only six—including FOS, FOSL2, JUN, JUNB, JUND (all
215 with palindromic motifs), and SPI1 (their interacting partner)—were characterized by
216 non-directional binding. Notably, 23 of these DBPs have been previously implicated in
217 insulator activity, CTCF association, or chromatin regulatory functions affecting
218 enhancer–promoter interactions (EPIs), with 13 of them directly interacting with CTCF
219 (**Table S7**). These proteins are multifunctional, serving as transcriptional regulators,
220 repressors, chromatin regulators, or cofactors for epigenetic modifier enzymes.

221 Collectively, our findings indicate that CTCF coordinates EPI regulation through
222 a network of diverse DBPs and that both directional and non-directional binding
223 preferences underlie the regulatory capacity of these proteins. The involvement of these
224 multifunctional DBPs in chromatin architecture and long-range genomic interactions
225 emphasizes the complex and modular organization of chromatin regulatory networks.

226 Extending these analyses to the spatial organization of chromatin, we found that
227 insulator-associated DBPs are frequently co-localized at both ends of chromatin
228 interaction sites, supporting their role in mediating long-range chromatin architecture.
229 Across 30,369 chromatin interactions defined by 400-bp regions (separated by >5 kb
230 and excluding TSS-proximal areas), DNA-binding sites for 44 (45%) and 97 (100%)
231 insulator-associated DBPs were identified at both ends or at either end of an interaction
232 pair, respectively. For 10,384 interactions defined by 200-bp regions, binding sites for
233 30 (31%) and 97 (100%) DBPs were detected at both or either end. These analyses,
234 based on orientation-aware pairings (FR, RF, FF, RR) (**Fig. 3e**), likely underestimate the
235 full extent of insulator DBP involvement due to incomplete coverage of chromatin

236 interaction data [25]. Canonical insulator proteins such as CTCF, RAD21, and SMC3
237 exhibited strong directional binding within these regions, consistent with ChIA-PET
238 results. To further validate the predictions of candidate insulator-associated DBPs, we
239 next sought independent lines of evidence—including histone modification signatures
240 characteristic of insulator function—to corroborate their roles in chromatin boundary
241 and architecture.
242

243 **Epigenomic validation of insulator activity**

244 Epigenomic analyses provide independent support for the predicted regulatory
245 roles of insulator-associated DBPs. We identified insulator-associated DBP binding sites
246 by detecting asymmetry in H3K27me3 signal profiles across their upstream and
247 downstream regions, reflecting boundaries between active and repressed chromatin
248 domains. Specifically, loci where the mean H3K27me3 signal intensity differed by more
249 than twofold between upstream and downstream regions, with the higher value
250 exceeding the first quartile of the genome-wide distribution, were selected to ensure
251 robust discrimination and to filter out low-intensity noise. This approach builds on prior
252 observations that CTCF and USF1 occupy the edges of repressive chromatin domains
253 (marked by H3K27me3) versus active regions (marked by H2AK5ac) [26,27], making
254 H3K27me3 a suitable marker for broad transcriptional repression.

255 DNA-binding sites of predicted insulator-associated DBPs were significantly
256 enriched at chromatin boundaries defined by asymmetric H3K27me3 signals. Known
257 insulator-associated factors, such as CTCF, RAD21, and SMC3, showed the highest

258 number of such sites (**Table S4**). Quantitatively, the median and mean counts of binding
259 sites with greater than twofold H3K27me3 signal differences between upstream and
260 downstream regions were significantly higher for predicted insulator-associated DBPs
261 than for other DBPs (Mann–Whitney U test, $p < 0.05$; **Table 2**, **Table S4**, **Fig. 4a**, **b**).
262 These results demonstrate that insulator-associated DBPs preferentially occupy genomic
263 positions flanked by distinct chromatin states, providing further evidence for their
264 function in defining regulatory domains.

265 Furthermore, predicted insulator-associated DBPs are localized to transcriptional
266 transition regions and chromatin boundaries. Analysis of RNA-seq data from HFF cells
267 showed that these DBPs are significantly enriched at sites where transcript signal
268 intensities differ by more than 1.5-fold between upstream and downstream regions, with
269 the higher value exceeding the first quartile of the transcriptome-wide distribution
270 (**Table 2**, **Table S4**, **Fig. 4a**, **c**). Additionally, DNA-binding sites of these DBPs more
271 frequently overlapped with chromatin boundary regions, as defined by insulation scores
272 from chromatin interaction data [28], than those of non-insulator DBPs (**Table 2**, **Fig. 4d**).
273 These results underscore the role of insulator-associated DBPs in partitioning active and
274 inactive chromatin domains and functionally positioning at regulatory boundaries. To
275 explore the fine-scale genomic context of these sites, we next conducted genome
276 browser inspections.

277 Genomic inspection using the UCSC Genome Browser revealed that predicted
278 insulator-associated DBPs consistently occupy sites with insulator-like features,
279 including chromatin boundary regions identified by Micro-C data and transitional

280 regions between H3K27me3-marked and transcriptionally active domains. Many of
281 these binding sites were located at the edges of H3K27me3-enriched domains and were
282 frequently distant from canonical insulator proteins such as CTCF, RAD21, and SMC3
283 (**Fig. 4a**, **Fig. S3**). For each DBP, up to three representative binding sites in
284 insulator-associated contexts were manually selected (**Fig. S3a**, **S3f**). Notably, MyoD1
285 was found at a chromatin boundary near the SELENON (SEPN1) gene—both involved
286 in skeletal muscle differentiation [29] —suggesting a possible regulatory role for MyoD1
287 via insulator mechanisms (**Fig. 4a**).

288 Expanding this analysis to alternative repressive histone modifications,
289 insulator-associated DBPs were also enriched at the edges of H3K9me3-enriched
290 domains (**Fig. S3b**), supporting their ability to define chromatin boundaries beyond
291 those marked by H3K27me3. The distribution of H3K9me3 differed from that of
292 H3K27me3 in certain regions, highlighting the context-dependent nature of insulator
293 activity. Additionally, predicted DBP binding sites frequently formed spatial clusters at
294 insulator regions and were often associated with chromatin interaction anchors (**Fig. 4f**,
295 **Fig. S3c**, **S3d**). This clustering aligns with the role of architectural proteins in
296 organizing TAD boundaries, while discrete insulator motifs contribute to intra-TAD
297 looping [30]. To further clarify the relationship between these epigenetically defined
298 insulator sites and chromatin architecture, we compared their locations with chromatin
299 boundaries identified by chromatin interaction data.

300 Comparative analyses showed that histone- and transcription-defined insulator
301 sites often coincided with, or were located near, chromatin boundaries identified by

302 interaction data. In HFF cells, 29% of H3K27me3-based insulator sites (type 1) and
303 37% of H3K27me3 plus transcription-based sites (type 2) overlapped with 25-kb
304 boundary regions, whereas only 6% and 7% overlapped with 5-kb boundaries,
305 respectively, reflecting differences in boundary resolution (**Fig. 4f**, **Fig. S3**).
306 Genome-wide, DNA-binding sites of the 97 predicted insulator-associated DBPs were
307 found at type 1 insulator sites assigned to 3,170 genes, at type 2 sites assigned to 1,044
308 genes, and at chromatin interaction–defined boundaries assigned to 6,275 genes. Among
309 these, 1,212 genes (38%) harbored both type 1 insulator and boundary sites, and 389
310 genes (37%) contained both type 2 insulator and boundary sites. These data indicate that
311 while chromatin interaction boundaries and epigenetically or transcriptionally defined
312 insulator sites are not completely overlapping, they frequently target overlapping gene
313 sets and are co-located within the same regulatory landscapes, supporting their
314 complementary roles in chromatin compartmentalization.
315

316 Independent and cooperative clustering of insulator-associated DBPs 317 at chromatin regulatory landmarks

318 Predicted insulator-associated DBPs are enriched at critical transcript regulatory
319 regions, such as transcription start sites (TSSs), transcription termination sites (TTSs),
320 and splice sites, suggesting potential roles in shaping transcript structure. Statistical
321 analyses showed that these DBPs are significantly more abundant at splice sites than
322 other DBPs (Mann–Whitney U test, $p < 0.05$; **Fig. 4e**, **Table 2**), with pronounced
323 enrichment at the first and last introns compared to internal introns (**Table 2**, **Table S9**).

This finding is consistent with prior reports that variation in TSS/TTS usage is a key driver of tissue-specific isoform diversity [31]. Examination of **Fig. 4a** and **Fig. S3** further revealed clustering of multiple insulator-associated DBP binding sites within single loci, suggesting potential cooperative functions in establishing insulator elements. Notably, transcribed regions with overlapping H3K27me3 enrichment and alternative transcriptional repression contained DBP sites between active and repressed domains (**Fig. 5**, **Fig. S3e**), indicating that insulator-associated DBPs may delineate boundaries between transcriptional states, thereby influencing selection of TSSs, TTSs, and exons in a context-dependent manner.

Chromatin interaction–associated DBPs, including several predicted insulator factors, are strongly enriched near chromatin interaction centers, emphasizing their potential in supporting chromatin looping. High-resolution Micro-C data from HFF cells identified DBP binding sites located within ±500 bp of interaction anchors [32]. Canonical organizers CTCF, RAD21, and SMC3 exhibited sharp enrichment at these sites (**Fig. 6a–c**), while YY1 and ZNF143 displayed broader or distinct distributions (**Fig. 6d**, **e**), and MAZ—a DBP implicated in CTCF-independent insulator activity [33]—demonstrated a unique peak (**Fig. 6f**). Even after removing binding sites overlapping CTCF, RAD21, and SMC3, other predicted insulator-associated DBPs remained consistently enriched at interaction centers, supporting their independent contributions to 3D chromatin structure.

Analysis of 228 DBPs with over 1,000 binding sites in HFF cells revealed that nearly 90% exhibited distinct enrichment near chromatin interaction sites (**Table S8**,

346 **Fig. S4**), displaying sharp, broad, or canonical CTCF-, RAD21-, or SMC3-like peak
347 profiles. Although DBPs such as MeCP2 and PRDM9, with fewer than 1,000 binding
348 sites, did not exhibit this pattern, these findings demonstrate that most DBPs—including
349 many predicted insulator-associated proteins—are spatially positioned to participate in
350 genome architecture regulation.

351 Co-localization analyses revealed that most predicted insulator-associated DBPs
352 are spatially segregated from canonical insulator proteins. Using 10-bp bins across ±500
353 bp windows centered at interaction anchors and ChIP-seq–validated binding sites, we
354 identified 245 DBP pairs (involving 135 DBPs) with ≥30% co-localization (Selection
355 Criteria S1 in Fig. S1; **Table S5**). CTCF was highly co-localized with RAD21 (39%)
356 and SMC3 (81%), whereas YY1/ZNF143 had much lower rates (<6%) with canonical
357 factors (**Table 3**). Some DBPs (CTCFL, RXRA, PPARG, NEUROD1, ZBTB7A)
358 exhibited moderate co-localization (>10%), but the majority of predicted
359 insulator-associated DBPs showed minimal overlap (<10%), suggesting functional
360 independence. Spatial clustering of multiple DBP binding sites was also observed at
361 putative insulator loci (**Fig. S3d**), and co-localization was especially strong at 5-kb
362 boundary-associated sites (197 pairs, 101 DBPs, ≥30% overlap; **Table S6**), suggesting
363 that these DBPs, like CTCF and cohesin, may assemble cooperatively at boundaries to
364 mediate domain partitioning.

365 Overlap analysis with ChIP-seq data showed that >98% of CTCF, RAD21, and
366 SMC3 binding sites overlapped their own ChIP-seq peaks under all conditions, thereby
367 validating site definitions. In contrast, nearly all other predicted insulator-associated

368 DBPs showed <50% overlap with CTCF, RAD21, or SMC3 peaks, even when data
369 from various cell types were considered. This low co-occupancy supports the notion
370 that predicted DBPs function largely independently of classical insulator complexes and
371 may establish chromatin boundaries by distinct mechanisms.

372 Interestingly, a distinct group of predicted insulator-associated DBPs—including
373 BACH2, FOS, ATF3, NFE2, and MAFK—exhibited binding site distributions around
374 chromatin interaction centers similar to canonical insulator proteins, but had less than
375 1% direct spatial overlap within 10-bp bins (**Fig. S4c**). However, these factors showed
376 strong mutual co-localization (**Table 3**), suggesting the existence of parallel,
377 functionally distinct DBP clusters that independently regulate chromatin structure [343536].
378 Thus, multiple DBP clusters with similar positional preferences may exert parallel
379 effects on chromatin remodeling or loop accessibility via independent regulatory
380 pathways.

381

382

383 **Discussion**

384 Here, we present the first, to our knowledge, computational method and
385 genome-wide analysis for the detection of insulator-associated DBPs. Transcriptional
386 regulation mediated by the binding of DBPs at sites located far from TSSs was modeled
387 using our deep learning approach based on eQTL data. Through this framework, the
388 effects of DNA-binding sites of insulator-associated DBPs on EPIs and gene expression
389 were estimated, and insulator-associated DBPs were predicted. We further identified

390 DBPs, including known insulator-associated proteins, that exhibit insulator-associated
391 functions and regulate transcription, thereby affecting EPIs as reported in previous
392 studies. The DNA-binding sites of the predicted insulator-associated DBPs
393 demonstrated orientation bias; unexpectedly, the significant differences in the
394 orientation bias of CTCF DNA-binding sites based on DeepLIFT scores (contribution
395 scores) were found to be consistent with the percent differences in the orientation of
396 CTCF DNA-binding sites at CTCF-mediated chromatin interaction sites. Furthermore,
397 statistical testing of the contribution scores revealed that the directional binding sites of
398 DBPs contributed more significantly to the prediction of gene expression levels than the
399 non-directional sites. The DNA-binding sites of the predicted insulator-associated DBPs
400 were located at potential insulator sites more frequently than those of other DBPs, as
401 indicated by differences in H3K27me3 signal intensities and transcript levels, and by
402 their overlap with boundary sites identified by chromatin interactions. We further
403 observed that only a small percentage of the DNA-binding sites of insulator-associated
404 DBPs co-localized with the DNA-binding sites of known insulator-associated proteins
405 such as CTCF, RAD21, and SMC3 around chromatin interaction sites. Finally, we
406 discovered that DBPs other than the well-established insulator proteins, as well as the
407 directionality of their DNA-binding sites, are associated with insulator function.

408 Our deep learning analysis accurately recapitulated the percent differences in the
409 FR, FF, RR, and RF orientations of CTCF DNA-binding sites at chromatin interaction
410 sites, by incorporating the genomic locations of DNA-binding sites for putative
411 insulator-associated DBPs. The other predicted insulator-associated DBPs also

412 demonstrated orientation bias in their DNA-binding sites. These results suggest that the
413 directional bias of DNA-binding sites of insulator-associated DBPs may contribute to
414 insulator function and chromatin regulation through structural interactions among DBPs,
415 other proteins, DNA, and RNA. For example, the N-terminal amino acids of CTCF
416 have been shown to interact with RAD21 in chromatin loops [8-10]. Cohesin complexes
417 that interact with convergent CTCF sites, that is, the N-terminus of CTCF, might be
418 protected from WAPL, whereas those that interact with divergent CTCF sites, i.e., the
419 C-terminus of CTCF, may not be protected from WAPL, which could release cohesin
420 from chromatin and thus disrupt cohesin-mediated chromatin loops [37]. As another
421 example, although sequence-based predictions of EPIs underestimated actual
422 interactions, an activity-by-contact (ABC) approach, based on a simple distance model,
423 performed well, suggesting that factors in addition to sequence influence EPIs [38,39].
424 Furthermore, CTCF-mediated chromatin looping induces the formation of
425 transcriptional condensates and phase separation between the inner and outer regions of
426 the loops, resulting in insulator function [40-42]. Therefore, structural features that drive
427 phase separation, rather than functional differences in DBPs themselves, may underlie
428 insulator activity. CEBPA, among the identified insulator-associated DNA-binding
429 proteins, has also been reported to participate in transcriptional condensates and phase
430 separation [43]. The FOXA1 pioneer factor functions as an initial chromatin-binding and
431 chromatin-remodeling factor and has also been shown to form biomolecular
432 condensates [44].

433 Other types of chromatin regulation are also likely to be related to the structural

434 interactions of molecules. In the boundary-pairing (insulator-pairing) model, molecules
435 bound to insulators physically pair with their partners, either head-to-head or
436 head-to-tail, with different degrees of specificity at the termini of TADs in flies (**Fig. 7**).
437 Although experiments have not clarified how partners locate each other, the mechanism
438 is unlikely to require loop extrusion, as is presumed in CTCF- and cohesin-mediated
439 chromatin looping. Both homologous and heterologous insulator–insulator pairing
440 interactions are thought to be central to the architectural functions of insulators.
441 Importantly, the nature of insulator–insulator interactions is orientation-dependent [45-47].
442 Regional micro-capture (Micro-C) analyses have revealed that microcompartment
443 interactions are largely independent of cohesin and transcription [5]. Depletion of cohesin
444 results in a subtle reduction in longer-range enhancer–promoter interactions, while
445 CTCF depletion can lead to rewiring of regulatory contacts [48]. Another group reported
446 that enhancer–promoter interactions and transcription are largely maintained upon
447 depletion of CTCF, cohesin, WAPL, or YY1; instead, cohesin depletion reduced
448 transcription factor binding to chromatin, suggesting that cohesin may enable
449 transcription factors to find and bind their targets more efficiently [49]. Furthermore, loop
450 extrusion is not essential for enhancer–promoter interactions but contributes to their
451 robustness, specificity, and the precise regulation of gene expressio [48]. The 'loop
452 extrusion' hypothesis is motivated by in vitro observations. Experiments in yeast, using
453 cohesin variants unable to extrude DNA loops but capable of topologically entrapping
454 DNA, suggested that in vivo chromatin loops can form independently of loop extrusion.
455 Instead, transcription was shown to promote loop formation and act as an extrinsic

456 motor that extends these loops and defines their final positions [50]. Maz and MyoD1,
457 among the identified insulator-associated DNA-binding proteins, have been shown to
458 form loops without CTCF [4,33,51]. Furthermore, protein post-translational modifications
459 (PTMs) facilitate control over the molecular valency and strength of protein–protein
460 interactions [52]. O-GlcNAcylation as a PTM has been reported to inhibit CTCF binding
461 to chromatin [53]. We found that the identified insulator-associated DNA-binding proteins
462 tend to cluster at potential insulator sites (Fig. 4f and Supplementary Fig. 3c). These
463 proteins may interact and actively regulate chromatin interactions, transcriptional
464 condensation, and phase separation via PTMs. Our analyses will contribute to revealing
465 novel functional and structural roles of DBPs in insulator functions, chromatin
466 regulation, and phase separation, and to advancing their experimental analysis.

467 High-resolution chromatin interaction data from Micro-C assays demonstrated
468 that most of the predicted insulator-associated DBPs exhibited peaks in the
469 DNA-binding site distribution around chromatin interaction sites, suggesting that these
470 DBPs are involved in chromatin interactions and highlighting the high resolution of
471 Micro-C data. Base pair–level resolution has been reported using Micro Capture-C [32].
472 However, due to the high resolution of Micro-C assays, it is challenging to obtain
473 genome-wide chromatin interaction data at a reasonable and accessible cost, whereas
474 Micro Capture-C and Region Capture Micro-C assays focus on specific genomic
475 regions to increase the coverage of chromatin interactions. The extent of chromatin
476 interaction coverage varies depending on the specific Micro-C experimental design. In
477 addition, chromatin interactions include types of interactions that are not associated

478 with EPIs and insulator functions [54-56]. Therefore, we analyzed and predicted
479 insulator-associated DBPs independently of chromatin interaction data. Our deep
480 learning analysis successfully learned enhancer–promoter interactions (EPIs) that affect
481 gene expression levels and accurately recapitulated the percent differences in the FR, FF,
482 RR, and RF orientations of CTCF DNA-binding sites at chromatin interaction sites.

483 When comparing insulator sites separated by active and repressive domains with
484 boundary sites identified by chromatin interaction data, we defined insulator sites based
485 on the total signal levels of H3K27me3 histone modification marks and transcript
486 expression levels within a region. Some insulator sites may not display smooth,
487 continuous distributions of histone marks and expression, which could account for
488 differences between insulator and boundary sites. Another possible explanation is that
489 boundary sites are more closely associated with topologically associating domains
490 (TADs) of chromosomes than are insulator sites [30]. Boundary sites represent regions
491 identified based on the separation of numerous chromatin interactions. In contrast, we
492 found that multiple DNA-binding sites of insulator-associated DNA-binding proteins
493 were frequently clustered at insulator sites and were associated with distinct nested and
494 focal chromatin interactions, as reported by Micro-C assays [5]. These interactions may
495 be transient and relatively weak, such as tissue-, cell type-, condition-, or
496 lineage-specific interactions [57].

497 AI (Artificial Intelligence) is often regarded as a black box, since the reasons and
498 causes of its predictions are difficult to determine. To address this issue, various tools
499 and methods have been developed to enhance model interpretability, collectively

500 referred to as Explainable AI. DeepLIFT is considered one such tool for Explainable AI.
501 However, DeepLIFT does not provide the specific reasons or causal mechanisms
502 underlying a prediction. Instead, it computes scores representing the contribution of
503 each input feature to the prediction outcome. In this study, analysis of numerous
504 contribution scores (i.e., DeepLIFT scores) for directional DNA-binding sites of
505 putative insulator-associated DBPs in the prediction of transcript expression levels
506 enabled accurate prediction of the percent differences in the FR, FF, RR, and RF
507 orientations of CTCF DNA-binding sites at chromatin interaction sites, based on the
508 contribution scores of DNA-binding sites in putative enhancer and open chromatin
509 regions. It should be noted that some contribution scores may not be accurate, and
510 predictions of insulator-associated DBPs or DNA-binding sites based on a small number
511 of scores may be unreliable. To enable comparison of contribution scores calculated
512 from different input–output pairs within our deep learning model, we normalized the
513 contribution scores for each input–output pair as previously described [58]. The analysis
514 of large sets of contribution scores will facilitate the estimation of the effects of other
515 molecules and functions involved in gene regulation.

516 One of the next computational challenges will be to predict and refine the
517 potential set of insulator-associated DNA-binding sites for multifunctional DBPs. To
518 validate machine learning predictions and identify insulator-associated DNA-binding
519 sites in the genome for subsequent experimental investigation, information on histone
520 modifications such as H3K27me3 and transcript levels will be valuable, and other types
521 of epigenomic data can also be utilized for this purpose. For example, as a reference for

522 an application of this approach, Liu et al. focused on CTCF-binding sites and performed
523 various assays in human and mouse cell cultures, as well as in mice, with the goal of
524 applying these sites in gene therapy [59]. The analysis of insulator-associated DBPs may
525 ultimately contribute to the development of technologies for precise gene expression
526 control. In summary, our novel computational method has identified
527 insulator-associated DBPs—including previously characterized factors—and revealed
528 that the directional bias of these DBPs is potentially involved in chromatin looping,
529 phase separation, transcriptional condensate formation, and insulator function through
530 the structural interactions of DBPs. These findings will contribute to a more
531 comprehensive and global understanding of transcriptional regulation and gene
532 expression.
533
534

535 # Methods

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

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

538 To predict DNA-binding sites that may also function as insulator sites in the
539 human genome, we modified a previously developed deep learning method originally
540 designed to predict transcript expression levels across tissues and cell types based on
541 DNA-binding sites of DBPs in promoter regions and other genomic regions (Fig. 2a;
542 Fig. S4) [14]. To minimize the influence of cellular heterogeneity inherent in tissue
543 samples, we utilized gene expression data from four distinct human cell types for which

544 extensive experimental datasets are publicly available: human foreskin fibroblasts
545 (HFFs; GSM941744), monocytes (Blueprint RNA-seq FPKM data, file
546 'C0010KB1.transcript_quantification.rsem_grape2_crg.GRCh38.20150622.results',
547 available at http://dcc.blueprint-epigenome.eu/#/files), human mammary epithelial cells
548 (HMECs; 'wgEncodeCshlLongRnaSeqHmecCellPapAlnRep1.bam', ENCODE project
549 via UCSC Genome Browser http://genome.ucsc.edu/encode/), and neural progenitor
550 cells (NPCs; GSM915326, ENCFF529SIO). FPKM (Fragments Per Kilobase of exon
551 per Million mapped reads) values representing gene expression levels were retrieved or
552 calculated based on GENCODE v19 annotations from the ENCODE portal [60]
553 (https://www.encodeproject.org/) using the RSeQC tool [61]
554 (http://rseqc.sourceforge.net/). Following the data processing pipeline described by
555 Tasaki et al. [14], we calculated $\log_2$-transformed FPKM values (adding 0.25 to avoid
556 negative infinity), determined median $\log_2$-FPKM values across the four cell types, and
557 computed $\log_2$-fold changes relative to the median expression for use as outputs in the
558 deep learning model.
559

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

561 To investigate the effect of insulator sites on EPIs and gene expression, the
562 DNA-binding sites in putative enhancer and open chromatin regions located away from
563 promoter regions were added as input data in the deep learning method. To identify the
564 DNA-binding sites of DBPs in promoter and enhancer regions, ChIP-seq data of the
565 DBPs were used and were updated to GTRD database release v20.06

566 (https://gtrd.biouml.org) [62]. The binding site of a DBP regulating the expression level of
567 each transcript was selected using eQTL data (GTEx_Analysis_v8_eQTL.tar) from
568 GTEx portal release v8 (https://gtexportal.org). If a DNA-binding site was located
569 within a 100-bp region around a single-nucleotide polymorphism (SNP) of an eQTL,
570 we assumed that the DBP regulated the expression of the transcript corresponding to the
571 eQTL. If the SNP of an eQTL was not located on the same chromosome as the
572 transcribed region of the eQTL (i.e., *trans*-eQTL), the eQTL data were not used. The
573 genomic positions of the eQTLs were converted to the corresponding positions in the
574 hg19 version of the human genome using the LiftOver tool
575 (http://hgdownload.soe.ucsc.edu/downloads.html#liftover).
576 The original model of the input data of DNA-binding sites in promoter regions
577 used for the deep learning analysis consisted of the lengths of a DBP ChIP-seq peak in
578 each 100-base bin in a region between 2 kbp upstream and 1 kbp downstream of the
579 transcription start site (TSS) of each gene or transcript [14]. To add the input data of
580 DNA-binding sites of DBPs in more upstream and downstream regions of a TSS and to
581 reduce the number of bins, we modified the model and first computed the length of a
582 DBP ChIP-seq peak in a 100-base bin and then selected the maximum length in each
583 interval of 100 bins in a region between 1 Mbp upstream and downstream of a TSS.
584 This step outputted the 200 longest DBP ChIP-seq peaks in the examined regions.
585
586 **Training step for machine learning analysis**
587 In the training of the modified deep learning model, we used the same test,

588 training, and validation data as in the original method: all 2,705 genes encoded on
589 chromosome 1 as test data, training data (22,251 genes), and validation data (2,472
590 genes) randomly selected from the other genes [14]. The input data for promoters and
591 putative enhancer regions were concatenated into a total of 230 bins of the lengths of
592 1,310 DBP ChIP-seq peaks, consisting of 30 bins for promoters and 200 bins for
593 putative enhancers because the DeepLIFT tool 'RevealCancel rule on dense layers and
594 the Rescale rule on conv layers (genomics default)', with high scoring accuracy, [24]
595 cannot be applied to a neural network with multiple input layers (Fig. S5). Ubuntu
596 (version 20.04), Python (version 3.8.10), TensorFlow (version 2.6.0), NVIDIA CUDA
597 toolkit (version 11.4.2), GeForce RTX 3080 GPU with 10 GB memory, and Intel Core
598 i7-10700 CPU with 128 GB memory were used for training and prediction, and the
599 original scripts were modified to run on newer versions of the software. The training
600 method was the same as the original, and the training time was two to four hours. In this
601 study, we focused on the analysis of HFF cells and assessed the training process using
602 all the input data as the test data, which slightly improved the prediction of gene
603 expression. Therefore, we used the trained model for the analysis of insulator-associated
604 DBPs.
605

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

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

608 To identify the direction of DBP-binding motifs in the DNA-binding sites of
609 predicted insulator-associated DBPs, we searched for DBP-binding motifs in open

610 chromatin regions in HFF cells (GSE170964) using the PIQ tool [17] and the position
611 frequency matrices of the DBP-binding motifs. The motif sequences were collected
612 from databases and previous studies, specifically, TRANSFAC (2019.2), JASPAR
613 (2018), UniPROBE (2018), DNA motifs from high-throughput SELEX, DNA motifs
614 from ENCODE DBP ChIP-seq data, and HOCOMOCO versions 9 and 11 [18-23]. To
615 reduce the number of false-positive prediction of DNA-binding sites, only
616 DNA-binding sites that overlapped the ChIP-seq peaks of the same DBP (GTRD
617 database release v20.06 [https://gtrd.biouml.org]) were used for further analyses. The
618 genomic positions of ChIP-seq and DNase-seq data in public databases were converted
619 to those provided in the hg19 version of the human genome using the LiftOver tool.

620 To confirm that CTCF DNA-binding sites were located at chromatin interaction
621 sites, we examined the overlap between CTCF ChIP-seq peaks and the sites of *in situ*
622 ChIA-PET chromatin interactions mediated by CTCF in human foreskin fibroblast
623 (HFF) cells (4D Nucleome data portal,
624 https://data.4dnucleome.org/files-processed/4DNFI73AFN12/). The directional
625 difference between RAD21 DNA-binding sites located at chromatin interaction sites
626 was also examined using ChIA-PET chromatin interactions mediated by RAD21 in
627 human fibroblast from skin (ENCLB147AJZ). RAD21 DNA-binding sites within 100
628 bp of chromatin interaction sites were associated with chromatin interactions. To
629 eliminate DNA-binding sites that overlapped EPI, we obtained ChIP-seq peaks of
630 histone modifications, such as H3K4me3, H3K27ac, and H3K4me1, associated with
631 promoter and enhancer regions in HFF cells (4D Nucleome data portal,

632 4DNFIJOK1BWR, 4DNFI442BQLO, and 4DNFIAGV3X99).

633

634 **Preparation of input data considering the DNA-binding sites of an**
635 **insulator-associated DBP**

636 To prepare the input data of the DNA-binding sites of DBPs in enhancers and
637 promoters for the deep learning analysis, first, a transcript regulated by a DBP bound to
638 enhancers and promoters was predicted based on the single nearest gene association rule
639 (Fig. 2c; McLean et al. [15] Fig. S2c), as we found that the association rule predicted EPIs
640 more accurately than the other rules in our previous study [15]. The maximum length of
641 the regulatory domain for each transcript was set to 1 Mbp upstream and downstream of
642 the TSS. Then, considering the genomic locations and effects of the insulator sites of a
643 DBP on EPIs, the length of a DBP ChIP-seq peak in a 100-base bin was calculated from
644 a TSS to the DNA-binding site of a predicted insulator-associated DBP. To reduce the
645 computational time of the analysis, we first calculated the length of a DBP ChIP-seq
646 peak in the regulatory domains of all transcripts in a 100-base bin. Then, for each
647 interval of 100 bins between a TSS and a DNA-binding site of a predicted
648 insulator-associated DBP in the regulatory domain, we computed the maximum length
649 of each DBP ChIP-seq peak in 100 bins on the basis of the precomputed length of the
650 DBP ChIP-seq peak in each bin. In the case of a set of 100 bins containing the
651 DNA-binding site of a predicted insulator-associated DBP, we computed the maximum
652 lengths from the bin closest to a TSS to the first bin containing the DNA-binding site of
653 the insulator-associated DBP, taking into account the orientation of the DNA-binding

site.

Prediction of insulator-associated DBPs using DeepLIFT tool

The DeepLIFT tool (version 0.6.13.0) was used to calculate the contribution (DeepLIFT score) of each DNA-binding site to the prediction of the expression level of a transcript [24]. The contribution scores of DNA-binding sites of DBPs in putative enhancers and promoters upstream and downstream of each transcript were normalized by the sum of the absolute values of all contribution scores for the transcript, as the range of contribution scores changes according to the range of the predicted output (i.e., the expression level of a transcript), as described in a previous study [58]. To estimate the effect of the DNA-binding sites of a predicted insulator-associated DBP on the prediction of transcript levels, positive normalized contribution scores for transcripts with DNA-binding sites in upstream or downstream regions were subjected to comparisons between all combinations of two different orientations (i.e., all combinations of two orientations among FR, RF, forward-forward [FF], and reverse-reverse [RR] orientations, and no orientation [X]: FR vs. RF, FR vs. X, RF vs. X, FF vs. RR, FF vs. X, RR vs. X, FF vs. FR, FF vs. RF, RR vs. FR, and RR vs. RF) among the DNA-binding sites of a predicted insulator-associated DBP by using a statistical test (Mann–Whitney *U* test, FDR-adjusted p value < 0.05 and p value < 10^{-3}). For EPIs that were common to the predictions based on two different orientations of DNA-binding sites of a predicted insulator-associated DBP, the DNA-binding sites of DBPs in putative enhancers and promoters among the common EPIs were omitted in

676 the comparison of positive normalized contribution scores between the two different
677 orientations. Based on the results of the statistical analysis and the mean differences in
678 contribution scores, we predicted an insulator-associated DBP on the basis of the
679 directional bias or lack of orientation bias of DNA-binding sites; for example, to predict
680 an insulator-associated DBP with a DNA-binding site in the FR orientation, we selected
681 a DBP for which the contribution scores among all combinations of FR vs. RF, FR vs.
682 FF, FR vs. RR, and FR vs. X were significant (Mann–Whitney *U* test, FDR-adjusted *p*
683 value < 0.05 and *p* value $< 10^{-3}$) and showed positive (high) mean differences among
684 the contribution scores of all the combinations.
685
686 **Search for insulator-associated DNA-binding sites based on histone modifications**
687 **and gene expression levels**
688 The genomic positions in the BAM files of H3K27me3 and H3K9me3 ChIP-seq
689 data (ENCFF267JBO, ENCFF528UZJ) from HFF cells were converted to the
690 corresponding positions in the hg19 version of the human genome using the CrossMap
691 tool [63] and then converted to WIG files using bam2wig.pl from Bio::ToolBox
692 (https://github.com/tjparnell/biotoolbox) to visualize the data on the UCSC Genome
693 Browser through custom tracks (http://www.genome.ucsc.edu/cgi-bin/hgCustom) and a
694 track hub (https://genome.ucsc.edu/cgi-bin/hgHubConnect). The BAM file of RNA-seq
695 data (GSM941744) of HFF cells was converted to a WIG file using bam2wig.pl in
696 Bio::ToolBox. The genomic positions in the BED file of the boundaries and the
697 distribution of insulation scores identified through a Micro-C assay (4D Nucleome data

698 portal, https://data.4dnucleome.org/files-processed/4DNFI5OZ7TBE/ and
699 4DNFIPMHBA41) were converted to the corresponding positions in the hg19 version
700 of the human genome.
701
702 **Distribution of DNA-binding sites around chromatin interaction sites**
703 We examined the distribution of the DNA-binding sites of predicted
704 insulator-associated DBPs in regions within ±500 bp of the center of a chromatin
705 interaction site using chromatin interaction data from HFF cells obtained via a Micro-C
706 assay (4D Nucleome data portal,
707 https://data.4dnucleome.org/files-processed/4DNFI1MCF9GZ/). For the
708 insulator-associated DBPs other than CTCF, RAD21, and SMC3, the DNA-binding
709 sites that do not overlap with those of CTCF, RND21, and SMC3 were used to examine
710 their distribution around interaction sites. The number of DNA-binding sites of
711 predicted insulator-associated DBPs that cooccurred with DNA-binding sites of known
712 insulator-associated DBPs, such as CTCF, RAD21, and SMC3, was counted in each
713 10-bp bin in the regions within ±500 bp of the center of a chromatin interaction site.
714
715 **Insulator-associated DNA-binding sites near splice sites**
716 Insulator-associated DNA-binding sites were searched within 200 bp of the center
717 of splice sites using the file of the genomic positions of alternative transcripts in the
718 UCSC Genome Browser
719 (https://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/all_mrna.txt.gz ;

720 https://genome.ucsc.edu/cgi-bin/hgTables?db=hg19&hgta_group=rna&hgta_track=mrna&hgta_table=all_mrna&hgta_doSchema=describe+table+schema). We performed the
721
722 statistical test to estimate the enrichment of insulator-associated DNA-binding proteins
723 at the splicing sites compared with the other DNA-binding proteins (Mann–Whitney *U*
724 test, *p* value < 0.05).
725
726 **Statistical analysis**
727 Comparison of the distribution of DeepLIFT scores between different orientations
728 of DNA-binding sites was performed using the Mann-Whitney *U* test with the
729 scipy.stats.mannwhitneyu library. Two-tailed *p* values were calculated. For multiple
730 comparisons, adjusted *p* values were calculated by false discovery rate (FDR) using the
731 Benjamini-Hochberg procedure. *Z* scores for the Mann-Whitney *U* test were not output
732 by the library. Therefore, we calculated *z* scores by z-standardizing the *U* values. The
733 code is made available on GitHub at https://github.com/reposit2/insulator/tree/main.
734
735 **Data availability**
736 Human foreskin fibroblast (HFF) cell RNA-seq data were obtained from GSM941744.
737 Human monocyte RNA-seq data were obtained from the Blueprint DCC portal
738 http://dcc.blueprint-epigenome.eu/#/files at
739 'C0010KB1.transcript_quantification.rsem_grape2_crg.GRCh38.20150622.results'. The
740 RNA-seq data of human mammary epithelial cells (HMEC) were found on ENCODE at
741 UCSC http://genome.ucsc.edu/encode/ at

742 'wgEncodeCshlLongRnaSeqHmecCellPapAlnRep1.bam'. Human neural progenitor cell
743 (NPC) RNA-seq data were obtained from GSM915326. The DNase-seq data of HFF
744 cells were obtained from GSE170964. The *in situ* ChIA-PET chromatin interaction data
745 mediated by CTCF in HFF cells were available at the 4D Nucleome data portal,
746 https://data.4dnucleome.org/files-processed/4DNFI73AFN12/. The ChIP-seq data of
747 H3K4me3, H3K27ac, and H3K4me1 of HFF cells were found at the 4D Nucleome data
748 portal, 4DNFIJOK1BWR, 4DNFI442BQLO, and 4DNFIAGV3X99. The ChIP-seq data
749 of H3K27me3 and H3K9me3 in HFF cells were obtained from the ENCODE portal,
750 ENCFF267JBO and ENCFF528UZJ. The chromatin interaction data and boundary data
751 of HFF cells by the Micro-C were obtained from the 4D Nucleome data portal,
752 https://data.4dnucleome.org/files-processed/4DNFI1MCF9GZ/ and
753 https://data.4dnucleome.org/files-processed/4DNFI5OZ7TBE/. ChIP-seq data of TFs
754 were downloaded from the v20.06 release of the GTRD database
755 (https://gtrd.biouml.org). eQTL data (GTEx_Analysis_v8_eQTL.tar) were downloaded
756 from the v8 release of the GTEx portal (https://gtexportal.org). DNA binding motifs of
757 TFs were collected from TRANSFAC (2019.2), JASPAR (2018)
758 (https://jaspar.genereg.net/), UniPROBE (2018)
759 (http://the_brain.bwh.harvard.edu/uniprobe/), high-throughput SELEX
760 (https://doi.org/10.1016/j.cell.2012.12.009), transcription factor binding sequences of
761 ENCODE TF ChIP-seq data (http://compbio.mit.edu/encode-motifs/), and
762 HOCOMOCO version 9 and 11 (https://autosome.org/HOCOMOCOS/)
763 (https://hocomoco11.autosome.org/). The genomic positions of alternative transcripts in

764 the UCSC Genome Browser

765 (https://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/all_mrna.txt.gz).

766

767 **Code availability**

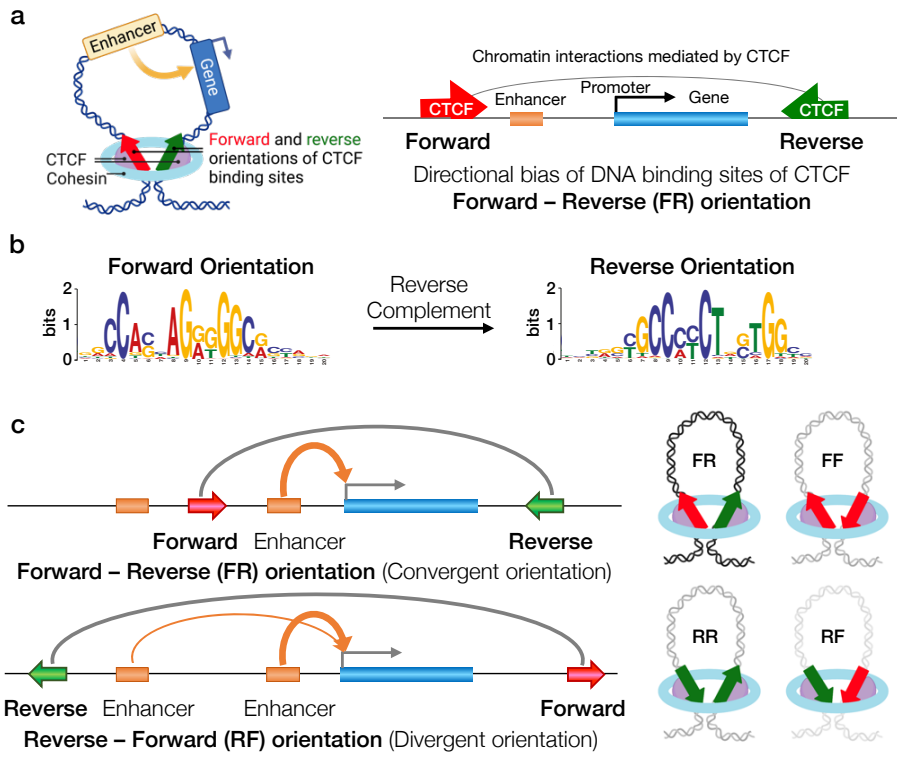
768 Demonstration codes are made available at

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

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

771

772 **Figures**



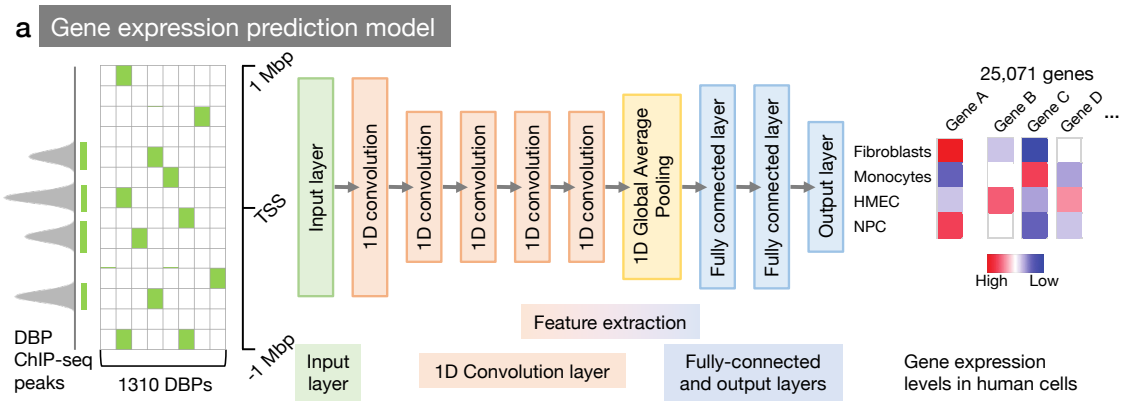
773

774 **Fig. 1. CTCF insulator function mediates enhancer–promoter interactions. a** CTCF

775 DNA-binding sites with a forward–reverse (FR) orientation are frequently observed at

776 chromatin interaction sites. The forward and reverse orientations correspond to

complementary sequences, which are clearly distinguishable in the reference genome. **b** Position weight matrices (PWMs) of CTCF binding motifs. The forward and reverse motifs are complementary on opposite DNA strands. A sequence logo is a graphical representation of nucleotide conservation, where the height of each letter reflects its frequency at that position (adapted from Wikipedia 'Sequence logo'). **c** CTCF can block interactions between an enhancer and a promoter (or the transcription start site [TSS] of a gene), thereby restricting enhancer activity within a defined chromatin domain (region between red and green arrows). The red and green arrows indicate the DNA-binding sites of a putative insulator-associated DBP, such as the FR orientation of CTCF sites illustrated here. The numbers of DNA-binding sites with FR, RF, FF, and RR orientations differ, as reflected by the intensity of chromatin loops (right panel). Images of double-stranded DNA loops were adapted from "Transcriptional Regulation by CTCF and Cohesin" at BioRender.com (2023).



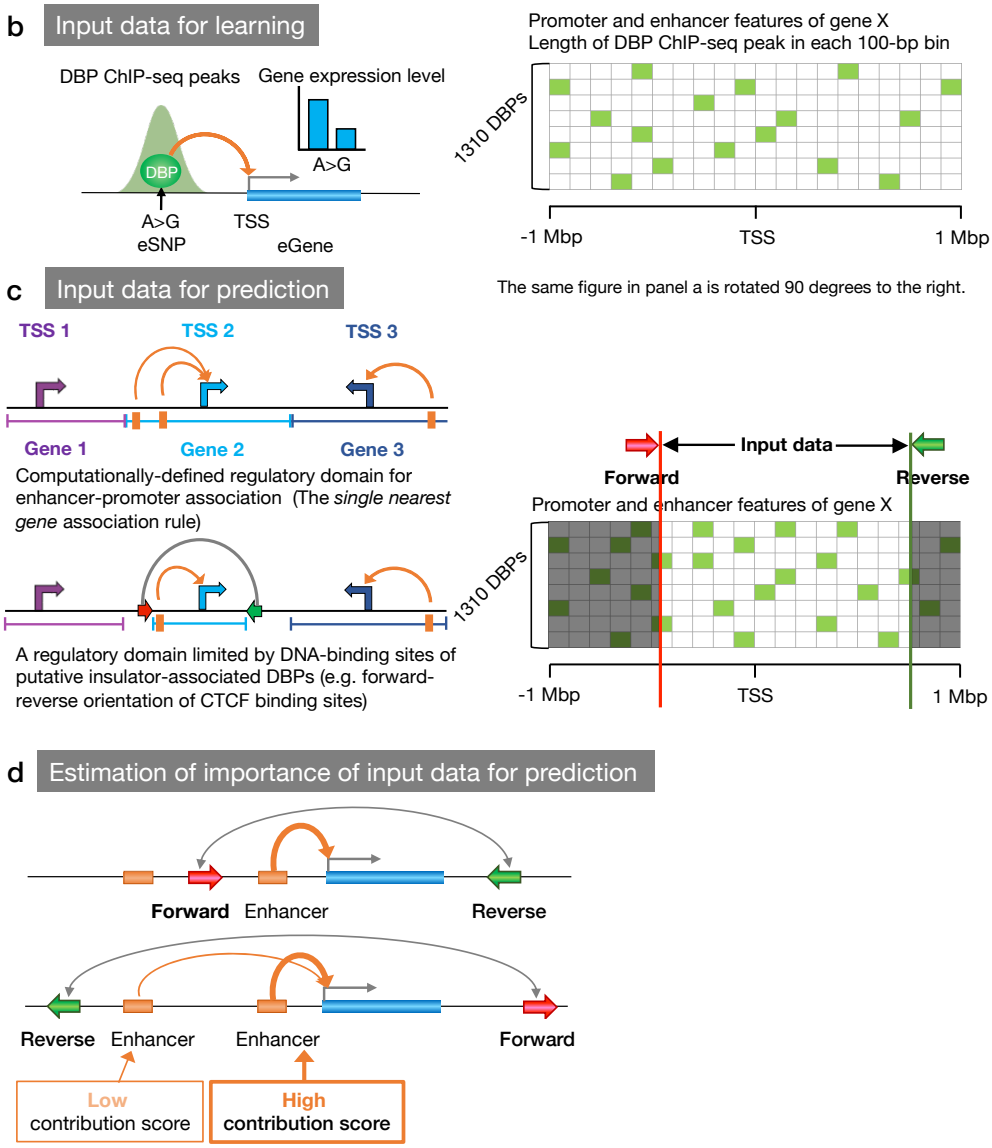


Fig. 2. Overview of the prediction pipeline for insulator-associated DBPs. a We employed a machine learning method, DEcode, with modifications. Gene expression levels in human cells were predicted using the genomic locations of DBP binding sites at promoters and putative enhancers as input features. **b** In the machine learning phase, a DBP bound to a putative enhancer overlapping an eSNP (expression quantitative trait locus SNP) was linked to a gene (eGene) potentially regulated by the DBP, based on eQTL data. For each gene, DBP ChIP-seq peak locations were binned (green boxes) and

802 summarized in a matrix as input: Promoters were represented by 30 bins of 100 bp each
803 spanning the region from -2 kbp to +1 kbp relative to the TSS. Enhancer–gene
804 regulatory regions were covered by 200 bins of 10 kbp each, spanning -1 Mbp to +1
805 Mbp from the TSS, excluding the promoter. **c** For prediction of putative
806 insulator-associated DBPs, a DBP bound to a putative enhancer was associated with a
807 gene based on a computationally defined regulatory domain, independent of eQTL data.
808 The single nearest gene association rule defines a regulatory domain as the region
809 between the TSS of a gene and the midpoints to the TSS of the nearest upstream and
810 downstream genes. The TSSs and regulatory domains for each gene are indicated with
811 purple, light blue, or blue arrows and bracketed lines, respectively. The regulatory
812 domain for enhancer–promoter associations was further constrained by DNA-binding
813 sites (red and green arrows) of putative insulator-associated DBPs. For each gene, DBP
814 ChIP-seq peak locations were binned (green boxes in non-gray regions) and
815 summarized in a matrix as input data. **d** Estimation of the directional effect of
816 DNA-binding sites (red and green arrows) of putative insulator-associated DBPs on
817 gene expression prediction. A DBP whose binding (orange box) is highly informative
818 for gene expression prediction receives a high contribution score.
819
820

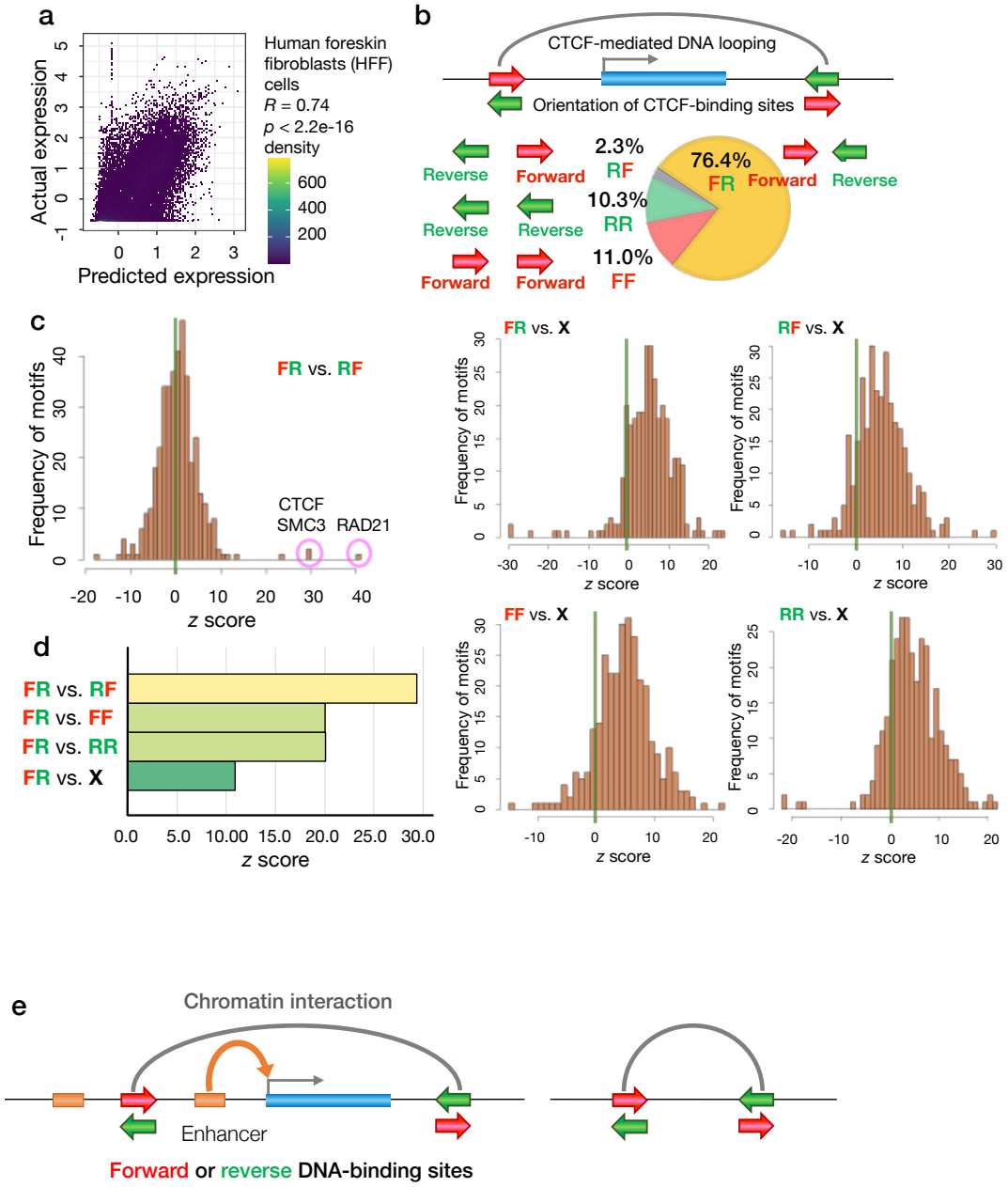
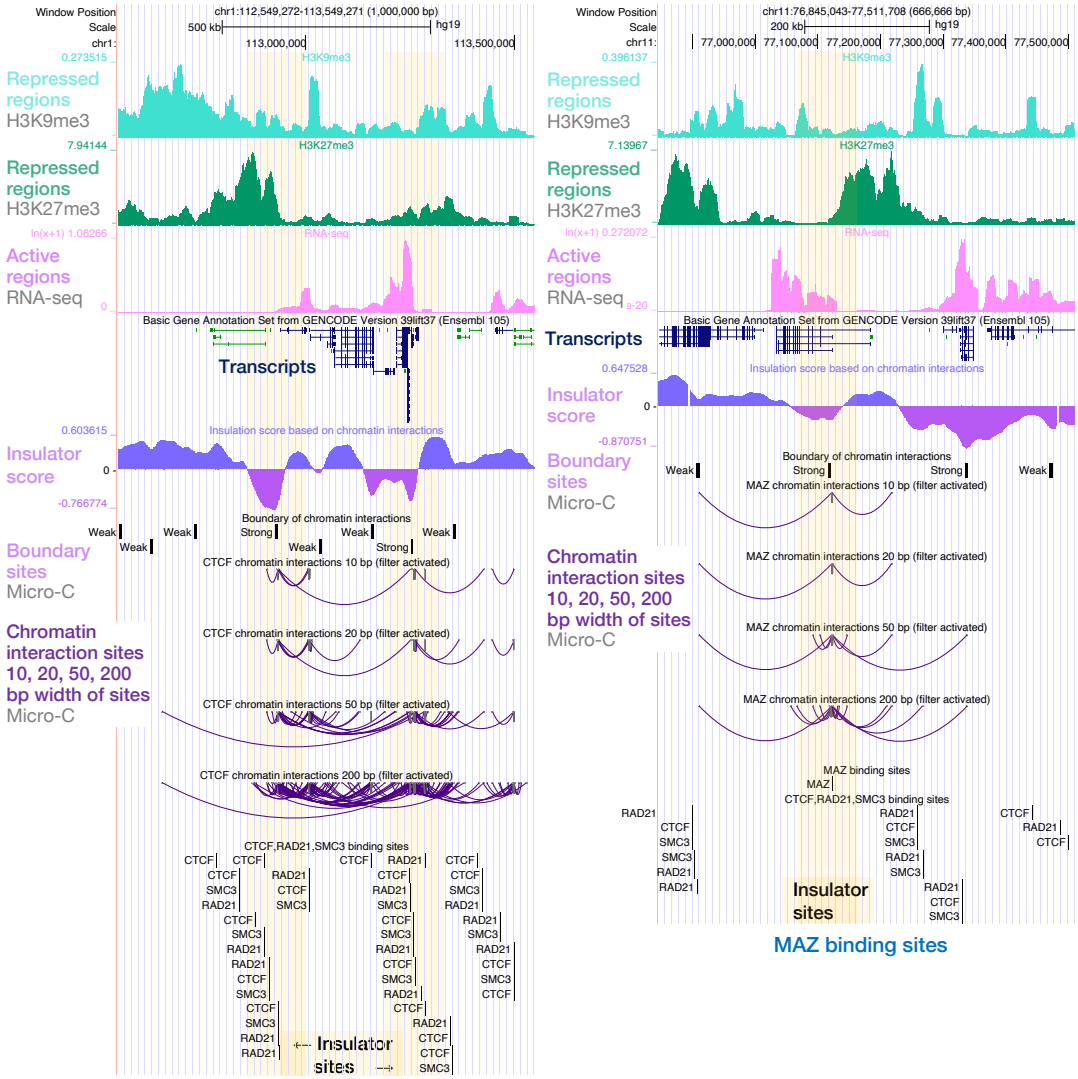


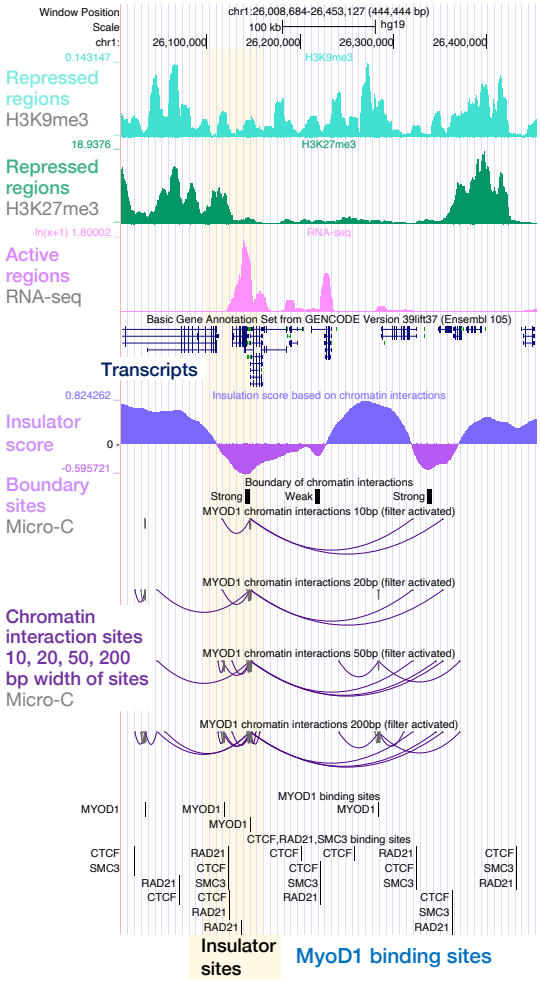
Fig. 3. Estimation of the effect of directional regulatory motifs on gene expression. **a** The predicted $\log_2$-transcripts per million ($\log_2$-TPM) values of genes were compared with actual expression levels in human foreskin fibroblast (HFF) cells using Spearman's rank correlation coefficient. **b** The forward–reverse (FR) orientation of CTCF-binding sites is predominant (76.4%) at chromatin interaction sites, while forward–forward (FF),

829 reverse–reverse (RR), and reverse–forward (RF) orientations are less frequent (11.0%,
830 10.3%, and 2.3%, respectively) (Guo Y et al., Cell 2015) [3]. **c** Distribution of z-scores
831 from the Mann–Whitney U test for DeepLIFT scores comparing FR and RF orientations
832 of protein DNA-binding sites. The right panel displays the distribution of z-scores
833 comparing non-directional (X) sites with directional orientations (FR, RF, FF, or RR).
834 Results are shown according to Selection Criteria S2 in Supplementary Fig. 1 and Note.
835 **d** *Z*-scores from the Mann–Whitney U test for DeepLIFT scores in pairwise
836 comparisons of CTCF-binding sites with four different orientations (FR, RF, FF, RR)
837 and non-directional (X) sites. The differences in *z*-scores mirrored the percentages of
838 CTCF-binding site orientations at chromatin interaction sites. **e** The DNA-binding sites
839 of insulator-associated proteins were identified at both sites or at either site within pairs
840 of long-range chromatin interactions.
841
842

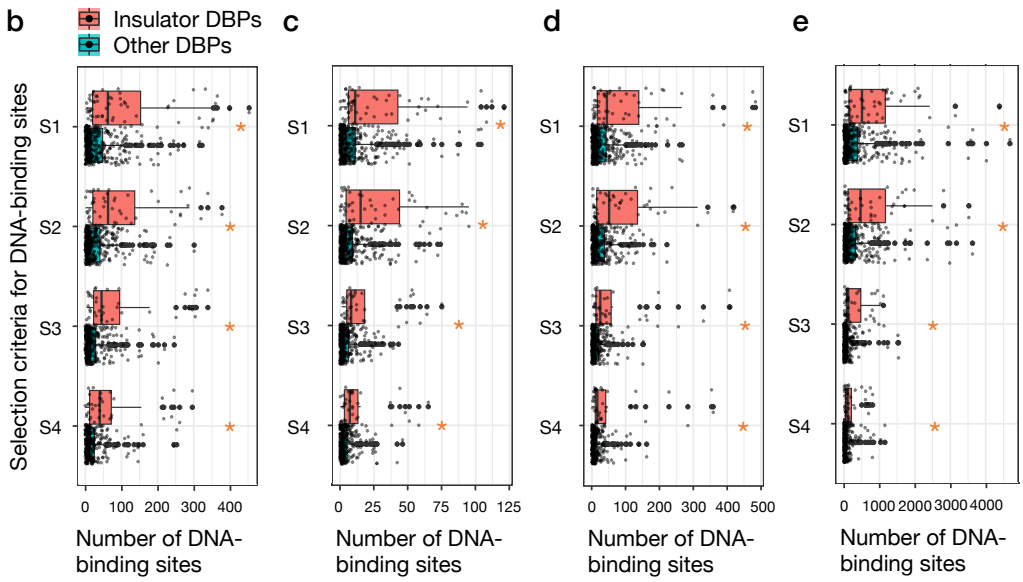
a


Window Position
chr1:112,549,272-113,549,271 (1,000,000 bp)
Scale
500 kb
hg19
chr1:
113,000,000
113,500,000
Repressed regions H3K9me3
H3K9me3
Repressed regions H3K27me3
H3K27me3
Active regions RNA-seq
RNA-seq
Basic Gene Annotation Set from GENCODE Version 39lift37 (Ensembl 105)
Transcripts
Insulation score based on chromatin interactions
Insulator score
Boundary of chromatin interactions
Boundary sites Micro-C
Weak
Strong
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
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
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)
MAZ binding sites
Insulator sites
MAZ binding sites

843


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
Repressed regions H3K9me3
H3K9me3
Repressed regions H3K27me3
H3K27me3
Active regions RNA-seq
RNA-seq
Basic Gene Annotation Set from GENCODE Version 39lift37 (Ensembl 105)
Transcripts
Insulation score based on chromatin interactions
Insulator score
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

844


b
Insulator DBPs
Other DBPs
c
d
e
Selection criteria for DNA-binding sites
S1
S2
S3
S4
0 100 200 300 400
0 25 50 75 100 125
0 100 200 300 400 500
0 1000 2000 3000 4000
Number of DNA-binding sites

845

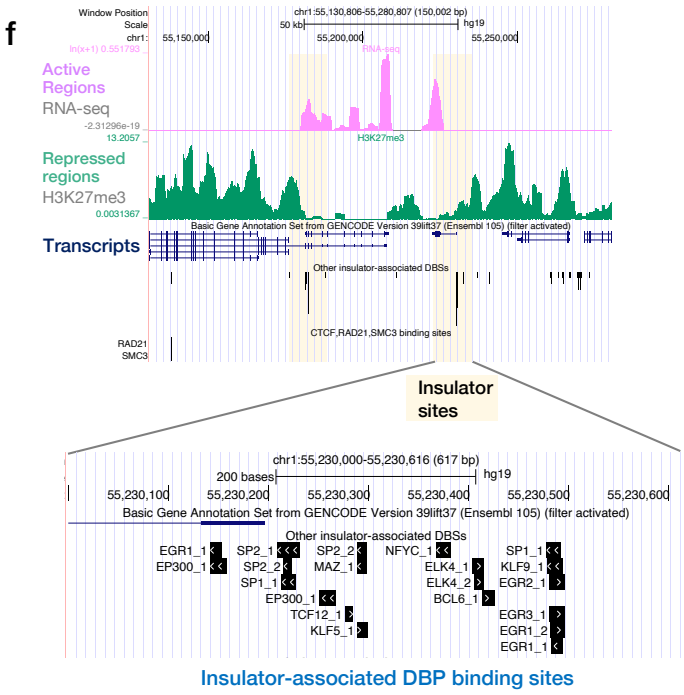


Fig. 4. DNA-binding sites of predicted insulator-associated DBPs at potential insulator sites. a Distribution of histone modification marks H3K27me3 (green) and H3K9me3 (turquoise), along with transcript levels (pink), in the upstream and downstream regions of a potential insulator site (light orange). These histone modification marks are characteristic of transcriptionally repressed regions. Insulator sites are positioned at the boundary between transcriptionally active and repressed regions (indicated by light orange vertical bars). Blue and green lines in the GENCODE track denote gene transcripts. Insulator scores and boundary regions were estimated from chromatin interaction data available via the 4D Nucleome Data Portal. Dark blue arcs represent chromatin interactions within 10 bp, 20 bp, 50 bp, and 200 bp of the binding sites of CTCF (upper left panel), MAZ (upper right panel), and MyoD1 (lower left panel), respectively. The binding sites of CTCF, RAD21, SMC3 (upper left), MAZ

859 (upper right), and MyoD1 (lower left) are displayed as black bars. Analyses were
860 performed using the Homo sapiens genome assembly GRCh37 (hg19) from the
861 Genome Reference Consortium. **b** Number of DNA-binding sites of predicted
862 insulator-associated DBPs where H3K27me3 signal intensities differed between
863 upstream and downstream regions. * indicates $p < 0.01$ (Mann–Whitney U test) (Table
864 2). Boxplots show the median (centerline), interquartile range (boxes), whiskers (up to
865 1.5× IQR), and all outliers (points). All data points are shown. **c** Number of
866 DNA-binding sites of predicted insulator-associated DBPs where both H3K27me3
867 signal intensities and transcript levels differed between upstream and downstream
868 regions. **d** Number of DNA-binding sites of predicted insulator-associated DBPs located
869 within boundary regions assigned by chromatin interaction data from human foreskin
870 fibroblast (HFF) cells. **e** Number of DNA-binding sites of predicted insulator-associated
871 DBPs overlapping within ±200 bp of splice sites. Insulator-associated DBP binding sites
872 were found within this window significantly more frequently than for other
873 DNA-binding proteins. **f** Cluster of DNA-binding sites of insulator-associated DBPs.
874
875

Fig. 5. Insulator-associated DBPs potentially regulate alternative transcription. Distribution of the H3K27me3 histone modification mark (green) and transcript levels (pink) is displayed. The H3K27me3 mark is characteristic of transcriptionally repressed regions. Insulator sites are positioned at the boundaries between transcriptionally active and repressed regions (indicated by light orange vertical bars). Blue lines in the GENCODE track indicate gene transcripts. Chromatin interactions overlapping a 200-bp window centered on the binding sites of RBPJ (left panel) and SP1 (right panel) are shown as dark blue arcs. The binding sites of RBPJ (left), SP1 (right), and CTCF, RAD21, SMC3 (both panels) are shown as black bars. The DNA-binding sites of insulator-associated DBPs were frequently observed near transcription start and termination sites. These results suggest that insulator-associated DBPs may differentially regulate transcribed regions that are repressed in alternative transcripts.

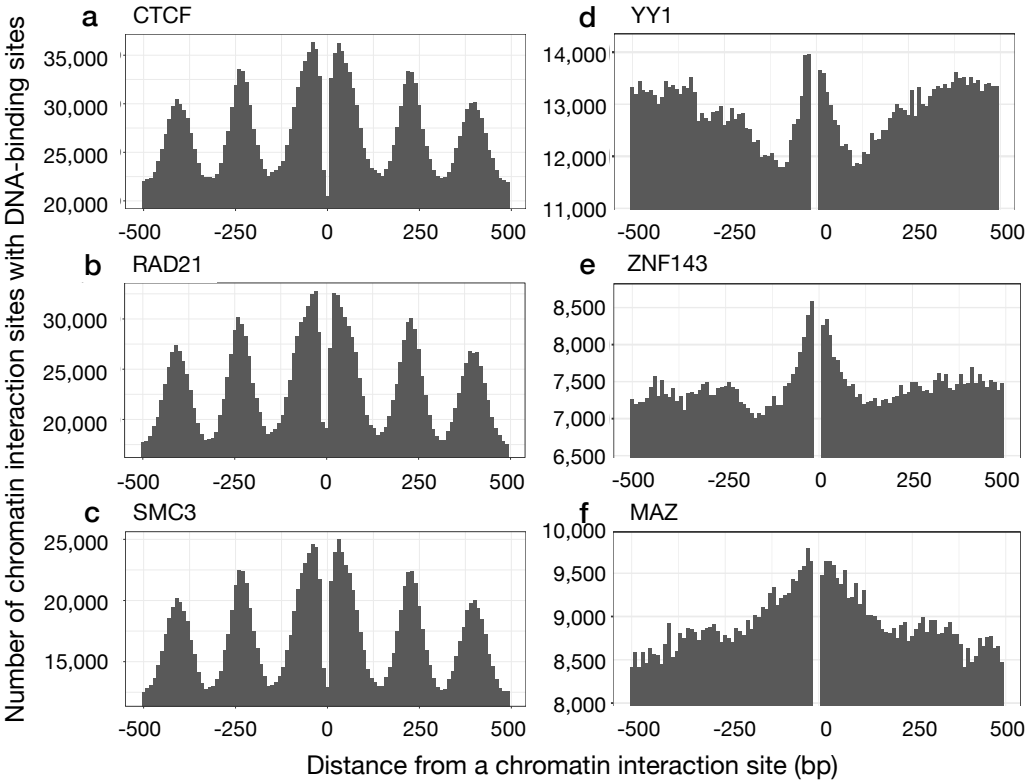


Fig. 6. Distribution of DNA-binding sites around chromatin interaction sites. a-f The x-axis represents the distance of a DNA-binding site from the center of a chromatin interaction site, while the y-axis indicates the number of chromatin interaction sites with DNA-binding sites located within a ±500 bp region. For the analyses of YY1, ZNF143, and MAZ, DNA-binding sites overlapping those of CTCF, RAD21, and SMC3 were excluded.

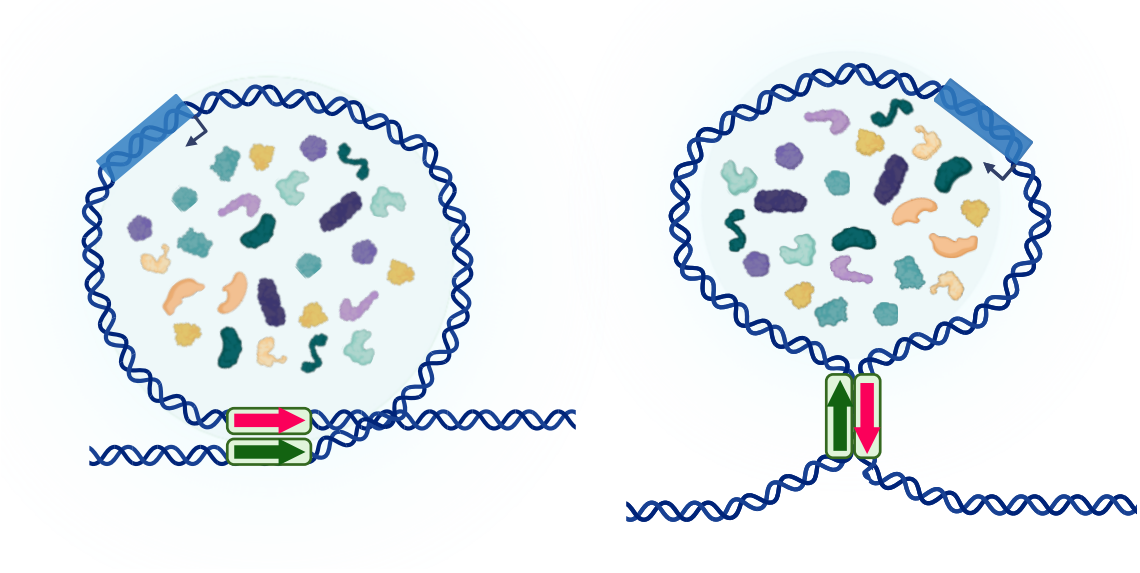


Fig. 7. Boundary-pairing (insulator-pairing) model for the regulation of enhancer –promoter interactions. Molecules bound to insulators physically pair with their partners, either head-to-head or head-to-tail, with varying specificity at the termini of topologically associating domains (TADs) in flies [45-47]. Although the precise mechanism by which partners locate each other remains unclear, it is unlikely to require loop extrusion. Homologous and heterologous insulator–insulator pairing interactions are central to the architectural functions of insulators, and these interactions are orientation-dependent. Homologous and heterologous interactions between insulator-associated DNA-binding proteins may similarly occur in humans. Chromatin looping can induce the formation of transcriptional condensates and phase separation, in addition to cohesin–CTCF-anchored chromatin looping [40-42]. Figure created with BioRender.com.

915 # Tables

916 **Table 1. Predicted insulator-associated DNA-binding proteins (DBPs) with**

917 **directional or nondirectional binding site preferences.**

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

918

919

920 *Z*-scores from the Mann–Whitney U test compare DeepLIFT scores between
921 forward–reverse (FR) and other orientations—reverse–forward (RF), forward–forward
922 (FF), reverse–reverse (RR), and non-oriented (X)—for protein DNA-binding sites.
923 Mean differences in DeepLIFT scores are also reported. Color scales represent values
924 for each DBP. Rows (Top, Middle, Bottom) correspond to: Top—Selection Criteria S1
925 (Supplementary Fig. 1); Middle—Selection Criteria S2 (Supplementary Fig. 1);
926 Bottom—List of predicted insulator-associated DBPs. DBPs highlighted in yellow are
927 associated with insulator function, CTCF, and/or other transcriptional regulatory
928 mechanisms known to affect enhancer–promoter interactions (EPIs) in previous studies.

929

930

931 **Table 2. Number of DNA-binding sites for predicted insulator-associated DBPs at**
932 **insulator sites and splice sites.**

933

		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	

934 Selection Criteria:

935 S1: DNA-binding sites of predicted insulator-associated DBPs.

936 S2: DNA-binding sites of predicted insulator-associated DBPs that overlap with a
937 chromatin interaction site located more than 5 kb upstream or downstream of a
938 transcribed region.

939 S3: DNA-binding sites of predicted insulator-associated DBPs that do not overlap with
940 ChIP-seq peaks for H3K4me3, H3K4me1, or H3K27ac.

941 S4: DNA-binding sites of predicted insulator-associated DBPs that both overlap a
942 chromatin interaction site (located more than 5 kb upstream or downstream of a
943 transcript) and do not overlap ChIP-seq peaks for H3K4me3, H3K4me1, or H3K27ac.
944 *P*-values were calculated using the Mann–Whitney U test.
945
946
947 **Table 3. Pairs of DBPs that colocalize at equivalent distances from chromatin**
948 **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

949

950 Distance to start and end (bp) indicates the genomic distance (in base pairs) from the
951 center of a chromatin interaction site to the 10-bp bin with the highest number of DBP1
952 and DBP2 binding sites, among 100 bins (covering a 1-kbp region) surrounding the site.
953 A positive value denotes the distance to the downstream side of an insulator site, while a
954 negative value indicates the upstream side. # chromatin interactions with DBP1 and
955 DBP2 sites shows the number of chromatin interaction sites that contain both DBP1 and
956 DBP2 binding sites within ±500 bp of the interaction center. Ratio (DBP1 and 2)/DBP2
957 refers to the ratio of chromatin interaction sites containing both DBP1 and DBP2
958 binding sites to the total number of interaction sites with DBP2 binding sites.

959

960

961 **References**

962 1 Weintraub, A. S. *et al.* YY1 Is a Structural Regulator of Enhancer-Promoter
963 Loops. *Cell* **171**, 1573-1588 e1528, doi:10.1016/j.cell.2017.11.008 (2017).
964 2 Rao, S. S. *et al.* A 3D map of the human genome at kilobase resolution reveals
965 principles of chromatin looping. *Cell* **159**, 1665-1680,
966 doi:10.1016/j.cell.2014.11.021 (2014).
967 3 Guo, Y. *et al.* CRISPR Inversion of CTCF Sites Alters Genome Topology and
968 Enhancer/Promoter Function. *Cell* **162**, 900-910, doi:10.1016/j.cell.2015.07.038
969 (2015).
970 4 Ortabozkoyun, H. *et al.* CRISPR and biochemical screens identify MAZ as a
971 cofactor in CTCF-mediated insulation at Hox clusters. *Nature genetics* **54**,
972 202-212, doi:10.1038/s41588-021-01008-5 (2022).
973 5 Goel, V. Y., Huseyin, M. K. & Hansen, A. S. Region Capture Micro-C reveals
974 coalescence of enhancers and promoters into nested microcompartments. *Nature*
975 *genetics*, doi:10.1038/s41588-023-01391-1 (2023).
976 6 Bailey, S. D. *et al.* ZNF143 provides sequence specificity to secure chromatin

977 interactions at gene promoters. *Nat Commun* **2**, 6186, doi:10.1038/ncomms7186
978 (2015).
979 7 Lis, M. & Walther, D. The orientation of transcription factor binding site motifs
980 in gene promoter regions: does it matter? *BMC Genomics* **17**, 185,
981 doi:10.1186/s12864-016-2549-x (2016).
982 8 Nora, E. P. *et al.* Molecular basis of CTCF binding polarity in genome folding.
983 *Nat Commun* **11**, 5612, doi:10.1038/s41467-020-19283-x (2020).
984 9 Li, Y. *et al.* The structural basis for cohesin-CTCF-anchored loops. *Nature* **578**,
985 472-476, doi:10.1038/s41586-019-1910-z (2020).
986 10 Pugacheva, E. M. *et al.* CTCF mediates chromatin looping via N-terminal
987 domain-dependent cohesin retention. *Proc Natl Acad Sci U S A* **117**, 2020-2031,
988 doi:10.1073/pnas.1911708117 (2020).
989 11 Tao, H. *et al.* Computational methods for the prediction of chromatin interaction
990 and organization using sequence and epigenomic profiles. *Brief Bioinform* **22**,
991 doi:10.1093/bib/bbaa405 (2021).
992 12 Zhang, K., Li, N., Ainsworth, R. I. & Wang, W. Systematic identification of
993 protein combinations mediating chromatin looping. *Nat Commun* **7**, 12249,
994 doi:10.1038/ncomms12249 (2016).
995 13 Xie, X. *et al.* Systematic discovery of regulatory motifs in human promoters and
996 3' UTRs by comparison of several mammals. *Nature* **434**, 338-345,
997 doi:10.1038/nature03441 (2005).
998 14 Tasaki, S., Gaiteri, C., Mostafavi, S. & Wang, Y. Deep learning decodes the
999 principles of differential gene expression. *Nat Mach Intell* **2**, 376-386,
1000 doi:10.1038/s42256-020-0201-6 (2020).
1001 15 McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-regulatory
1002 regions. *Nature biotechnology* **28**, 495-501, doi:10.1038/nbt.1630 (2010).
1003 16 Consortium, G. T. The GTEx Consortium atlas of genetic regulatory effects
1004 across human tissues. *Science (New York, N.Y.)* **369**, 1318-1330,
1005 doi:10.1126/science.aaz1776 (2020).
1006 17 Sherwood, R. I. *et al.* Discovery of directional and nondirectional pioneer
1007 transcription factors by modeling DNase profile magnitude and shape. *Nature*
1008 *biotechnology* **32**, 171-178, doi:10.1038/nbt.2798 (2014).
1009 18 Wingender, E., Dietze, P., Karas, H. & Knuppel, R. TRANSFAC: a database on

1010 transcription factors and their DNA binding sites. *Nucleic acids research* **24**,
1011 238-241 (1996).
1012 19 Newburger, D. E. & Bulyk, M. L. UniPROBE: an online database of protein
1013 binding microarray data on protein-DNA interactions. *Nucleic acids research* **37**,
1014 D77-82, doi:10.1093/nar/gkn660 (2009).
1015 20 Portales-Casamar, E. *et al.* JASPAR 2010: the greatly expanded open-access
1016 database of transcription factor binding profiles. *Nucleic acids research* **38**,
1017 D105-110, doi:10.1093/nar/gkp950 (2010).
1018 21 Jolma, A. *et al.* DNA-binding specificities of human transcription factors. *Cell*
1019 **152**, 327-339, doi:10.1016/j.cell.2012.12.009 (2013).
1020 22 Kheradpour, P. & Kellis, M. Systematic discovery and characterization of
1021 regulatory motifs in ENCODE TF binding experiments. *Nucleic acids research*
1022 **42**, 2976-2987, doi:10.1093/nar/gkt1249 (2014).
1023 23 Kulakovskiy, I. V. *et al.* HOCOMOCO: towards a complete collection of
1024 transcription factor binding models for human and mouse via large-scale
1025 ChIP-Seq analysis. *Nucleic acids research* **46**, D252-D259,
1026 doi:10.1093/nar/gkx1106 (2018).
1027 24 Shrikumar, A., Greenside, P. & Kundaje, A. in *Proceedings of the 34th*
1028 *International Conference on Machine Learning* Vol. 70 (eds Precup Doina &
1029 Teh Yee Whye) 3145--3153 (PMLR, Proceedings of Machine Learning Research,
1030 2017).
1031 25 Lee, B. H., Wu, Z. & Rhie, S. K. Characterizing chromatin interactions of
1032 regulatory elements and nucleosome positions, using Hi-C, Micro-C, and
1033 promoter capture Micro-C. *Epigenetics Chromatin* **15**, 41,
1034 doi:10.1186/s13072-022-00473-4 (2022).
1035 26 Cuddapah, S. *et al.* Global analysis of the insulator binding protein CTCF in
1036 chromatin barrier regions reveals demarcation of active and repressive domains.
1037 *Genome research* **19**, 24-32, doi:10.1101/gr.082800.108 (2009).
1038 27 Huang, S., Li, X., Yusufzai, T. M., Qiu, Y. & Felsenfeld, G. USF1 recruits
1039 histone modification complexes and is critical for maintenance of a chromatin
1040 barrier. *Mol Cell Biol* **27**, 7991-8002, doi:10.1128/MCB.01326-07 (2007).
1041 28 Crane, E. *et al.* Condensin-driven remodelling of X chromosome topology
1042 during dosage compensation. *Nature* **523**, 240-244, doi:10.1038/nature14450

1043 (2015).
1044 29 Noda, Y., Okada, S. & Suzuki, T. Regulation of A-to-I RNA editing and stop
1045 codon recoding to control selenoprotein expression during skeletal myogenesis.
1046 *Nat Commun* **13**, 2503, doi:10.1038/s41467-022-30181-2 (2022).
1047 30 Matharu, N. K. & Ahanger, S. H. Chromatin Insulators and Topological
1048 Domains: Adding New Dimensions to 3D Genome Architecture. *Genes (Basel)*
1049 **6**, 790-811, doi:10.3390/genes6030790 (2015).
1050 31 Reyes, A. & Huber, W. Alternative start and termination sites of transcription
1051 drive most transcript isoform differences across human tissues. *Nucleic acids*
1052 *research* **46**, 582-592, doi:10.1093/nar/gkx1165 (2018).
1053 32 Hua, P. *et al.* Defining genome architecture at base-pair resolution. *Nature* **595**,
1054 125-129, doi:10.1038/s41586-021-03639-4 (2021).
1055 33 Xiao, T., Li, X. & Felsenfeld, G. The Myc-associated zinc finger protein (MAZ)
1056 works together with CTCF to control cohesin positioning and genome
1057 organization. *Proc Natl Acad Sci U S A* **118**, doi:10.1073/pnas.2023127118
1058 (2021).
1059 34 Ishihara, K., Oshimura, M. & Nakao, M. CTCF-dependent chromatin insulator
1060 is linked to epigenetic remodeling. *Molecular cell* **23**, 733-742,
1061 doi:10.1016/j.molcel.2006.08.008 (2006).
1062 35 Vierbuchen, T. *et al.* AP-1 Transcription Factors and the BAF Complex Mediate
1063 Signal-Dependent Enhancer Selection. *Molecular cell* **68**, 1067-1082 e1012,
1064 doi:10.1016/j.molcel.2017.11.026 (2017).
1065 36 Gong, Q. H., McDowell, J. C. & Dean, A. Essential role of NF-E2 in remodeling
1066 of chromatin structure and transcriptional activation of the epsilon-globin gene
1067 in vivo by 5' hypersensitive site 2 of the beta-globin locus control region. *Mol*
1068 *Cell Biol* **16**, 6055-6064, doi:10.1128/MCB.16.11.6055 (1996).
1069 37 Davidson, I. F. & Peters, J. M. Genome folding through loop extrusion by SMC
1070 complexes. *Nat Rev Mol Cell Biol* **22**, 445-464,
1071 doi:10.1038/s41580-021-00349-7 (2021).
1072 38 Karollus, A., Mauermeier, T. & Gagneur, J. Current sequence-based models
1073 capture gene expression determinants in promoters but mostly ignore distal
1074 enhancers. *Genome Biol* **24**, 56, doi:10.1186/s13059-023-02899-9 (2023).
1075 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 nano-scale regulatory interactions with variable dependence on cohesin and CTCF. *Nat Commun* **13**, 2139, doi:10.1038/s41467-022-29696-5 (2022).

49 Hsieh, T. S. *et al.* Enhancer-promoter interactions and transcription are largely maintained upon acute loss of CTCF, cohesin, WAPL or YY1. *Nature genetics* **54**, 1919-1932, doi:10.1038/s41588-022-01223-8 (2022).

50 Guerin, T. M., Barrington, C., Pobegalov, G., Molodtsov, M. I. & Uhlmann, F.

An extrinsic motor directs chromatin loop formation by cohesin. *EMBO J* **43**, 4173-4196, doi:10.1038/s44318-024-00202-5 (2024).

51 Wang, R. *et al.* MyoD is a 3D genome structure organizer for muscle cell identity. *Nat Commun* **13**, 205, doi:10.1038/s41467-021-27865-6 (2022).

52 Snead, W. T. & Gladfelter, A. S. The Control Centers of Biomolecular Phase Separation: How Membrane Surfaces, PTMs, and Active Processes Regulate Condensation. *Molecular cell* **76**, 295-305, doi:10.1016/j.molcel.2019.09.016 (2019).

53 Tang, X. *et al.* The PTM profiling of CTCF reveals the regulation of 3D chromatin structure by O-GlcNAcylation. *Nat Commun* **15**, 2813, doi:10.1038/s41467-024-47048-3 (2024).

54 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).

55 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).

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

57 Spurrell, C. H., Dickel, D. E. & Visel, A. The Ties That Bind: Mapping the Dynamic Enhancer-Promoter Interactome. *Cell* **167**, 1163-1166, doi:10.1016/j.cell.2016.10.054 (2016).

58 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).

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.