

1 **Computational prediction of CTCF/cohesin-based intra-TAD loops that insulate**
2 **chromatin contacts and gene expression in mouse liver**

3

4 Bryan J. Matthews and David J. Waxman*

5 Department of Biology and Bioinformatics Program, Boston University, Boston, MA 02215

6

7 * To whom correspondence should be addressed:

8 David J. Waxman

9 Dept. of Biology, Boston University

10 5 Cummington Mall

11 Boston, MA 02215

12 Email: djw@bu.edu

13

14

15 **ABSTRACT**

16 CTCF and cohesin are key drivers of 3D-nuclear organization, anchoring the megabase-scale Topologically
17 Associating Domains (TADs) that segment the genome. Here, we present and validate a computational method
18 to predict cohesin-and-CTCF binding sites that form intra-TAD DNA loops. The intra-TAD loop anchors identified
19 are structurally indistinguishable from TAD anchors regarding binding partners, sequence conservation, and
20 resistance to cohesin knockdown; further, the intra-TAD loops retain key functional features of TADs, including
21 chromatin contact insulation, blockage of repressive histone mark spread, and ubiquity across tissues. We
22 propose that intra-TAD loops form by the same loop extrusion mechanism as the larger TAD loops, and that
23 their shorter length enables finer regulatory control in restricting enhancer-promoter interactions, which enables
24 selective, high-level expression of gene targets of super-enhancers and genes located within repressive nuclear
25 compartments. These findings elucidate the role of intra-TAD cohesin-and-CTCF binding in nuclear organization
26 associated with widespread insulation of distal enhancer activity.

27

28 **INTRODUCTION**

29 The mammalian genome is organized into stereotypical domains, averaging ~700 kb in length, called
30 Topologically Associating Domains (TADs) [1, 2]. TADs are insulated chromatin domains whose genomic
31 boundaries are largely retained across tissues [1] and have been conserved during mammalian evolution [3, 4].
32 TADs provide a stable genomic architecture that constrains enhancer-promoter contacts, while allowing for

33 dynamic tissue-specific interactions that stimulate gene expression within TADs, thereby linking chromatin
34 structure and positioning to gene expression [5, 6].

35

36 Hi-C, an unbiased genome-wide chromosome conformation capture method [7], identifies TADs based on
37 their insulation from inter-domain interactions and by the increased frequency of intra-domain interactions that
38 occurs within individual TADs [1, 2]. TADs show substantial overlap with features of nuclear organization
39 identified using other approaches, including replication domains, lamina-associated domains, and A/B chromatin
40 compartments [4, 8, 9]. TADs impact gene expression by insulation, which limits a given gene's access to
41 regulatory regions [10]. While TAD structures are often shared across tissues within a species, some individual
42 TADs show tissue-specific differences in their spatial positioning within the nucleus, and in their overall activity,
43 transcription factor (TF) binding patterns, and patterns of expression of individual genes [4]. It is unclear to what
44 extent these large megabase-scale chromatin structures exert regulatory control over the multiple, often
45 variably-expressed, genes found within their boundaries.

46

47 Two key protein factors, CCCTC-binding factor (CTCF) and the multi-subunit cohesin complex, are the
48 primary architects of nuclear organization in mammals [11-13]. CTCF and cohesin cooperatively engage
49 genomic DNA via a loop extrusion complex, which is dynamically mobile within TAD boundaries and may help
50 organize TAD structure [12, 14, 15]. CTCF is an 11 zinc finger protein that stably binds DNA and can serve as
51 an insulating enhancer-blocker and a modulator of 3D chromatin structure [16]. Sites bound by both cohesin and
52 CTCF (cohesin-and-CTCF (CAC) sites) are associated with insulator function [5, 17] and are found at TAD
53 boundaries [1, 2]. In contrast, sites bound by cohesin but not CTCF (cohesin-non-CTCF (CNC) sites) are found
54 at tissue-specific promoters and enhancers [18] and may help to stabilize large TF complexes [19]. CAC
55 complexes are also associated with topoisomerase-II β (Top2b), which presumably relieves the torsional strain of
56 the extrusion complex [20].

57

58 Complete knockout of either CTCF or cohesin is embryonic lethal [21-23], whereas partial depletion of CTCF
59 or cohesin results in altered gene expression but has limited phenotypic impact, increasing radiation sensitivity,
60 DNA repair defects, and cell cycle arrest [11, 23, 24]. Complete removal of CTCF or cohesin-related factors,
61 achieved using inducible degradation systems, leads to a complete loss of virtually all loop structures in a highly
62 dosage-dependent manner [25-27]. Mutations affecting CAC loop anchors are frequently seen in cancer and

63 lead to dysregulation of adjacent genes, evidencing the functionality of these loops [28-30]. However, there are
64 many more CAC sites within TADs than at TAD boundaries, and it is not clear what factors differentiate loop-
65 forming CAC sites at TAD boundaries from other CAC sites in the genome.

66

67 Chromatin interactions can be studied by Hi-C analysis, which under standard conditions provides a
68 resolution of 25-100 kb and has been used to study nuclear organization at the level of megabase-scale TAD
69 structures. However, high resolution Hi-C datasets obtained using extreme deep sequencing (>25 billion reads)
70 have led to two key discoveries [15]. First, ~90% of DNA loops (“loop domains”, defined as local peaks in the Hi-
71 C contact matrix) are associated with both CTCF binding and cohesin binding, and 92% of such loops involve
72 inwardly oriented CTCF anchors [15]. Thus, loop anchors are bound at asymmetric CTCF motifs that face the
73 loop interior. This previously unappreciated feature of CTCF loops facilitates the identification of such loops *in*
74 *silico* [12, 31]. Furthermore, expression of neighboring genes changes in a predictable manner when CTCF
75 anchors are inverted or deleted by CRISPR/Cas9 genomic editing [5, 12, 13]. Second, extreme deep
76 sequencing Hi-C studies identify a much larger number of shorter loops than previously recognized (~10,000
77 loops with a median size of 185 kb) [15], many of which represent complex nested structures (e.g., isolated
78 cliques) [12]. The ability to distinguish between such substructures has led to predictions ranging from 10^3 to 10^6
79 loops per genome, depending on the 3C-based analysis method and the cutoff values employed [12, 32-34].
80 The presence of nested loop structures may be a general feature of topological nuclear organization, and the
81 ability to detect such structures is dependent on the method, resolution, and computational approach [32-36].

82

83 While sub topologies within TADs have been observed, it is unknown whether those interactions represent
84 enhancer-promoter loops or other looped structures, and whether they are mediated by cohesin, mediator, or
85 other architectural proteins [17, 37]. Short, < 200 kb CTCF-anchored loops, termed chromatin contact domains
86 or super-enhancer domains, have been identified in mouse embryonic stem cells (mESCs) by ChIA-PET
87 experiments that select for CTCF and cohesin binding sites (via immunoprecipitation of Smc1) [32, 38], and are
88 enriched for tissue-specific genes and enhancers [5, 32]. However, these genomic regions represent a minority
89 of CTCF-anchored DNA loops, and likely do not fully represent all of the nuclear topological domains evident in
90 high resolution Hi-C maps [15, 26, 39]. Given the inability to identify CAC-anchored intra-TAD loops from
91 standard, low resolution Hi-C data, we sought to build on the above advances and develop a computational
92 method to predict such subTAD-scale loops by using only 2D (CTCF and cohesin ChIP-seq binding activity) and

93 1D (CTCF motif orientation) information. Here we define intra-TAD loops anchored by cohesin and CTCF and
94 that contain at least one gene, which represent a superset encompassing super-enhancer and polycomb
95 domains [5]. These CAC-mediated intra-TAD loops are mechanistically distinct from short range enhancer-
96 promoter loops, and from longer range genomic compartmentalization [26, 27, 40], whose impact on gene
97 expression is also discussed.

98

99 Here we present, and then validate in three mouse tissues and two human cell lines, a computational method
100 to identify intra-TAD loops genome-wide. We elucidate the structural and functional features of the intra-TAD
101 loops identified, and those of the better-established TADs, including their impact on gene expression in a mouse
102 liver model. We show that, mechanistically, intra-TAD loops are anchored by loop extrusion CAC complexes
103 that are shared across tissues and show strong conservation. Further, we demonstrate that, at a functional
104 level, intra-TAD loops insulate repressive chromatin mark spread and thereby enable selective expression of
105 genes at a high level compared to their immediate genomic neighbors, notably genes targeted by super-
106 enhancers, and genes that are otherwise found in repressive nuclear compartments. These findings reveal how
107 intra-TAD loops harness many of the same mechanisms as TAD-scale loops but in ways that allow for greater
108 local control of gene expression.

109

110 **RESULTS**

111 **I. Features of TADs and their functional impact on gene expression**

112 *Features associated with TAD boundaries.* We characterized TADs identified in mouse liver [3] using matched
113 ChIP-seq datasets for CTCF and the cohesin subunit Rad21, which we obtained for a group of individual adult
114 male mouse livers. Genomic regions co-bound by cohesin and CTCF (CAC sites) were strongly enriched at
115 TAD boundaries (Figure 1A), consistent with [3]. In contrast, cohesin-non-CTCF sites (CNC sites) were weakly
116 depleted at TAD boundaries. We also observed strong enrichment for motif-oriented CTCF binding at TAD
117 boundaries (Figure 1-figure supplement 1A), consistent with recent reports and the loop extrusion model of
118 domain formation [3, 12, 15]. Next, we explored the impact of cohesin depletion on CAC sites associated with
119 TAD boundaries, following up on the finding that many cohesin binding sites are maintained upon knockout or
120 knockdown of components of the cohesin complex [19]. Figure 1B shows the distribution of cohesin-bound
121 regions that are either resistant or sensitive to haploinsufficiency of the cohesin subunit Rad21 in hepatocytes,
122 or to knockout of the cohesin subunit Stag1 in mouse embryonic fibroblasts (MEFs). In both cell types, sites

123 resistant to cohesin loss are enriched at TAD boundaries, while those sensitive to cohesin loss are more equally
124 distributed along the TAD length. This may explain the unexpected finding that domains and compartments are
125 largely maintained after depletion of cohesin [17, 37, 41].

126

127 Given the frequent conservation of TAD boundaries between tissues in both mouse and human [1, 2, 4], we
128 compared regions bound by CTCF in mouse liver to 15 other mouse tissues from the ENCODE Project [42].
129 CTCF sites that are shared across 12 or more tissues showed 3-4 fold enrichment at TAD boundaries relative to
130 the center of the TAD, whereas CTCF binding sites unique to liver, or shared with only one other tissue, showed
131 no such enrichment (Figure 1C). TAD boundaries were also enriched for CpG hypomethylation, which was most
132 pronounced at TAD anchor CTCF motifs (Figure 1D, Figure 1-figure supplement 1F). CpG methylation is greater
133 at CAC sites not involved in TAD or intra-TAD loop anchors (Figure 1-figure supplement 1F; see below), and
134 could represent an additional layer of epigenetic regulation of CAC-based loop formation. By comparison, the
135 TAD boundary enrichment seen for CTCF and cohesin was absent for >50 other liver-expressed TFs whose
136 binding site distribution within TADs we examined (Figure 1E, and data not shown). Tbp and E2f4, which are
137 characterized by promoter-centric binding [43, 44], are two notable exceptions (Figure 1-figure supplement 1B).
138 Consistent with this, TAD boundaries were enriched for promoters of protein-coding genes, including promoters
139 that do not overlap CAC sites, and for histone marks associated with promoters but not enhancers Figure 1-
140 figure supplement 1C-E).

141

142 *TADs segregate the genome into compartmentalized units.* TADs have the ability to insulate the spread of
143 repressive histone marks and also enhancer-promoter interactions, referred to as enhancer blocking [1, 5, 37].
144 By these dual mechanisms, TADs can exert control over tissue-specific gene expression, despite the TADs
145 themselves being largely structurally invariant across tissues. As TADs are defined based on their insulation of
146 chromatin contacts, we investigated their impact on chromatin mark spread. We examined four broad histone
147 marks associated with either transcriptional repression (H3K9me3, H3K27me3) or activation (H2AK5ac,
148 H3K36me3). We also examined Global Run-on Sequencing data to identify actively transcribed regions of the
149 genome, as well as Lamina Associated Domain (LAD) coordinates to visualize areas of the genome associated
150 with the nuclear periphery. Figure 1F shows a heat map representation of a 1 Mb window around each TAD
151 boundary in mouse liver, clustered using k-means clustering (k=4) based on H3K9me3 and H2AK5ac ChIP-seq
152 data and on the Eigen value of the Hi-C principal component analysis (PCA), which provides an estimate of

153 active versus inactive genomic compartments [7]. A subset comprised of 1,439 liver TAD boundaries (40.9% of
154 all boundaries) represents transitions from inactive to active chromatin compartments, or vice versa (Figure 1F;
155 2nd and 3rd clusters). Also shown is an example of a transitional TAD boundary on chromosome 13, where there
156 is a shift from active to inactive chromatin marks separated by inversely-oriented CTCF binding sites (Figure
157 1G). Using the clusters shown in Figure 1F, each TAD was designated as active, weakly active, inactive, or
158 weakly inactive, based on the signal distribution around the boundary and the eigenvalue from Hi-C PCA
159 analysis along the length of the TAD (see Methods and listing of TADs in Supplementary File 1A). Striking
160 differences in gene expression were seen between active and inactive compartment TADs (median liver
161 expression 1.095 FPKM for 12,258 genes in active TADs vs. 0.003 FPKM for 4,643 genes in inactive TADs;
162 Figure 1H).

163
164 We sought to determine the tissue-specificity of the genes in active vs. inactive TADs. We used the
165 expression level of each gene across ENCODE tissues to calculate Tau scores, a robust metric for tissue
166 specificity [45, 46]. Tau values close to 1 are highly tissue specific, while lower values (< ~0.3) are widely
167 expressed and considered housekeeping genes (see Methods). A greater fraction of genes located in inactive
168 TADs are tissue-specific compared to genes in active TADs (Figure 1I). Overall, only 939 (20.2%) of all 4,643
169 genes in inactive TADs are expressed in liver (FPKM >1) vs. 6,290 (51.3%) of the 12,258 genes in active TADs.
170 Furthermore, genes whose TSS is close to a TAD boundary (i.e., TSS within 2% of the total TAD length in either
171 direction from the boundary) tend to be less tissue specific than the genomic average (Figure 1-figure
172 supplement 1G). Active transcription may be a key driver of dynamic cohesin movement in the nucleus [47], and
173 RNA polymerase II, *in vitro*, is capable of translocating cohesin rings along DNA [48]. Thus, the ubiquitous
174 expression of genes at TAD boundaries could be either a driver or an initiator of loop extrusion, although the
175 exact mechanism remains unknown.

176

177 **II. Identification of intra-TAD loops**

178 *Predicting intra-TAD loops from TAD-internal CTCF and cohesin binding sites.* While CAC sites that are tissue
179 ubiquitous, cohesin knockdown-resistant, or species-conserved show a clear 2 to 5-fold enrichment at TAD
180 boundaries (Figure 1), a large majority of such sites are TAD-internal and presumably do not contribute to TAD
181 formation. Overall, only 14.7% of liver CTCF binding sites are associated with TAD boundaries (Figure 2-figure
182 supplement 1A), consistent with other reports [1], and only 23% of the CTCF-bound regions that retain all four of

183 the above features are within 25 kb of a TAD boundary (Figure 2-figure supplement 1A). We considered two
184 possibilities: 1) TAD-internal CAC sites form intra-TAD loops that are too short to be detected in standard Hi-C
185 datasets; and 2) additional factors associated with TAD boundary CTCF sites differentiate them from other such
186 binding sites in the genome (see Discussion). To examine the first possibility, we modified an algorithm for
187 analysis of CTCF loops [31] and adapted it to predict subTAD-scale loops *in silico*, using CTCF and cohesin
188 peak strength and CTCF orientation as inputs (Figure 2A). Our approach builds on the finding that >90% of
189 CTCF-based loops are formed between inwardly-oriented CTCF sites [15]. Each mouse liver CAC site was
190 given a score that represents its peak strength and CTCF motif score, and an orientation was assigned based
191 on whether the non-palindromic CTCF motif was present on the (+) strand or the (-) strand, considering the
192 highest scoring CTCF motif at each CAC site. Scanning the genome, each (+) strand CAC peak was connected
193 to putative downstream (-) strand CAC sites. Low scoring CAC peaks were removed and the process was
194 iteratively repeated until the top 20,000 candidate loops remained. The set of loops was then filtered, as detailed
195 in Methods, to take into account cohesin scoring, and to ensure TSS overlap and <80% TAD overlap, to restrict
196 our definition of intra-TAD loops to TAD-internal CAC-mediated loops that contain at least one TSS. Applying
197 this algorithm to each of 4 matched pairs of liver ChIP datasets for CTCF and cohesin, we identified a set of
198 9,543 intra-TAD loops present in all four liver samples, with a median length of 151 kb. The set of intra-TAD
199 loops identified differs substantially from the generally shorter and much larger number of CTCF loops predicted
200 by the original algorithm [31] (Figure 2-figure supplement 1B; see Methods). Functionally, anchors of the shorter
201 loops predicted using the method of [31] show less insulation and weaker directional interactions than the intra-
202 TAD loop anchors identified in our study (Figure 2-figure supplement 1C,D; also see below). Moreover, 91% of
203 our predicted intra-TAD loops were wholly contained within a single TAD, versus only 67% for a random shuffled
204 control (Figure 2-figure supplement 1E). Figure 2B illustrates intra-TAD loop structures within TADs along a
205 segment of chromosome 2 and highlights their substantial overlap with “CTCF-CTCF” DNA loops identified by
206 ChIA-PET analysis of cohesin-mediated interactions in mouse embryonic stem cells (mESCs) using antibodies
207 to the cohesin subunit Smc1 [5]. The final set of 9,543 liver intra-TAD loops includes 1,632 intra-TAD loops
208 (17.1%) that share one CTCF loop anchor with a TAD boundary (i.e., an intra-TAD loop nested in a TAD with a
209 potential shared anchor). Consistent with these findings, high resolution Hi-C data in mouse CH12 cells reveals
210 the presence of single, multiple, and more complex nested intra-TAD loops that were predicted in mouse liver
211 (Figure 2-figure supplement 1F-H).

212

213 *Intra-TAD loop anchors share many properties of TAD anchors.* We examined the set of predicted intra-TAD
214 loops and their CAC site anchors to investigate their impact on genome structure and gene regulation. For these
215 analyses, we excluded from the intra-TAD anchor group the 1,632 intra-TAD anchors that are shared with TAD
216 anchors to ensure that the groups compared are mutually exclusive (Figure 2-figure supplement 2A). We first
217 sought to determine if the intra-TAD loop anchors show conserved CTCF binding across multiple ENCODE
218 tissues, as seen for TADs in Figure 1C. Figure 2C shows the tissue distribution of CTCF binding at CTCF
219 binding sites found at intra-TAD loop anchors in liver, where an X-axis value of 1 indicates the liver CTCF
220 binding site is occupied by CTCF in only one other tissue, and a value of 15 indicates binding occurs in all 15
221 mouse tissues where CTCF ChIP-seq data is available. Results show that a large majority of intra-TAD loop
222 anchors are bound by CTCF in at least 10 of the 15 mouse tissues examined. Indeed, CTCF binding at the
223 TAD-internal intra-TAD loop anchors is more deeply conserved across mouse tissues than that at TAD
224 boundaries. In contrast, CTCF sites not associated with cohesin binding (lone CTCF sites), and to a lesser
225 extent CAC sites not at intra-TAD or TAD loop anchors (other CAC sites) showed much greater tissue specificity
226 for CTCF binding (Figure 2C).

227

228 The enrichment of knockdown-resistant cohesin binding sites at TAD boundaries, seen in Figure 1B, may
229 explain the persistence of domain structures following CTCF or cohesin depletion [17, 41]. Further, we found
230 that 80% of TAD anchors and 90% of intra-TAD loop anchors are resistant to the loss of cohesin binding in
231 Rad21^{+/-} mice vs. only 52.8% for all CAC sites (Figure 2D). Moreover, a large fraction of TAD and intra-TAD
232 loop anchors, 70.6% and 77.6%, respectively, are comprised of “triple sites”, where cohesin and CTCF are co-
233 bound with Top2b, a potential component of the loop extrusion complex [20], vs. only 46.6% for the set of all
234 CTCF sites (Figure 2D). Top2b binding appears to be associated with cohesin binding rather than with CTCF
235 binding, as it is present at enhancer-like CNC sites much more frequently than at CTCF sites in the absence of
236 cohesin (Figure 2-figure supplement 2B).

237

238 TAD and intra-TAD loop anchors show greater sequence conservation within the core 18 bp CTCF motif than
239 other CTCF sites (Figure 2E). Analysis of sequences surrounding the CTCF core motif did not provide evidence
240 for loop anchor-specific motif usage or cofactor binding (Figure 2-figure supplement 2C,D). Downstream from
241 the core CTCF motif (within the loop interior) we observed a shoulder of high sequence conservation, as well as
242 additional complex motif usage, likely due to the multivalency of CTCF-DNA interaction, as described in [49].

243 TAD and intra-TAD loop anchors showed broader and more complex CTCF motif usage outside of the core
244 (Figure 2-figure supplement 2C,D); however, only a small minority of sites contained any specific motif in this
245 flanking region. It is less clear to what extent this broader motif usage is a general property of strongly-bound
246 CTCF regions or of the loop anchors themselves. In fact, we observed a consistent positioning of the cohesin
247 peak just downstream of the CTCF peak, independent of whether the CTCF site was predicted to participate in
248 loop formation or not (i.e. at both loop anchors and “Other CAC” sites; Figure 2F), in accordance with the loop
249 extrusion model and other observations [15, 18, 19, 50].

250

251 *Intra-TAD loops in other mouse tissues and in human cells.* Given the highly tissue-conserved binding of CTCF
252 at sites that we predicted to serve as intra-TAD loop anchors in liver (Figure 2C), we sought direct experimental
253 evidence for the presence of these loops in two other mouse cell types, ESCs and NPCs, where domain and
254 loop definitions have been established based on high-resolution Hi-C datasets [51]. Predicted loops were similar
255 in number and size across cell types, with substantial overlap between intra-TAD loops predicted in liver
256 compared to loops identified experimentally by Hi-C in mESCs or NPCs (62-63%; Figure 2-figure supplement
257 3A). We also observed 57-63% overlap of our predicted intra-TAD loops across the 3 cell types with CAC-
258 anchored loops identified in mESC using ChIA-PET for the cohesin subunit Smc1 [5, 32, 35] (Figure 2B, Figure
259 2-figure supplement 3A).

260 A majority of the predicted liver intra-TAD loops were also found in newer cohesin HiChIP datasets (ChIP for
261 Smc1a followed by Hi-C; data not shown) [52]. The substantial overlap between intra-TAD loops across mouse
262 cell types is very similar to the overlap between TADs, indicating a similar level of tissue ubiquity for intra-TAD
263 loops as for TADs (Figure 2-figure supplement 3B). Further, loops that were predicted in multiple cell types had
264 stronger interactions than tissue-specific or other loops, as determined by Smc1 ChIA-PET in mESC (Figure 2-
265 figure supplement 3C). A recent pre-print corroborates this result in human cells, where tissue-shared CTCF
266 loops were much stronger than tissue-specific CTCF loops [53].

267

268 Overall ~75% of intra-TAD loops that we identified in mouse liver are experimentally observed in at least one
269 other cell type (mESCs, CH12, or mouse NPCs). Further, 48.5% (4,632) of the computationally predicted liver
270 intra-TAD loops were also present in both mESCs, and NPCs (Supplementary File 1B). By comparison, 26.2%
271 and 21.5% of CTCF loops in HeLa and K562 cells, respectively, were tissue-specific, suggesting that the ~25%
272 of loops without support in at least one other tissue likely represent bona fide (albeit weaker) liver-specific intra-

273 TAD loops [53]. Examples of both shared and tissue-specific intra-TAD loops with supporting high resolution Hi-
274 C interactions are shown in Figure 2-figure supplement 3D-F.

275

276 We also predicted intra-TAD loops for two human cell lines, GM12878 and K562, and then compared our
277 predictions to loop domains and contact domains identified in these cells [15]. We predicted more loops in
278 human cells (~15,000 loops that contain a TSS) than in mouse (~10,000 loops with a TSS), owing in part to the
279 lack of a TAD overlap filter. We observed substantially more overlap of the predicted set of intra-TAD loops with
280 loop domains (40-54%) or with K562 cell CTCF ChIA-PET interactions (60-65%) than with contact domains (26-
281 35%; Figure 2-figure supplement 4A). Biologically, this difference makes sense, as ChIA-PET and our intra-TAD
282 prediction method both define a CTCF/CAC mediated interactome. Intra-TAD loops are more commonly shared
283 between K562 and GM12878 cells (67-73% shared) as compared to loop domains (46-66%) or contact domains
284 (37-57%) (Figure 2-figure supplement 4B); this can be compared to the much smaller difference in percentage
285 overlap (only 2-6%) between TADs and intra-TAD loops seen in mouse cells (Figure 2-figure supplement 3B).
286 As in mouse, intra-TAD loops that were predicted in both K562 and GM12878 cells interacted more strongly
287 than K562-specific or other loops, as determined by CTCF ChIA-PET in K562 cells (Figure 2-figure supplement
288 4C). Further, we found evidence for bona fide tissue-specific and shared intra-TAD loops (Figure 2-figure
289 supplement 4D-F), as well as an example of a tissue-specific enhancer-promoter interaction in GM12878 cells
290 within a larger tissue-specific intra-TAD loop (Figure 2-figure supplement 4E).

291

292 *Intra-TAD loops show strong, directional interactions and insulate chromatin marks.* TADs are proposed to
293 impact gene expression via two types of insulation: by insulation of chromatin interactions (also called enhancer
294 blocking) and by segregation of chromatin domains, primarily insulation of repressive histone mark spread. We
295 investigated whether intra-TAD loops demonstrate these dual insulating properties, canonically ascribed to
296 TADs.

297

298 CTCF sites were divided into three groups, based on whether they were predicted to anchor TAD loops or
299 intra-TAD loops, or were not predicted to interact (non-anchor CTCF sites) based on our algorithm. Figure 3A
300 shows the extent to which individual CTCF sites within each group show directional interactions towards the
301 loop interior based on Hi-C data, as determined using a chi-squared metric derived from the same directionality
302 index used to predict TADs genome-wide. This inward bias index quantifies the strength of interaction from a 25

303 kb bin immediately downstream from the CTCF motif, with a positive sign indicating a downstream (inward) bias
304 with regard to motif orientation, i.e., towards the center of a TAD/intra-TAD loops in the case of loop anchors.
305 TAD and intra-TAD anchors both show strong directional interactions compared to non-anchor CTCF sites, and
306 TAD anchors show stronger interactions than intra-TAD loop anchors. For the non-anchor CTCF sites, only
307 those sites containing a strong CTCF motif (FIMO score > 10) were considered, and the predicted directionality
308 was oriented relative to this motif (as there is no left versus right anchor distinction).

309

310 To test specific examples of insulation, we used available high resolution Hi-C for mESC to perform virtual
311 4C analysis [51] at select loop anchors (Figure 3B, Figure 3-figure supplement 1). This allowed us to visualize
312 the distribution of interactions originating from an intra-TAD loop anchor, and compare them to those originating
313 from an adjacent upstream region (outside of the loop). For virtual 4C viewpoints placed downstream of a *left*
314 intra-TAD loop anchor (i.e., within an intra-TAD loop; IN), interaction reads were shifted in favor of the
315 downstream direction, which comprised 58.8-79.9% of the Hi-C read pairs. This is comparable to the skew
316 observed for 4C-seq experiments performed at TAD anchors [13, 54] (Figure 3-figure supplement 1A,B). In
317 contrast, interactions were generally skewed in the upstream direction for viewpoints placed upstream of the
318 same set of intra-TAD loop anchors (Figure 3-figure supplement 1, OUT). These shifts in the distribution of
319 interactions further support the insulating nature of these intra-TAD anchors, and were seen both for intra-TADs
320 that are tissue-specific (Albumin, Sox2) and for those that are common (Hnf4a, Scd1) across the three tissues
321 we examined (liver, NPC and mESC cells).

322

323 To visualize features of these anchors, aggregate liver Hi-C profiles spanning 1 Mb around each anchor were
324 generated for each group of CTCF sites (Figure 3C). Each group was further subdivided into a left (upstream)
325 and a right (downstream) anchor based on its CTCF motif orientation. All non-anchor CTCF sites used for
326 comparison in Figures 3A and 3C were required to contain a CTCF motif to assign directionality, and CTCF
327 peaks were required to be present in a minimum of 2 individual biological replicates. By aggregating many sites,
328 we can visualize the overall interaction properties of each group of sites at high resolution (5 kb bins), revealing
329 features that are much harder to discern at an individual CTCF site (Figure 3C). TAD and intra-TAD loop
330 anchors both show strong enrichment of interactions towards the loop interior when compared to CTCF sites
331 that were not predicted to participate in loop formation (non-anchor CTCF sites). Furthermore, intra-TAD
332 anchors show less enrichment of long-range contacts compared to TAD anchors, likely because of their shorter

333 length compared to TADs. This may also explain the lower inward bias scores of intra-TAD loops seen in Figure
334 3A. Depletion of interactions that span across loop anchors (dark blue density above anchor points in Figure 3C)
335 was seen for both TAD and intra-TAD anchors, however, this local insulation was substantially greater for TAD
336 anchors. This may be due to the compounding impact of adjacent TAD loops, i.e., the end of one TAD is often
337 close to the beginning of an adjacent TAD loop anchor (median distance between TAD anchors = 33.5 kb,
338 Figure 3-figure supplement 2A). GO term analysis of genes whose TSS fall within these inter-TAD regions
339 revealed enrichment for housekeeping genes (ribosomal, nucleosome, and mitosis-related gene ontologies),
340 whereas neighboring genes found just within the adjacent TADs were enriched for distinct sets of GO terms
341 (Figure 3-figure supplement 2B; Supplementary File 3B, Figure 2-figure supplement 2C). The nearby but
342 oppositely oriented TAD anchors flanking the inter-TAD regions likely contribute to the more bidirectional
343 interaction pattern for TADs seen in Figure 3C. For instance, the left TAD anchor plot in Figure 3C shows a well-
344 defined pattern of interaction enrichment downstream from the anchor, but also a more diffuse enrichment
345 upstream contributed by upstream loop anchors located at various distances. In contrast, non-anchor CTCF
346 sites do not show strong directional interactions and only very weak distal contact depletion. CNC-bound
347 regions are predominantly found at enhancers [19] and do not show any discernable patterns of insulation or
348 focal and directional interactions (Figure 3C, *bottom*). Thus, the weak contact depletion spanning each of the
349 CTCF-containing groups shown in Figure 3C is likely real, and not an artifact of the background model or other
350 noise.

351

352 To determine if intra-TAD loops share with TADs the ability to block histone mark spread and establish
353 broad, insulated chromatin domains of activity and repression, we analyzed the distribution of two repressive
354 marks in relation to CTCF binding sites: H3K27me3 and H3K9me3. An insulation score based on Jansen
355 Shannon Divergence (JSD) [55] was calculated for ChIP signal distribution within a 20 kb window around each
356 CTCF peak. A low JSD value indicates less divergence from a string representing perfect insulation (i.e., high
357 signal on one side of peak, and low or no signal on the other). This scoring allows for a direct comparison of
358 different classes of CTCF sites, or other TF-bound sites, in terms of their insulation properties. In addition to
359 TAD and intra-TAD loop anchors, we examined three other sets of sites as controls: other CAC sites, sites
360 bound by CTCF alone, and CNC sites. Figure 3D shows the cumulative distribution of JSD scores for each set
361 of sites, where a leftward shift indicates greater insulation across the site for the chromatin mark examined. For
362 H3K27me3 signal distribution, TAD and intra-TAD loop anchors showed the greatest insulation, but were not

363 significantly different from each other (Kolmogorov-Smirnov (KS) test; $p=0.52$). The same general trend was
364 observed for H3K9me3 signal insulation; however, intra-TAD loop anchors actually showed greater insulation
365 than TAD anchors and all other groups (KS test; $p<0.001$). CNC sites consistently showed the least insulation of
366 both repressive histone marks, as expected. Sites where CTCF is bound alone showed a small but significant
367 increase in insulation compared to CNC sites, as did the CAC group. As a control, the distribution of IgG signal
368 (input signal) showed much less insulation overall and no significant differences between the various classes of
369 CTCF/cohesin binding sites (Figure 3-figure supplement 2C).

370

371 Figure 3E and 3F show heat map representations of H3K27me3 ChIP-seq signal around the top 500 active
372 and top 500 inactive intra-TAD loops, based on a ranked list of JSD insulation scores. These represent intra-
373 TAD loops that have significantly lower (or higher) H3K27me3 signal in the loop interior based on the combined
374 rank of JSD insulation scores for each anchor, respectively ($p < 0.05$, two-sided t-test). For example, Figure 3F
375 shows intra-TAD loops with lower H3K27me3 signal within the loop than in neighboring regions (and thus the
376 left anchor transitions from high to low signal, while the right anchor transitions from low to high signal). No such
377 pattern was seen for the IgG (control) ChIP-seq signals for these same regions, indicating this is not an artifact
378 of the sequence read mappability of these regions (Figure 3-figure supplement 2D).

379

380 **III. Impact of intra-TAD loops on *cis* regulatory elements in mouse liver**

381 *Classifying open chromatin regions and defining super-enhancers in mouse liver.* Many TADs, and also intra-
382 TAD loops, are structurally conserved across tissues, yet the activity of enhancers and promoters contained
383 within these looped structures is highly tissue-specific [5, 56]. Accordingly, it is important to understand the
384 ability of TADs and intra-TAD loops to insulate active enhancer interactions, i.e. enhancer-blocking activity. To
385 address this issue, we examined the distribution of CTCF and cohesin binding sites in relationship to promoters
386 and enhancers across the genome, as well as the impact of TADs and intra-TAD loops on their associated
387 genes and enhancers.

388

389 We previously identified ~70,000 mouse liver DNase hypersensitive sites (DHS), whose chromatin
390 accessibility is in part determined by TF binding and their flanking histone marks [57, 58]. To assign a function
391 for each DHS, we classified each DHS according to the ratio of two chromatin marks, H3K4me1 and H3K4me3,
392 which are respectively associated with enhancers and promoters [59]. The ~70,000 DHS were grouped into five

393 classes: promoter, weak promoter, enhancer, weak enhancer and insulator (Figure 4A, Figure 4-figure
394 supplement 1A; see Methods). Promoter-DHS were defined as DHS with a H3K4me3/H3K4me1 ratio ≥ 1.5 , and
395 enhancer-DHS by a H3K4me3/H3K4me1 ratio ≤ 0.67 (as first described in [60]). DHS with similar signals for
396 each mark (H3K4me3/H3K4me1 ratio between 0.67 and 1.5) were designated weak promoter-DHS, based on
397 their proximity to TSS and comparatively lower expression of neighboring genes (Figure 4-figure supplement
398 1B,C). DHS with low signal for both marks were classified as weak enhancers, or as insulators, in those cases
399 where they overlapped a CTCF peak with a comparatively strong ChIP-seq signal (Figure 4-figure supplement
400 1A). Using this simplified five DHS class model, we observed that enhancer-DHS bind cohesin largely in the
401 absence of CTCF, while promoter-DHS are bound by both CTCF and cohesin (Figure 4B). Additionally,
402 H3K27ac is enriched at promoter-DHS and enhancer-DHS but not at weak enhancer-DHS, which are less open
403 (lower DNase-seq signal) (Figure 4B) and more distal (Figure 4-figure supplement 1B). In contrast to enhancer-
404 DHS, insulator-DHS have a well-defined bimodal distribution of tissue-specific vs. tissue-ubiquitous sites based
405 on comparisons across 20 ENCODE tissues (Figure 4-figure supplement 1D). This supports the proposal that
406 insulators are a unique class of intergenic regions, and not simply enhancers bound by CTCF.

407

408 A subset of intra-TAD CAC loops are well characterized as insulators of tissue-specific genes with highly
409 active enhancer clusters, termed super-enhancers [5, 61]. To determine if some intra-TAD loops correspond to
410 these “super-enhancer domains” we first identified super-enhancers in mouse liver. We used 19 publicly
411 available mouse liver H3K27ac ChIP-seq datasets to score clusters of individual enhancers + weak enhancers
412 DHS identified above (Figure 4C). Super-enhancers were identified separately in male and female liver, as well
413 as in male liver at various circadian time points, to take into account these three key sources of natural variation
414 in gene expression in mouse liver [57, 62]. In total, we identified 503 core super-enhancers, i.e., super-
415 enhancers that show strong signal regardless of sex or time of day (Figure 4-figure supplement 2A). Core super-
416 enhancers represent 14.1% of all enhancer regions in the genome (6,680 of 47,372 constituent enhancers +
417 weak enhancers), and 2.8% of all enhancer clusters, defined as groups of enhancers within 12.5 kb of one
418 another (503 of all 17,964 enhancer clusters).

419

420 Genes, both protein coding and lncRNA genes, that neighbor super-enhancers are more highly expressed
421 and tissue-specific when compared to all genes, or to all genes neighboring typical enhancers (KS test p-value <
422 0.0001; Figure 4D). Consistent with this, only 6.8% of genes proximal to liver super-enhancers are targets of

423 super-enhancers in mESCs or pro-B cells (Figure 4E), whereas 55.2% of genes proximal to liver typical
424 enhancers are proximal to typical enhancers in the other two cell types. Super-enhancers showed much greater
425 accumulation of RNA polymerase 2, despite the lack of the promoter mark H3K4me3 (Figure 4F) and are
426 transcribed to yield eRNAs (Figure 4-figure supplement 2B). GO terms associated with genes targeted by either
427 typical enhancers or super-enhancers are enriched for liver functions (such as oxidoreductase activity),
428 however, super-enhancer target genes also show enrichment for transcription regulator activity and steroid
429 hormone receptor activity (Figure 4G). These data support the model that super-enhancers drive high
430 expression of select liver-specific genes, including transcriptional regulator genes (Supplementary File 2).

431

432 Strikingly, 72.2% of core super-enhancers (363/503) overlap either an intra-TAD loop or a TAD that contains
433 only a single active gene (defined as a gene expressed at FPKM ≥ 1 and with a promoter-DHS within 5 kb of the
434 TSS; Supplementary File 1B and Supplementary File 2B) (see, e.g., Figures 5C and 5D, below). By comparison,
435 only 43.6% (17,742/40,692) of typical enhancers are insulated in a similar manner (data not shown). We also
436 observed an enrichment of single-TSS intra-TAD loops ($n=3,142$) over a random shuffled set (Figure 4-figure
437 supplement 2C), which could represent tissue-specific genes that are regulated by super enhancers in liver or in
438 other tissues. Genes within these single-TSS intra-TAD loops (Supplementary File 1B), were enriched for
439 ontologies related to transcriptional regulation and phosphorylation (Figure 4-figure supplement 2D,
440 Supplementary File 3D,E). This is consistent with a model of intra-TAD loops as functionally inducible units of
441 gene expression, allowing selective transcription in a given tissue or in response to cell signaling events.

442

443 To determine the impact of intra-TAD loops on the expression of genes with neighboring super-enhancers,
444 we considered two possible gene targets for each super-enhancer, with the requirement that the TSS of each
445 gene target be within 25 kb of one of the individual enhancers that constitute the overall super-enhancer: one
446 gene target is located within the intra-TAD loop, and the other gene target crosses an intra-TAD loop anchor
447 (Figure 5A, scheme at *top*; genes inside (red) and genes outside (black) of the intra-TAD loop). We
448 hypothesized that the true gene target of the super-enhancer will be more highly expressed. Indeed, we found
449 that genes neighboring super-enhancers and found within the same intra-TAD loop are significantly more highly
450 expressed than the alternative potential gene targets, located outside of the intra-TAD loop (Figure 5A).
451 Similarly, when comparing the tissue specificity of genes within TADs and intra-TAD loops to a random shuffled
452 set of regions, we observed less variance in the Tau value (index of tissue specificity) for genes within

453 TAD/intra-TAD loops compared to the shuffled set (Figure 5B; KS test p-value < 0.0001). Thus, groups of genes
454 within intra-TAD loops are more uniformly tissue specific (as in the case of some super-enhancer-adjacent
455 genes) or tissue ubiquitous (as in the case of groups of housekeeping genes). Examples are shown in Figure
456 5C and 5D, which illustrate the impact of a TAD loop on the expression of *Cebpb* and the impact of an intra-TAD
457 loop on the expression of *Hnf4a*. The TSS of two other nearby genes, *Ptpn1* and *Ttpa1*, are closer in linear
458 distance to the adjacent super-enhancer than *Cebpb* and *Hnf4a*, respectively, however, any super-enhancer-
459 promoter interactions involving *Ptpn1* and *Ttpa1* would need to cross a TAD or intra-TAD loop boundary. In both
460 cases, the genes within the super-enhancer-containing TAD or intra-TAD loop are expressed at least 10-fold
461 higher than genes outside the loop. Therefore, based on the 3D-structure imposed by these TADs and intra-
462 TAD loops, one predicts that the super-enhancers are restricted from interacting with *Ptpn1* or *Ttpa1*, in
463 agreement with the comparatively low expression levels of those genes.

464

465 Given the ability intra-TAD loops to insulate repressive histone marks (Figure 3D-3F), we considered whether
466 intra-TAD loops enable high expression of genes within otherwise repressed genomic compartments. As seen in
467 Figure 1H, a minority of genes within inactive TADs are expressed (939 genes expressed at FPKM > 1). The
468 obesity-related gene stearoyl-CoA desaturase-1 (*Scd1*) is one such gene. Figure 5E shows a transitional TAD
469 boundary, with genes in the upstream TAD expressed and associated with low levels of H3K27me3 repressive
470 histone marks. Six genes are in the downstream TAD, but only one of these genes, *Scd1*, is expressed (Figure
471 5E, *bottom*). The high expression of *Scd1* (FPKM > 100) can be explained by its localization in an active intra-
472 TAD loop that is insulated from the repressive mark H3K27me3 compared to the rest of the TAD. This same
473 structural organization was seen for 291 of the 939 expressed genes found in inactive TADs (Figure 1H, above),
474 which are contained within intra-TAD loops. It is unclear what other mechanisms allow for selective expression
475 of the other 648 genes (Supplementary File 1E).

476

477 *4C-seq analysis of super-enhancer contacts at Alb promoter.* To test directly for the insulation of an intra-TAD
478 loop containing a super-enhancer, we performed 4C-seq analysis for the promoter of albumin (*Alb*), the most
479 highly expressed gene in adult mouse liver. 4C-seq is designed to identify all chromatin contacts originating
480 from a single genomic region (the promoter of *Alb* in this case), known as the 4C viewpoint. Using 4C-seq, we
481 captured many highly specific, reproducible interactions with the *Alb* promoter, a majority of which are localized
482 across an upstream ~50 kb region (Figure 6A, Figure 6-figure supplement 1A,B). 40% of the interactions are

483 localized within DHS that are constituent enhancers of the *Alb* region super-enhancer. Furthermore, >80% of
484 chromatin contacts are within the *Alb* super-enhancer, and >98% are contained within the *Alb* intra-TAD loop, in
485 both male and female liver. These interactions become more diffuse with increasing genomic distance from the
486 viewpoint, which may represent dynamic interactions, or alternatively, may reflect averaging across a
487 heterogeneous cell population. Comparing the combined interaction profiles between male and female livers, we
488 observed highly reproducible results, with 92.9% of male interactions also present in female liver (Figure 6-
489 figure supplement 1C).

490

491 Looking beyond local interactions, we observed 4C interaction frequencies that span several orders of
492 magnitude, going from local (intra-TAD) to *cis* (intra-chromosomal) and *trans* (genome-wide) interactions. Thus,
493 for the *Alb* promoter, the TAD 4C signal per TAD was up to >1000 RPKM for local interactions, >100 RPKM for
494 *cis* interactions within 3 TADs adjacent to the viewpoint, and ~10 RPKM beyond that (Figure 6B, 6C). *Trans*
495 interactions were almost exclusively < 10 RPKM (Figure 6B, 6C). Using a separate background model for far-*cis*
496 and *trans* 4C signals, we categorized TADs as either high, medium, low, or non-interacting (see Methods). As
497 *Alb* is the most highly expressed gene in liver and is proximal to a strong super-enhancer, we expected that
498 distal interacting regions would also be active genomic regions, as proposed in the transcription factory model of
499 nuclear compartmentalization [63]. Indeed, we found that >80% of the distal high interacting TADs (both far-*cis*
500 and *trans*) were active TADs, and >90% were either active or weakly active TADs. In contrast, only 16.6% of the
501 non-interacting TADs in *cis* and 25.9% of those in *trans* were active TADs (Figure 6D). Furthermore, genes in
502 the interacting regions are more highly expressed than genes in *Alb* 4C non-interacting regions, and the vast
503 majority are found in active TADs (Figure 1F), as determined by analysis of the Hi-C data alone (Figure 6-figure
504 supplement 1D,E).

505

506 **DISCUSSION**

507 We present, and then validate in multiple mouse and human cell models, a computational method that uses
508 2D (ChIP-seq) and 1D (DNA sequence) information to predict 3D-looped intra-TAD structures anchored by
509 cohesin and CTCF (CAC sites), and we provide evidence that the intra-TAD loops predicted underpin a general
510 mechanism to constrain the interactions between distal enhancers and specific gene targets. While select
511 instances of CAC-mediated loop insulation within TADs have been described [5, 64, 65], our work establishes
512 that this phenomenon is a more general feature of genomic organization and regulation than previously

513 appreciated. The intra-TADs described here are nested, CAC-anchored loops whose formation may be a result
514 of extrusion complex pausing within larger domains (i.e., TADs); these loops act to constrain the promoter
515 contacts available to a given distal enhancer, and correspondingly, the distal enhancer contacts available to a
516 given promoter [35]. We also provide evidence that the loop-forming CTCF sites, but not other CTCF sites, are
517 highly insular. This insulation is apparent from the blockage of repressive histone mark spread and by the
518 inhibition of chromatin contacts across intra-TAD loop and TAD boundaries. The impact of this insulation is
519 highlighted for super-enhancer regions, such as the super-enhancer upstream of *Alb*, where local insulation by
520 CAC-anchored intra-TAD loops both enables and constrains strong near-*cis* interactions, which facilitate the
521 high expression of *Alb* and presumably also other liver-expressed genes regulated by super-enhancers. Weaker
522 *trans* interactions with distal active regions were also observed, and are likely driven by a distinct mechanism,
523 such as aggregation of transcription factories or super-enhancers [26, 66].

524

525 Genomic interactions occur at three levels: (1) compartmentalization, where inactive regions localize to the
526 nuclear periphery and active chromatin compartments aggregate toward the center of the nucleus in *cis* or *trans*
527 in a largely cohesin-independent manner, as proposed in the transcription factory model [7, 15, 41, 66]; (2)
528 CAC-dependent looping, which generates many tissue-invariant scaffolds along the linear chromosome [1, 12,
529 35]; and (3) enhancer-promoter looping within CAC-loops, which may be directed by cohesin non-CTCF (CNC)
530 sites, mediator, or tissue-specific TFs [5, 18, 19]. If TADs define the broad domain within which a cohesin-driven
531 extrusion complex generally operates, then we have presented a simple method to identify loops within this
532 region that form as a result of dynamic loop extrusion movement and pausing at additional loop anchors. We
533 have used the term intra-TAD loops, also referred to as sub-TADs, to highlight their subdivision of TAD-internal
534 genomic space, although they are functionally similar to loop domains, isolated cliques, and insulated
535 neighborhoods, which tend to overlap or be contained within TADs [12, 15, 35]. Our computational method
536 cannot predict CTCF-independent loops, such as those mediated cohesin alone (enhancer-promoter loops),
537 although such loops are likely constrained by CAC driven intra-TADs, as was highlighted by our *Albumin* 4C-seq
538 results.

539

540 The method for CAC-mediated intra-TAD loop identification described here builds on the strong preference
541 for inward-facing CTCF motifs evident from high resolution Hi-C data [12, 15], and will be most useful for the
542 identification of intra-TAD CAC loops for the large number of cell lines and tissues that lack high resolution Hi-C

543 data. In these cases, intra-TAD loop domains cannot be identified because there is not sufficient local
544 enrichment to calculate a corner score with the arrowhead algorithm [15]. Further, while we used TAD
545 boundaries from standard resolution liver Hi-C data to filter out longer CAC loops, the frequent conservation of
546 TADs across both tissues and species [1, 3] broadens the applicability of our method to cell types, and perhaps
547 to new species, for which Hi-C data is not available and TAD boundaries have not been determined. Thus, even
548 in the absence of TAD coordinates, our method identifies TAD and intra-TAD looping events, which provide an
549 invaluable first approximation for understudied organisms. As we have tuned our parameters to identify loop
550 structures comparable in size and number to those found previously in mouse and human models, the
551 parameters used to filter an initial set of loop anchors may need to be adjusted for other model organisms.

552

553 We have used both CTCF and cohesin peak strength as the primary predictor of intra-TAD loop strength,
554 which is a reasonably good predictor of interactions [12, 31]. An alternative machine learning approach to
555 predicting CTCF/cohesin-mediated interactions, posted as an on-line preprint during review of our manuscript
556 [53], uses data from up to 77 genomic-derived features to predict CTCF-mediated loops in three human cell
557 lines. A key finding from this work was that cohesin strength was consistently the most predictive feature of
558 CTCF loops, followed by CTCF binding strength [53]. This method also captures loops that lack convergent
559 CTCF motif orientation, which represent as few as 8% of the total for loop domains [15], or as many as 20% in
560 the case of Insulated Neighborhoods [28]. However, the identification of this subset of loops comes at the
561 expense of requiring a minimum of 16 features for a given cell type, whereas our approach only requires three
562 features (CTCF motif, CTCF ChIP-seq, and cohesin ChIP-seq data). Importantly, the three features used by our
563 method represent 3 of the top 4 predictive features identified in [53].

564

565 The computational method presented here, which was validated in both mouse and human cell models,
566 provides a practical alternative to using high resolution Hi-C libraries for the identification of intra-TAD loops.
567 High resolution Hi-C requires extremely deep sequencing, which is costly, both in terms of computational and
568 experimental laboratory resources, and has only been achieved for a small number of cell lines [15, 34, 51].
569 Strategies to reduce the need for extreme deep sequencing to identify interactions at high resolution have been
570 proposed [36, 67, 68], and are beginning to make higher resolutions possible in more systems, however, the
571 sequencing depth and cost will likely remain out of reach for many labs. Antibody enrichment for select genomic
572 regions followed by chromosome conformation capture, as implemented in ChIA-PET, is an experimental

573 alternative to intra-TAD prediction. ChIA-PET and other 3C-based antibody enrichment methods select for
574 genomic regions that are highly bound by the protein(s) of interest (e.g., CTCF and cohesin), and can therefore
575 identify “many to many” interactions, instead of the “all to all” interactions identified by Hi-C; these methods are
576 therefore more practical than Hi-C, in terms of their sequencing depth requirements [33]. However, ChIA-PET
577 still requires ~10-fold more extensive deep sequencing per sample (~400 million reads) than is needed to obtain
578 the CTCF and cohesin ChIP-seq data utilized in our computational analysis to identify intra-TAD loops. Further,
579 as ChIA-PET uses antibody to select for genomic regions bound by CTCF and/or cohesin, it is difficult to
580 differentiate strength of antibody binding to the anchor proteins from strength of chromatin interaction between
581 the anchors. Of the various CTCF loops described in the literature, insulated neighborhoods are most similar to
582 the intra-TAD loops described here. Insulated neighborhoods are proposed to rectify the observation of smaller
583 and more abundant loops, evident in ChIA-PET datasets, with the established TAD model of large loops from
584 Hi-C experiments [5, 15, 38].

585

586 The TAD and intra-TAD loop anchors identified here together comprise 27% of all liver CTCF binding sites,
587 consistent with the 30% of murine ESC CTCF peaks that overlap insulated neighborhood anchor regions [35].
588 The precise mechanism that differentiates these CTCF sites, which anchor intra-TAD and TAD loops, from the
589 larger number of non-anchor CTCF binding sites present in any given tissue is unknown. Further, it is unclear
590 what role the typically weaker remaining ~70% of CTCF sites play in organizing the nucleus. Some of these
591 non-(intra-)TAD anchor CTCF sites may serve other, unrelated functions, given the ability of CTCF to interact
592 with other TFs, bind RNA, and regulate splicing mechanics [69-73]. Alternatively, some of these CTCF sites may
593 anchor loops present in only a minority of cells in the population analyzed, which would account for their overall
594 weaker signals. Early single cell Hi-C experiments suggested that TADs are present in virtually all individual
595 cells [74], however, more recent studies indicate cell-to-cell variability in TADs within a given cell population,
596 although the presence of distinct active and inactive genomic compartments is common across most individual
597 cells [40, 75]. Truly high-resolution elucidation of single cell intra-TAD structures may not be possible due to the
598 intrinsic limitation of two potential ligation events per fragment in any given cell.

599

600 We found that CAC sites are found at insulators and also at promoters, which we defined as DNase
601 hypersensitive sites (DHS) with high a histone-H3 K4me3/K4me1 ratio, whereas CNC sites are primarily at
602 enhancers and weak enhancers. Others find that promoters, when defined as the set of all TSS upstream

603 regions (including those not at a DHS), are bound by cohesin alone [18, 19]. Further, we found that CTCF-
604 bound open chromatin regions distal to promoters (insulator-DHS) show features that distinguish them from
605 other classes of open chromatin (promoter-DHS and enhancer-DHS), including the absence of enhancer marks
606 and their general conservation across tissues. Thus, these insulator-DHS regions are not simply enhancers with
607 CTCF bound. Supporting this, insulator regions consistently show less intrinsic enhancer activity than weak
608 enhancers in *in vivo* enhancer screens [76]. It is less clear what role CTCF binding in the absence of cohesin
609 plays in the nucleus, as we found such sites lack insulating activity and also lack strong directional interactions.
610 As CTCF binding is always intrinsically directional, due to its non-palindromic motif, the absence of directional
611 interactions from CTCF-non-cohesin sites suggests that the directionality of interactions with CTCF sites at TAD
612 and intra-TAD loop anchors is conferred by other factors associated with the extrusion complex, such as
613 cohesin [12, 14] or Top2b [20]. However, our findings suggest that the interactions of Top2b involve binding to
614 cohesin, and not CTCF, as indicated by the high frequency of CNC sites bound by Top2b vs. very low frequency
615 of Top2b binding at CTCF-non-cohesin sites (Figure 2-figure supplement 2B). Furthermore, binding by Top2b
616 does not distinguish TAD anchors from intra-TAD loop anchors. Indeed, by all metrics tested, we found no TF or
617 motif that differentiates TAD anchors from intra-TAD loop anchors, although the existence of some unknown
618 differentiating factor cannot be ruled out. Cohesin can stabilize large protein complexes comprised of up to 10
619 distinct TFs at enhancers [19], and could thus facilitate the binding of other unknown proteins to the loop
620 extrusion complex.

621

622 Cohesin is continuously recycled throughout the genome by loading and release factors [47], and so it is
623 unclear how insulator activity is effectively maintained at TAD and intra-TAD loop anchors in such a dynamic
624 environment. We found that CNC sites, which are primarily found at enhancers, consistently show the least
625 insulation of repressive histone marks, just as they show the least insulation of chromatin contacts. This
626 provides further evidence that TAD and intra-TAD loop anchors are functionally unique sites, and are not a
627 moonlighting feature of CTCF bound to enhancer regions. Furthermore, while enhancers are strongly enriched
628 for genetic non-coding variants, genetic variations at loop anchors are rare [35]. Mutations that occur at loop
629 anchors can result in dramatic phenotypes like polydactyly or tumorigenesis [77] and often occur in cancer [28].
630 Disruption of specific, individual CAC-mediated loop anchors using genomic editing tools results in aberrant
631 chromatin contacts and misregulation of neighboring genes in a largely predictable manner, although some
632 redundancy may occur when multiple nearby anchors are present [5, 35, 64].

633

634 The computational method for intra-TAD loop discovery, described here, is a substantial improvement over
635 prior implementations of computational loop prediction [12, 31]. The loops we identified were longer and fewer in
636 number (~9,500 vs ~60,000), showed much stronger insulation of chromatin interactions and greater insulation
637 of repressive histone marks, and displayed considerably greater overlap with cohesin-mediated loops identified
638 by ChIA-PET using antibody to the cohesin subunit Smc1 (Figure 2-figure supplement 1). Key features of our
639 computational method include the consideration of both CTCF and cohesin binding strength, as noted above, as
640 well as TAD structure and consistency across biological replicates. Our use of both CTCF and cohesin binding
641 strength in predicting intra-TAD loops is supported by a recent study of CTCF sites nearby the mouse α -globin
642 gene cluster, where the presence of CTCF alone was not sufficient to predict DNA loop interactions, and where
643 insulation by individual CAC sites ranged widely – from none to moderate to very strong insulation – in direct
644 proportion to the strength of CTCF binding, as revealed by deletion of individual CTCF sites [65]. Furthermore,
645 we developed a simple extension of our method that predicts TAD anchors when given a set of TAD boundaries
646 (Supplementary File 1C), and thereby overcome the limitation in identifying TAD anchors from low resolution,
647 standard sequencing depth Hi-C datasets. We were thus able to identify well-defined inter-TAD regions, which
648 we found are enriched for unique gene ontologies, notably, housekeeping genes with ribosomal, nucleosome,
649 and mitosis-related functions. A further extension of our findings would be the explicit use of the intra-TAD and
650 refined-TAD loop coordinates defined here to improve gene target assignments for distal regulatory elements,
651 based on the insulating capacity of these CAC-anchored looped domains. Such an approach may be beneficial
652 for the many model systems where distal enhancer activity is the clear driver of tissue specificity or a given
653 disease state [35]. The ability to identify intra-TAD loops based solely on CTCF motif orientation and CTCF and
654 cohesin ChIP-seq binding data, and then use these loops to improve gene target assignments for distal
655 regulatory elements is likely to constitute a substantial improvement over “nearest gene” and other, more
656 nuanced target assignment algorithms, such as GREAT [78].

657

658 In conclusion, our studies reveal that while TAD structures are readily apparent in routine Hi-C experiments,
659 their structural organization and functional impact on the genome is not unique. Structurally, the 9,543 TAD-
660 internal sub-loops that we identified for mouse liver have strong cohesin-and-CTCF-bound anchors and appear
661 to be formed by the same loop extrusion mechanism responsible for TAD formation. Functionally, we
662 hypothesize that these intra-TAD loops contribute to nuclear architecture as intra-TAD scaffolds that further

663 constrain enhancer-promoter interactions. We further show that these intra-TAD loops maintain key properties
664 of TADs, most notably insulation of chromatin interactions and insulation of repressive histone mark spreading.
665 The insulation provided by intra-TAD loops may enable high expression of super-enhancer target genes, as
666 illustrated for *Alb* in mouse liver, as well as high expression of individual genes within otherwise inactive TADs,
667 as exemplified by *Scd1* and the many other single gene intra-TAD loops that we identified. Given the increasing
668 interest in interactions of genes with distal enhancers and other intergenic sequences, the rapid and cost-
669 effective method described here for identification of intra-TAD structures that constrain long-range chromatin
670 interactions may prove invaluable in many areas of genomic research.

671

672

673 **MATERIAL AND METHODS**

674 **Animals and processing of liver.** Adult male and female CD-1 mice (ICR strain; RRID:MGI:5659424) were
675 purchased from Charles River Laboratories (Wilmington, MA) and were housed in the Boston University
676 Laboratory Animal Care Facility. Animals were treated using protocols specifically reviewed for ethics and
677 approved by Boston University's Institutional Animal Care and Use Committee (IACUC; protocol 16-003). Livers
678 were collected from 8-week-old mice euthanized by cervical dislocation and rinsed in cold PBS. Livers were
679 homogenized in a Potter-Elvehjem homogenizer using high sucrose homogenization buffer (10 mM HEPES (pH
680 7.5), 25 mM KCl, 1 mM EDTA, 2 M sucrose, 10% glycerol, 0.05 mM DTT, 1 mM PMSF, 0.15 mM spermine,
681 0.2% (v/v) spermidine, 1 mM Na orthovanadate, 10 mM NaF, and Roche Complete Protease Inhibitor Cocktail)
682 to prevent aggregation of nuclei and preserve chromatin structure. The resulting slurry was transferred on top of
683 a 3 ml cushion of homogenization buffer followed by ultracentrifugation at 25,000 RPM for 30 min at 4°C in an
684 SW41 Ti rotor to pellet the nuclei and remove cellular debris. The supernatant was carefully decanted to remove
685 liquid, and residual solid debris was removed from the tube walls using a sterile spatula and a dampened
686 Kimwipe. Nuclei were resuspended in 1 ml of crosslinking buffer (10 mM HEPES buffer (pH 7.6), 25 mM KCl,
687 0.15 mM 2-mercaptoethanol, 0.34 M sucrose, 2 mM MgCl₂) and transferred to a 1.5 ml Eppendorf tube. To
688 ensure consistent crosslinking, tubes were incubated for 3 min at 30°C prior to the addition of formaldehyde to a
689 final concentration of 0.8% (v/v). Samples were incubated in a 30°C water bath for 9 min with periodic mixing.
690 Crosslinking was halted by the addition of 110 µl of 1 M glycine, followed by a 5 min incubation at room
691 temperature. The crosslinked material was layered on top of 3 ml of high sucrose homogenization buffer and
692 then centrifuged as above. The crosslinked nuclear pellet was resuspended at 4°C in 1 ml of 1X

693 Radioimmunoprecipitation assay (RIPA) buffer (50 mM Tris-HCl, pH 8.0, 150 mM NaCl, 1% IPEGAL, 0.5%
694 deoxycholic acid) containing 0.5% SDS and protease inhibitors until homogenous by gentle pipetting.

695

696 **Sonication.** Crosslinked nuclei in RIPA buffer containing 0.5% SDS were transferred to 15 ml polystyrene tubes
697 (BD Falcon # 352095) for sonication using a Bioruptor Twin instrument (UCD-400) according to the
698 manufacturer's instructions. Briefly, samples were sonicated at 4°C for 30 sec ON and 30 sec OFF at high
699 intensity for a total of 75 cycles. Sonicated material was transferred to 1.5 ml Eppendorf tubes, and large debris
700 was cleared by centrifugation at 18,000 x g for 10 min at 4°C. The bulk of this material was snap frozen in liquid
701 nitrogen and stored at -80°C for immunoprecipitation, except that a small aliquot (15 µl) was removed to quantify
702 material and ensure quality by gel electrophoresis, as follows. Aliquots from each sample were adjusted to 0.2
703 M NaCl, final concentration, then incubated for 6 h at 65°C. After a three-fold dilution in nuclease-free water, 5
704 µg of RNase A (Novagen: #70856) was added and samples were incubated for 30 min at 37°C. Samples were
705 then incubated for 2 h at 56°C with 20 µg of Proteinase K (Bioline; BIO-37084). This material was then
706 quantified in a dilution series using PicoGreen assay (Quanti-iT dsDNA Assay Kit, broad range, Invitrogen) and
707 analyzed on a 1% agarose gel to ensure the bulk of material was within 100-400 bp.

708

709 **Chromatin immunoprecipitation (ChIP).** Immunoprecipitation of crosslinked, sonicated mouse liver chromatin
710 and downstream steps were as described previously [57]. Protein A Dynabeads (30 µl; Invitrogen: 1002D) were
711 incubated in blocking solution (0.5% bovine serum albumin in PBS) with 5 µl of antibody to CTCF (Millipore #07-
712 729; RRID:AB_441965) or to the cohesin subunit Rad21 (Abcam #992; RRID:AB_2176601) for 3 h at 4°C. As a
713 control, 1 µl of non-specific rabbit IgG was used (Santa Cruz: sc-2027). Bead immune-complexes were washed
714 with blocking solution, followed by overnight incubation with 70 µg of liver chromatin. After washing with 1X
715 RIPA (containing 0.1% SDS) and reverse crosslinking as described above, DNA was purified using the
716 QIAquick Gel Extraction Kit (Qiagen #28706) and quantified on a Qubit instrument with high sensitivity detection
717 (Invitrogen DNA HS# Q32854), with ChIP yields ranging from 1-25 ng. Samples were validated by qPCR using
718 primers shown in Supplementary File 4B.

719

720 **Library preparation and sequencing.** ChIP libraries were prepared for sequencing using NEBNext Ultra II
721 DNA Library Prep Kit for Illumina according to the manufacturer's instructions (NEB, cat. #E7645). All samples

722 were subjected to double-sided SPRI size selection prior to PCR amplification (Agencourt AMPure XP;
723 Beckman Coulter: A63882). Samples were assigned unique barcodes for multiplexing, and subjected to 8
724 rounds of PCR amplification with barcoded primers (NEB, cat. #E7335). Samples were sequenced either on an
725 Illumina Hi-Seq 4000 instrument at the Duke Sequencing Core or an Illumina Hi-Seq 2000 instrument at the MIT
726 BioMicroCenter, giving 50 bp single end reads at a depth of ~11-19 million reads per sample. A total of four
727 CTCF and three Rad21 (cohesin) ChIP-seq samples were analyzed, representing four male mouse livers. The
728 fourth liver CTCF sequencing sample, sample G133_M9, did not have a matching cohesin ChIP-seq dataset
729 from the same liver sample, and was therefore matched to a merged sample comprised of all three cohesin
730 ChIP-seq replicates (merged at the fastq file level, with processing described below). Raw and processed
731 sequencing data are available at GEO accession number GSE102997.

732

733 **General ChIP-seq analysis pipeline.** Sequence reads were demultiplexed and mapped to the mouse genome
734 (build mm9) using Bowtie2 (version 2.2.9), allowing only uniquely mapped reads. Peaks of sequencing reads
735 were identified using MACS2 (version 2.1.1) as regions of high signal over background. Peaks were filtered to
736 remove blacklisted genomic regions (www.sites.google.com/site/anshulkundaje/projects/blacklists). Genomic
737 regions called as peaks that contain only PCR duplicated reads, defined as >5 identical sequence reads that do
738 not overlap any other reads, were also removed. All BigWig tracks for visualization in a genome browser were
739 normalized for sequencing depth, expressed as reads per million mapped reads (RPM) using Deeptools
740 (version 2.3.3). Unless otherwise indicated, all pairwise comparisons presented in the Figures were performed
741 using a Kolmogorov–Smirnov test, where **** indicates $p \leq 0.0001$, *** $p \leq 0.001$, ** $p \leq 0.01$, and * $p \leq 0.05$.

742

743 **Motif analysis.** Motifs within CTCF peak regions were identified using MEME Suite (version 4.10.0; FIMO and
744 MEME-ChIP options). FIMO was used to assign CTCF motif orientation and motif scores for CAC sites and to
745 discover individual motif occurrences. *De novo* motif discovery was carried out using MEME-ChIP using default
746 parameters (Figure 2-figure supplement 2C,D). Similar results were obtained using Homer (version 4.8).
747 Alternative CTCF motifs were downloaded from CTCFBSDB 2.0
748 (http://insulatordb.uthsc.edu/download/CTCFBSDB_PWM.mat), however, these did not substantially change
749 any results performed using the core JASPAR motif (MA0139.1). These motifs were explicitly used in Figure 2-
750 figure supplement 2D, where no difference between intra-TAD loop anchor and TAD anchor motif usage was
751 observed.

752

753 **4C-seq protocol.** Four male and four female mouse livers were processed for *Albumin*-anchored 4C-seq
754 analysis using published protocols, with some changes for primary tissue [79]. To adapt the protocol for liver,
755 care was taken to rapidly isolate single liver cells or nuclei suspensions prior to crosslinking. Specifically, two
756 approaches to crosslinking were taken and both gave similar results. One male mouse liver and one female
757 mouse liver sample were processed through the crosslinking step as described for the ChIP protocol, above,
758 prior to quantification of nuclei. The other liver samples (3 males and 3 females) were crosslinked as follows.
759 Half of a liver (~0.5 g) was dissected from each mouse, the gall bladder was removed, and the liver was rinsed
760 with PBS. The liver was then minced and rapidly processed with 10 strokes in a Dounce homogenizer in PBS
761 containing protease inhibitors (PBS-PI; 1X Roche Complete Protease Inhibitor Cocktail; Roche #11697498001).
762 The resulting slurry was passed through a 40-micron cell strainer (Corning #431750), then pelleted and rinsed
763 with PBS-PI (centrifugation at 1,300 RPM for 5 min at 4°C). Following an additional spin, the cell pellet was
764 resuspended well in 9 ml of PBS-PI at room temperature. 270 µl of 37% formaldehyde was added to give a final
765 concentration of 1%, and crosslinking was carried out for 10 min with nutation at room temperature. The
766 remaining formaldehyde was quenched with 1.25 ml of 1 M glycine. Crosslinked cells were pelleted and rinsed
767 with PBS twice (as above) prior to lysis. The supernatant was removed following the second wash, and cell
768 pellets were resuspended in 8 ml of lysis buffer (50 mM Tris (pH 7.5), 150 mM NaCl, 5 mM EDTA, 0.5% NP-40,
769 1% TX-100, and 1X Complete Protease Inhibitor Cocktail) and incubated on ice for 40 min with occasional
770 mixing. Lysed cells were spun down at 2,000 RPM for 5 min at 4°C then washed twice with PBS-PI, as above.
771 Nuclei were pelleted, quantified using an Invitrogen Countess instrument, and snap frozen in 10 million nuclei
772 aliquots. Primary digestion of 10 million nuclei with 50,000 U of DpnII (NEB: #R0543) was performed overnight
773 at 37°C in 450 µl of NEBuffer 3 (NEB: #B7003S; 100 mM NaCl, 50 mM Tris-HCl, 10 mM MgCl₂, 1 mM DTT, pH
774 7.9) with agitation at 900 RPM. After confirming digestion by agarose gel electrophoresis, DpnII was inactivated
775 with SDS (2%, final concentration) and the samples then diluted 5-fold in 1X ligation buffer (Enzymatics
776 #B6030). 200 U of T4 DNA ligase was added and primary ligation was carried out overnight at 16°C
777 (Enzymatics #L6030). Ligation was confirmed by analysis of a small aliquot on an agarose gel, and reverse
778 crosslinking was conducted by overnight incubation with 600 µg proteinase K at 65°C. After RNase A digestion
779 and phenol/chloroform cleanup, samples underwent secondary digestion with 50 U of Csp6I (Fermentas
780 #ER0211) overnight at 37°C in 500 µl of 1X Buffer B (Fermentas: #BB5; 10 mM Tris-HCl (pH 7.5), 10 mM

781 MgCl₂, 0.1 mg/ml BSA). Csp6I was then heat inactivated for 30 min at 65°C. Samples were diluted 10-fold and
782 secondary ligation was carried out as above, overnight at 16°C. The final PCR template was purified by
783 phenol/chloroform clean up, followed by QiaPrep 2.0 column cleanup (Qiagen #27115). PCR reactions were
784 performed using inversely-oriented 4C primers specific to the *Alb* promoter (sequences shown in bold, below)
785 with dangling 5' half adapter sequences (Reading primer:
786 ACACTCTTTCCCTACACGACGCTCTTCCGATCT**GGTAAGTATGGTTAATGATC**; Non-reading primer:
787 GACTGGAGTTCAGACGTGTGCTCTTCCGATCT**CTCTTTGTCTCCCATTTGAG**). This design has two
788 advantages: 1) the addition of barcodes in a secondary reaction allows a primer to be reused across samples;
789 and 2) it avoids barcoding at the start of 4C read, which would reduce the mappable read length available for
790 downstream analyses. 4C templates were amplified using Platinum Taq DNA polymerase (Invitrogen
791 #10966026) under the following conditions: 94°C for 2 min, 25 cycles at (94°C 30 s, 55°C 30 s, 72°C 3 min),
792 then 4°C hold. A total of eight liver samples were analyzed (four males, M1-M4; and four females, F1-F4). For
793 liver samples M3, M4, F3 and F4, eight identical PCR reactions processed in parallel were pooled to limit the
794 impact of PCR domination in any single reaction. For liver samples M1, M2, F1 and F2, two PCR reactions for
795 each sample were sequenced separately then pooled at the fastq file level for downstream analyses. We
796 observed that the 8 PCR pool reactions gave a more reproducible profile than the single PCR reactions. After
797 pooling, 4C samples were purified using AMPure XP beads (Beckman Coulter: #A63882) at a 1.5:1 ratio of
798 beads to sample, washed with 75% ethanol, dried, and resuspended in 0.1X TE buffer to elute the DNA. 4C-seq
799 samples were multiplexed and PCR amplified using standard NEB barcoded primers (NEB #E7335), as was
800 done for ChIP-seq library preparations, but for five additional PCR cycles rather than the 8 cycles used for ChIP
801 libraries (total of 30 cycles of PCR per sample: 25 cycles with viewpoint-specific primers followed by 5 cycles
802 with viewpoint-generic barcoded NEB primers). 4C libraries were sequenced on an Illumina Hi-Seq 2500
803 instrument at the New York Genome Center giving 125 bp long paired end reads. Samples were each
804 sequenced to a depth of ~1-5 million reads. Raw and processed sequencing data are available at GEO under
805 accession number: GSE102998.

806

807 **4C-seq data analysis.** All *Alb* viewpoint 4C-seq reads were filtered to ensure a match for the bait primers used,
808 then trimmed using a custom script to remove the first 20 nt of each read (Source Code File 1). Reads were
809 then mapped to the mouse mm9 reference genome using the Burrows-Wheeler aligner (bwa-mem) allowing for
810 up to two mismatches. The package r3Cseq [80] was used to analyze the distribution of signal in *cis*, both near

811 the bait and along chromosome 5. Reads were counted per restriction fragment to obtain the highest possible
812 resolution. Data shown in the main text figures are for the intersection of four male and four female mouse
813 livers, merged according to sex using the intersection option, meaning that the 4C interactions shown are those
814 present in all four samples for a given sex. This produces both a normalized read depth signal (reads per million
815 per restriction fragment) and an associated p-value for the interaction, taking into account distance from
816 viewpoint and reproducibility across replicates (Figure 6-figure supplement 1A,C). A comprehensive view of all
817 replicates is presented in Figure 6-figure supplement 1A. For all pairwise comparisons, a correlation between
818 samples was calculated genome-wide using the UCSC utility bigWigCorrelate with default settings. To account
819 for high signal immediately surrounding the viewpoint, this analysis was only conducted for regions >10 kb from
820 the viewpoint fragment. To analyze more distal *cis* interactions, we first calculated the normalized 4C signal
821 observed per TAD along chromosome 5, in units of RPKM per TAD. We observed a robust logarithmic decay of
822 signal with increasing distance from the viewpoint TAD (Figure 6C; $R^2=0.719$). Interacting TADs were defined
823 according to observed over expected 4C signal relative to this background model. Interacting TADs were
824 designated as follows: high, defined as regions with >2-fold enrichment over this background model (observed/
825 expected); medium, defined as 1.5 to 2-fold enrichment; and low, between 1.5-fold enrichment and 1.5-fold
826 depletion. Non-interacting TADs showed >1.5-fold depletion of signal. For the *Alb* viewpoint, we identified 17
827 high, 20 medium, 128 low, and 30 non interacting *cis* TADs along chromosome 5. We required a more
828 comprehensive background model to analyze interactions in *trans*. The tool 4Cker was used for its adaptive
829 windowing and Hidden Markov model approach [78]. The count tables from r3C-seq were merged by sex and
830 imported, then *trans* analysis was conducted with the recommended parameters ($k=20$). The default output
831 identifies three classes of regions: interacting, low-interacting, and non-interacting. For our analysis, the
832 interacting group was divided into two equal-number groups: high-interacting and medium-interacting, based on
833 4C interaction strength in the male liver samples. *Trans* interacting regions tend to be large (median size of 1.8
834 Mb), therefore *trans* interacting TADs were defined as TADs wholly contained within these interacting regions.
835 This corresponds to a total of 659 (high), 618 (medium), 969 (low), and 77 (non) interacting *trans* TADs
836 genome-wide.

837

838 **CAC sites and scores.** Cohesin-and-CTCF (CAC) sites were defined as CTCF peaks that were present in at
839 least 2 of 4 individual mouse liver samples and that overlapped with a cohesin peak in any liver sample. CAC
840 sites were scanned for a CTCF motif (JASPAR motif MA0139.1) within the CTCF peak coordinates using the

841 FIMO tool in the MEME Suite (version 4.10.0). For a given CAC site, the highest scoring motif occurrence for
842 the canonical core CTCF motif (MA0139.1) was considered. A (+) strand orientation indicates that the motif is
843 found on the (+) genomic strand (Watson strand). Each CAC site was represented by two different scores: a
844 CTCF score = $p * (m/10)$, where p is the CTCF peak strength (MACS2 score) and m is the CTCF motif score, as
845 determined by FIMO; and a cohesin score = $p * (m/10)$, where p is the cohesin (Rad21) peak strength (MACS2
846 score) and m is the CTCF motif score, as determined by FIMO.

847

848 **intra-TAD loop prediction method.** We modified a published algorithm for CTCF-mediated loop prediction [31]
849 to predict intra-TAD loop structures. Key modifications to the algorithm include the following: incorporation of
850 cohesin ChIP-seq data in scoring, based on the finding that CTCF signal in the absence of cohesin is not
851 sufficient to predict chromatin interactions [65]; consideration of TAD structure, TSS overlap, and consistency
852 across biological replicates when filtering to obtain the final set of predicted loops; and a final target set of
853 approximately 10,000 intra-TAD loops, based on experimental results from high resolution Hi-C analyses [15].
854 First, CAC sites were identified from mouse liver ChIP-seq data for Rad21 and CTCF, obtained as described
855 above. Next, each chromosome was scanned for putative intra-TAD loops, formed between a (+) anchor
856 [upstream anchor, i.e., CAC site with a CTCF motif (JASPAR motif MA0139.1) on the (+) strand] at the start of a
857 loop and a (-) anchor [downstream anchor, i.e., CAC site with a CTCF motif on the (-) strand] at the end of a
858 loop, as described for prediction of intra-chromosomal CTCF loops in [31]. Scanning was initiated from the
859 beginning of each chromosome, and a list of putative (+) anchors was generated. Next: (1) if the next CAC site
860 encountered was a (-) anchor, the pair of (+) and (-) anchors was recorded as a putative intra-TAD loop. The (+)
861 anchor was paired with all subsequent, downstream (-) anchors until another (+) anchor was encountered, at
862 which point the list of putative intra-TAD loops was closed, ending with the last (-) anchor. Alternatively, (2) if the
863 next CAC sites encountered were (+) anchors, then all such (+) anchors were retained as putative upstream
864 anchors, until the next (-) anchor was reached, and then all such (+) anchors were paired (i.e., assigned to
865 loops) with all of the subsequent, downstream (-) anchors until a new (+) anchor was encountered, as described
866 under (1), at which point the list of putative loops was closed, ending with the last (-) anchor. A new scan for
867 putative intra-TAD loops was then initiated in a linear fashion, starting from the next (+) anchor until all
868 chromosomes were scanned and a set of putative intra-TAD loops was obtained. Chromosome scanning for
869 putative intra-TAD loops was then repeated as described above after removing 10% of the CAC sites – those
870 with the lowest CTCF scores (defined above). Chromosome rescanning was repeated iteratively until the

871 number of putative intra-TAD loops decreased to as close to 20,000 as possible (removing the lowest scoring
872 loops if needed so that all replicates had exactly 20,000 loops prior to merging). A parallel series of iterative
873 scans was carried out, except that 10% of the CAC sites with the lowest cohesin scores (defined above) were
874 removed at each iteration, to generate a second set of ~20,000 putative intra-TAD loops. The intersection of the
875 two sets of 20,000 putative intra-TAD loops was then determined. The same iterative process of intra-TAD loop
876 prediction was carried out independently for each of the n=4 individual mouse livers, based on an analysis of
877 matched CTCF and cohesin (i.e., Rad21) ChIP-seq datasets for each liver. Thus, for each liver sample, a single
878 putative intra-TAD loop set was generated from the intersection of two sets of predicted CAC-based loops, one
879 using CTCF ChIP-seq strength, and the other using cohesin (Rad21) ChIP-seq strength, with both scores
880 modified by the CTCF motif score. The overlap of these two putative intra-TAD loop sets was approximately
881 80%, and ranged from 15,999-16,892 loops for a given liver sample. Additional filters were then applied to
882 remove intra-TADs that did not contain either a protein-coding TSS or a liver-expressed multi-exonic lncRNA
883 TSS (as defined in [81]), as we were primarily interested in the impact of intra-TADs on gene expression and
884 regulation. Putative intra-TAD loops that overlapped >80% of a TAD, or whose (+) and (-) anchors are both TAD
885 anchors (defined below) were also excluded, as these loops could not be distinguished from TAD loops. These
886 two filters further reduced the putative intra-TAD loop sets to approximately 63% of the original 20,000 (ranging
887 from 12,395 to 12,962 loops across the four liver samples). A single merged ChIP-seq dataset (merged at the
888 fastq file level, separately for CTCF and for Rad21 datasets) was treated as a fifth dataset. It was run through
889 the full pipeline, above, and then sequentially intersected with the set of putative intra-TAD loops predicted for
890 each individual liver to obtain a final set of 9,543 intra-TAD loops identified in all 4 livers and also present in the
891 5th dataset (merged sample). Each intra-TAD loop was assigned an intra-TAD loop score equal to the geometric
892 mean of the (+) anchor's CAC site CTCF score and that of its (-) anchor. A second intra-TAD loop score, equal
893 to the geometric mean of the (+) anchor's CAC site cohesin score and that of its (-) anchor, was also assigned.
894 The CTCF and cohesin scores reported for each loop in the final intra-TAD loop lists (Supplementary File 1B)
895 are those obtained from the merged sample. Custom scripts for intra-TAD loop prediction are available in
896 Source Code File 1.

897

898 Loop predictions in two other mouse cell types (mESC and NPC) and in two human cell lines (GM12878 and
899 K562) was carried out as described above, with the following modifications during filtering. For the mouse cells,
900 ChIP-seq data from biological replicates (n=4 for mESC and n=3 for NPC) was obtained for CTCF and cohesin

901 (ChIP for the subunit Smc1). Further, TADs from the same cell type were used to filter based on TAD overlap
902 (using TAD boundaries from [51]). TSS overlap used the same definitions as above (RefSeq and multi-exonic
903 lncRNA TSS defined in mouse liver [81]). For human loop predictions, cohesin (ChIP for the subunit Rad21) and
904 CTCF ChIP-seq data were obtained from biological replicates for K562 cells and for GM12878 cells (n=5 for
905 both cell lines), and overlap with Refgene TSS (hg19) was used to filter the merged loops, as TADs were not
906 defined in [15]. Supplementary File 4A provides further details on data sources and accession numbers.

907

908 **RNA-seq Analysis.** Gene expression values for liver-expressed protein coding genes are $\log_2(\text{FPKM}+1)$ values
909 from [81]. Liver-expressed non-coding genes are expressed in FPKM based on the gene models and
910 expression values from [81]. To express the tissue specificity of a gene's expression across a panel of 21
911 mouse tissues (including liver), we used Tau, which was shown to be the most robust in a recent study [46].
912 Testis was excluded from this analysis because a large proportion of testis-expressed genes are highly tissue
913 specific. For each tissue, the maximum FPKM per gene between the two replicates was used. These FPKM
914 values were log transformed and a Tau value, ranging from 0 to 1, was calculated, where 1 represents high
915 tissue specificity: $\tau = \sum_{i=1}^n (1-y_i)^{n-1}$, where $y_i = x_i / \max_{1 \leq i \leq n}(x_i)$, n is the number of tissues, and x_i is the
916 expression of the gene in tissue i.

917

918 **General Hi-C Processing.** Hi-C data was processed using the HiC-Pro package (version 2.7.0) [82] for
919 mapping and read filtering, followed by Homer (version 4.8) for downstream analyses such as PCA analysis and
920 aggregate contact profiles. Biological replicates were merged to increase read depth. The default Homer
921 background model was used for all datasets, where the expected frequency of interactions takes into account
922 read depth between interacting bins and genomic distance. PCA was conducted using Homer with the
923 command "runHiCpca.pl -res 10000 -cpu 4 -genome mm9" to generate genome wide eigenvalues at 10 kb
924 resolution. The values changed marginally at 20, 40, or 50 kb, but the sign of the eigenvalue was unaffected,
925 i.e., there was no impact on whether a TAD was assigned as A compartment or B compartment.

926

927 **Peak distribution within TADs.** Published TAD coordinates in mouse liver [3] were converted from mouse
928 genome mm10 to mm9 using liftover with default parameters. Each TAD was then divided into 100 equal-sized
929 bins using the Bedtools command makewindows. Next, these bins were compared to the peak positions of
930 various publicly available ChIP-seq datasets using the Bedtools coverage command, and the number of peaks

931 per bin was counted. This resulted in a string of 100 values for each TAD, representing the number of ChIP-seq
932 peaks per bin, where the first value is the start of the TAD and the last value is the end of the TAD. Conducting
933 this analysis across all TADs yielded a matrix of 3,617 rows (one per TAD) x 100 columns (one per bin). To
934 generate the aggregate profiles shown in Figure 1A-1E, and in Figure 1-figure supplement 1B-E, the sum of
935 each column was taken and then normalized to account for differences in total peak count for the different
936 samples, factors, and chromatin marks analyzed. Normalization was conducted by taking the average of the
937 center five bins (bins 48-52) and dividing the bin sums by this normalizing factor. This allows the y axis to
938 represent bin enrichment relative to the center of the TAD, as shown.

939

940 **TAD activity and compartment assignment.** ‘TAD boundaries’ were defined at single nucleotide resolution as
941 the end of one TAD and the start of another (as defined in [3]), thus excluding the start of the first TAD in each
942 chromosome and the end of the last TAD. In contrast, all references to ‘TAD anchors’ refers to the CTCF sites
943 most likely to be anchoring TAD loops based on distance from the boundary and proper orientation (as
944 described in TAD Anchor Identification). Data sources for all ChIP-seq, GRO-seq, Hi-C, and other datasets are
945 described in Supplementary File 4. H3K9me3, H3K27me3, H2AK5ac, and H3K36me3 marks were processed
946 from the raw sequencing data (fastq files) through the standard ChIP-seq pipeline, above. H3K9me3 and
947 H3K27me3 mark data were expressed as $\log_2(\text{ChIP/IgG signal})$. Lamina-associated domain coordinates and
948 GRO-seq data were downloaded as pre-processed data. Heat maps were generated using Deeptools reference
949 point, with a bin size of 10 kb. TAD boundaries were grouped according to k-means clustering ($k=4$) using signal
950 within a 1 Mb window from 3 datasets: H3K9me3, H2AK5ac, and the eigenvalue of the Hi-C PCA analysis
951 (above). Based on these clusters, TADs were classified as active, weak active, weak inactive, or inactive, as
952 follows. If both the start and end boundary of a given TAD were classified as active, then the TAD was
953 designated active. Specifically, a TAD was considered “active” if the boundary at the start of a TAD fell into
954 clusters 1 or 2 (as marked in Figure 1F) and the boundary at the end of the same TAD fell into clusters 1 or 3.
955 The corresponding metric was applied to identify inactive TADs. If the two ends of a given TAD were not in
956 agreement, then the TAD was designated weakly active if the median Hi-C PC1 eigenvalue within the TAD was
957 positive, or weakly inactive if the median Hi-C PC1 eigenvalue was negative. Gene expression and tissue
958 specificity metrics represent expression or Tau values of genes whose TSS overlap active or inactive TADs.
959

960 **Additional Hi-C analysis.** Contact profiles around TAD, intra-TAD loop, and non-loop-anchor CTCF sites were
961 generated using Homer (v4.9) using the command analyzeHiC and the options “-size 500000 -hist 5000” to
962 generate interaction profiles for 1 Mb windows around CTCF sites with 5 kb resolution. TAD and intra-TAD loop
963 anchors were split into left and right anchors if they were at the start or the end of the predicted loop. Non-
964 anchor CTCF sites were defined as any remaining CTCF sites from the merged sample that also contained a
965 CTCF motif. Left and right groupings were determined based on the orientation of the strongest CTCF motif
966 within the non-anchor peak regions. The inward bias index (IBI) was modified from the more genome-wide
967 directionality index (DI) described in [1]. Both DI and IBI use a chi-squared statistic to determine the extent to
968 which Hi-C reads from a given region have a strong upstream or strong downstream bias. While DI is genome
969 wide, IBI focuses on the directionality of *cis* interactions (within 2 Mb) from a 25 kb window immediately
970 downstream of a CTCF peak relative to the motif orientation. A large positive value indicates a strong interaction
971 bias towards the loop center, as the motif orientation would predict. Values close to zero indicate a roughly
972 equal distribution of interactions upstream and downstream. By orienting the sign of the IBI value relative to the
973 CTCF motif directionality, we were able to group left and right loop anchors together.

974 Virtual 4C plots (Figure 3B and Figure 3-figure supplement 1) and Hi-C screenshots (Figure 2-figure supplement
975 1, Figure 2-figure supplement 3D-F, and Figure 2-figure supplement 4D-F) were generated using the 3D
976 Genome Browser (<http://promoter.bx.psu.edu/hi-c/index.html>). Virtual 4C plots used mESC Hi-C with 10 kb
977 resolution and a 25 kb viewpoint for ± 250 kb of the selected region. Screenshots were generated for mouse
978 (mESC, CH12, and NPC) and human (GM12878 and K562) cells using raw signal and 10 kb resolution. All Hi-C
979 datasets used were publicly available for mouse [15, 51] and human cells [15] (Supplementary File 4).

980

981 **TAD anchor identification.** TAD anchors were predicted for mouse liver using a modified version of the intra-
982 TAD loop prediction algorithm. The merged list of CTCF peaks was filtered to only consider peaks that were
983 found across all four biological replicates, that contained CTCF motifs, and that were within 50 kb of a TAD
984 boundary, as defined previously for mouse liver [3]. This 50 kb distance was chosen based on the ambiguity of
985 binned Hi-C data to more accurately determine the precise TAD boundary. Then, for each TAD boundary, all
986 pairs of (+) and (-) CTCF peaks were considered and scored based on their combined distance to the called
987 TAD boundary. Pairs of “+/-” peaks that were comprised of a (+) anchor upstream of a (-) anchor (i.e., CTCF
988 peak pairs that were not divergently oriented) were considered an invalid combination to define the end of one
989 TAD and the beginning of the next TAD, and were not considered. The valid pairs with the shortest combined

990 distance to the previously defined liver TAD boundary [3] were retained and all others were removed. If no valid
991 pair for a TAD boundary was identified, the single CTCF peak closest to the TSD boundary was retained as the
992 TAD anchor. A complete listing of TAD anchors is found in Supplementary File 1C, and a listing of inter-TAD
993 regions and associated gene ontology analysis is presented in Supplementary File 3.

994

995 **Alternative loop anchor analysis.** We sought to compare the relative insulation of loops identified by our
996 computational approach to alternative loops identified using the original core algorithm [31]. This provides an
997 objective measure to compare the performance of each implementation in identifying TAD-like loops and loop
998 anchors within TADs. To this end, we used the complete CTCF peak list from the merged sample as input and
999 implemented the loop prediction algorithm exactly as described previously (60% proportional peak cutoff, CTCF
1000 signal + motif scores as above, retaining only loops <200 kb) [31]. As summarized in Figure 2-figure supplement
1001 1B, this analysis yielded many more loops (60,678), which were shorter (median size of 61 kb) and showed less
1002 overlap with cohesin-mediated loops in the mESC ChIA-PET dataset (25.5% versus 63.2% for intra-TAD loops).
1003 59% of the intra-TAD loops characterized in our study are found in this larger loop list (termed “60k loops”). To
1004 characterize loops unique to the 60k loop set, we had to first filter out anchors found intra-TAD or TAD loops (to
1005 insure that each list was mutually exclusive, as above). Any 60k loop anchor within 50 kb of a TAD boundary
1006 was excluded from downstream analysis. Similarly, we also excluded any 60k loop with at least one anchor that
1007 coincided with an intra-TAD loop anchor. These mutually exclusive lists of intra-TAD loop anchors and the
1008 filtered set of 60k loop anchors (25,983 loop anchors in total; representing a subset of “Non Anchor CTCF” in
1009 the main text, and referred to as ‘26k loop anchors’ in Figure 2-figure supplement 1C,D) were then compared
1010 based on insulation of repressive histone marks (see “Repressive histone mark insulation”, below) and Hi-C
1011 interaction profiles (see “Additional Hi-C analysis”, above). Figure 2-figure supplement 1B,C compares the
1012 insular features of intra-TAD loop anchors to those of the 26k alternative loop anchors that are not intra-TAD
1013 loop or possible TAD anchors.

1014

1015 **Anchor/Loop overlap.** CTCF ChIP-seq data for 15 non-liver tissues from the ENCODE Project were
1016 downloaded (<https://genome.ucsc.edu/cgi-bin/hgTrackUi?db=mm9&g=wgEncodeLicrTfbs>) and intersected with
1017 replicates to form a single peak list for each tissue [42]. These single peak lists per tissue were then compared
1018 to liver CTCF peaks using the Bedtools multiinter command with the –cluster option to generate a union CTCF
1019 peak list for all tissues with a score representing the number of tissues in which a peak is present. “Lone” CTCF

1020 (CTCF sites lacking cohesin bound), other/non-anchor CAC sites, TAD anchors, and intra-TAD loop anchors
1021 were compared to this list to generate the histograms in Figure 2C (see also, Supplementary File 1C).
1022 Knockdown-resistant cohesin binding sites in liver were defined as Rad21 ChIP-seq peaks found in both wild-
1023 type (WT) and Rad21^{+/-} mouse liver, with knockdown-sensitive sites defined as Rad21 peaks found in WT liver
1024 that are absent in Rad21^{+/-} liver [19]. Similarly, knockdown-resistant cohesin binding sites in MEFs were defined
1025 as Smc1a ChIP-seq peaks present in both WT [18] and Stag1-knockout MEFs [83]. Knockdown-sensitive sites
1026 were defined as Smc1a peaks found in WT MEFs that are absent in Stag1-knockout MEFs. Phastcons 30-way
1027 vertebrate conservation scores were downloaded from the UCSC table browser and converted to BigWig tracks
1028 using ucscutils (version 20130327;
1029 <ftp://hgdownload.cse.ucsc.edu/goldenPath/mm9/phastCons30way/vertebrate>). Comparisons to mESC Smc1a
1030 ChIA-PET and Smc1a Hi-ChIP datasets [5, 52] were based on merged replicates, and reciprocal overlaps with
1031 intra-TAD loops were required (Bedtools intersect -wa -u -r -f 0.8 -a intraTADloops.bed -b mESC.bed). The
1032 mESC Smc1 ChIA-PET dataset was filtered to define “CTCF-CTCF” interactions as those with both anchor
1033 regions overlapping CTCF peak present in a minimum of 2 replicates (total of 3). Any remaining interactions
1034 were considered as CNC-mediated enhancer-promoter interactions for the analysis shown in Figure 2-figure
1035 supplement 3C.

1036
1037 **Repressive histone mark insulation.** To determine if a TAD or intra-TAD loop anchor CAC showed more
1038 insulation, or less insulation, than other classes of CTCF or cohesin binding sites, we used Jensen Shannon
1039 Divergence (JSD; [55]) to quantify the insulation of H3K27me3 and H3K9me3 ChIP-seq signals. Specifically,
1040 regions 10 kb upstream and 10 kb downstream of each peak in a given peak list (i.e., TAD anchors, CNC, etc.)
1041 were each divided into 50 bins of 200 bp each. The number of H3K27me3, H3K9me3, or IgG ChIP-seq reads
1042 within each bin was tallied, resulting in a vector of 50+50 values for each peak region. These were then
1043 compared to two test vectors representing complete (maximal) insulation: fifty 0's followed by fifty 1's, and fifty
1044 1's followed by fifty 0's. These are theoretical representations of low signal upstream of the peak followed by
1045 high signal downstream, and vice versa. Using a custom python script (Source Code File 1), the similarity
1046 between the experimentally derived vector and each of the test vectors was calculated, where a lower value
1047 represents less divergence from the test vector. The cumulative frequency distribution per group (anchors, CAC,
1048 CNC, etc.) is presented for the most similar test vector per peak in Figure 3D, 3E (K27me3 and K9me3) and

1049 Figure 3-figure supplement 2D (IgG). Heat maps show ChIP signal Z-transformed data across all CTCF-bound
1050 regions.

1051

1052 **Five Class DHS model.** The ~70,000 open chromatin regions (DHS) previously identified in mouse liver [58]
1053 were classified based on ChIP-seq signals for H3K4me1, H3K4me3, and CTCF within 1 kb of each DHS
1054 summit, obtained using the refinepeak option in MACS2. The general schematic is shown in Figure 4-figure
1055 supplement 1A. Promoter DHS were defined as DHS with a >1.5-fold ratio of H3K4me3 relative to H3K4me1
1056 ChIP-seq signal; enhancer DHS were defined as DHS with a <0.67-fold ratio of H3K4me3 relative to H3K4me1
1057 ChIP-seq signal, calculated as reads per million for each factor. Both DHS sets were filtered to remove DHS
1058 with <4 reads per million for both marks after subtracting IgG signal (Figure 4A). These cutoff values leave two
1059 remaining DHS groups, one with a roughly equal ratio between the two histone-H3 marks, and one with low
1060 signal (<4 reads per million) for both marks. The former DHS were classified as weak promoter DHS, based on
1061 their close proximity to RefSeq TSS and the low expression of neighboring genes (Figure 4-figure supplement
1062 1B,C). The remaining DHS group, characterized by low ChIP-seq signals, was largely intergenic but showed
1063 weak to undetectable levels of canonical histone marks. Low signal regions that overlapped a CTCF site with
1064 higher CTCF ChIP-seq signals than H3K4me1 signals were classified as potential insulators (Figure 4-figure
1065 supplement 1A). The remaining regions were designated weak enhancer-DHS based on their distance from
1066 TSS and their low levels of H3K27ac ChIP-seq signal compared to the enhancer-DHS group (Figure 4-figure
1067 supplement 1B). The majority of promoter-DHS and weak promoter-DHS were <1 kb from a TSS (Figure 4-
1068 figure supplement 1B). To compare the level of expression for genes with promoter-DHS or weak promoter-DHS
1069 (Figure 4-figure supplement 1C), the TSS was required to be within 10 kb of the DHS summit. Any gene with
1070 both a weak promoter-DHS and a promoter-DHS within 10 kb was categorized as being regulated by a
1071 promoter-DHS; thus, there was no overlap between weak promoter-DHS-regulated genes and promoter-DHS-
1072 regulated genes.

1073

1074 **Comparison of DHS classes across tissues.** All available mouse tissue DNase-seq peak regions were
1075 downloaded from the ENCODE Project website (<https://www.encodeproject.org/>) [42], ENCODE mm9 blacklist
1076 regions (<https://sites.google.com/site/anshulkundaje/projects/blacklists>) were removed, and the lists were
1077 merged to form a single reproducible peak list for each tissue, as follows. Due to variable replicate numbers
1078 across tissues, the following cutoffs were used to form merged DHS lists for each tissue. If a tissue had only two

1079 replicates (as was the case for 12 of the 20 non-liver tissues), we required that the DHS be present in both
1080 replicates. If a tissue had 3 or 4 replicates, then the DHS were required to be present in all or all but one
1081 replicate (this was the case for 7 of the 20 non-liver tissues). For whole brain tissue, the merged peak list
1082 required that a DHS was present in at least 5 of the 7 replicates. These regions were compared to each other
1083 using the Bedtools multiinter command with the `-cluster` option to generate a union DHS peak list for all tissues,
1084 where the score column represents the number of non-liver tissues in which a given region was found. For all
1085 liver DHS assigned to one of the above five DHS classes (Supplementary File 2A), each liver DHS summit was
1086 mapped to this all tissue union peak list, allowing only one match per summit up to 150 nt away. If a given liver
1087 DHS summit was >150 nt from the nearest DHS in any other tissue, it was given a score of “0” for liver-
1088 specificity. Otherwise the score represents the number of mouse tissues that the closest DHS was found in.

1089

1090 **Super-enhancer identification.** Super-enhancers were identified using the ROSE (Ranked Order of Super
1091 Enhancers) software package (http://younglab.wi.mit.edu/super_enhancer_code.html). ROSE takes a list of
1092 enhancer regions and mapped read positions as input to identify highly active clusters of enhancers. Default
1093 options were used, including 12.5 kb as the maximum distance for grouping (stitching) enhancers into putative
1094 super-enhancers, as well as reads per million normalization for all H3K27ac ChIP signal used for ranking
1095 enhancer clusters. The set of all enhancer-DHS and weak enhancer-DHS regions from the five class DHS
1096 model described above (Supplementary File 2A) was used as the region input list. A set of 19 publicly available
1097 H3K27ac ChIP-seq datasets from mouse liver was used as signal input (see Supplementary File 4 for sample
1098 information). This set includes datasets for male, female [57], and circadian time course (male only; [84]) mouse
1099 liver datasets. A strict intersection of super-enhancers identified across all 19 samples was used to define a set
1100 of 503 “core” super-enhancers in mouse liver using the Bedtools multiintersect command, as shown in Figure 4-
1101 figure supplement 2A. Any enhancer cluster (i.e., constituent enhancers within 12.5 kb, as above) not identified
1102 as a super-enhancer in any sample was termed a typical enhancer and considered as individual constituents
1103 only.

1104

1105 Gene targets for enhancers were assigned as the nearest gene (based on TSS position up to a maximum
1106 distance cutoff of 10 or 25 kb, as specified). Gene expression values and tissue specificity were defined as
1107 described above. Aggregate plots were generated using Deeptools (version 2.3.3). In Figure 4F, the scale-
1108 regions option of Deeptools was used to scale super-enhancers and typical enhancers to their median sizes of

1109 44 kb and 1 kb, respectively. Figure 4-figure supplement 2B used the reference-point mode of Deeptools and
1110 shows GRO-seq signal that overlaps eRNA loci as defined previously [62]. Super-enhancer and typical
1111 enhancer coordinates for mESC and ProB cells are from [61].

1112

1113 **DATA AVAILABILITY**

1114 Data generated and used in this study has been deposited in the Gene Expression Omnibus (GEO) under
1115 accession number GSE102999 (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE102999>). ChIP-seq
1116 data are available under the subseries GSE102997
1117 (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE102997>). 4C-seq data are available under the
1118 subseries GSE102998 (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE102998>). Published datasets
1119 used in this study are listed in Supplementary File 4.

1120

1121 **SUPPLEMENTARY DATA**

1122 Supplementary File 1: Intra-TAD loop and CTCF coordinates relevant to Figures 1-3.
1123 Supplementary File 2: Mouse liver DHS and Super-enhancer information relevant to Figure 4.
1124 Supplementary File 3: Genes and GO term enrichments for genes at refined TAD boundaries and in intra-TADs.
1125 Supplementary File 4: Publicly available datasets used and qPCR primers used to validate CTCF/Rad21 ChIPs.
1126 Source Code File 1: All code used for the prediction of intra-TAD loops and additional custom scripts.

1127

1128 **ACKNOWLEDGEMENTS**

1129 We thank Andy Rampersaud and Tisha Melia for their work building and standardizing the ChIP-seq and RNA-
1130 seq pipelines used in this study. We also thank Aram Shin for first piloting CTCF and Rad21 ChIP-seq
1131 experiments in this lab.

1132

1133 **FUNDING**

1134 This work was supported by National Institutes of Health grants [grant numbers DK33765, ES024421 to D.J.W];
1135 and by the National Science Foundation predoctoral fellowship [grant number DGE-1247312 to B.J.M]. Funding
1136 for open access charge: National Institutes of Health.

1137

1138 **CONFLICT OF INTEREST**

1139 The authors declare that they have no competing interests.

1140

1141 REFERENCES

- 1142 1. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B: **Topological domains in**
1143 **mammalian genomes identified by analysis of chromatin interactions.** *Nature* 2012, **485**:376-
1144 380.
- 1145 2. Nora EP, Lajoie BR, Schulz EG, Giorgetti L, Okamoto I, Servant N, Piolot T, van Berkum NL,
1146 Meisig J, Sedat J, et al: **Spatial partitioning of the regulatory landscape of the X-inactivation**
1147 **centre.** *Nature* 2012, **485**:381-385.
- 1148 3. Vietri Rudan M, Barrington C, Henderson S, Ernst C, Odom DT, Tanay A, Hadjur S: **Comparative**
1149 **Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture.** *Cell Rep*
1150 2015, **10**:1297-1309.
- 1151 4. Dixon JR, Jung I, Selvaraj S, Shen Y, Antosiewicz-Bourget JE, Lee AY, Ye Z, Kim A, Rajagopal N,
1152 Xie W, et al: **Chromatin architecture reorganization during stem cell differentiation.** *Nature*
1153 2015, **518**:331-336.
- 1154 5. Downen JM, Fan ZP, Hnisz D, Ren G, Abraham BJ, Zhang LN, Weintraub AS, Schuijers J, Lee TI,
1155 Zhao K, Young RA: **Control of cell identity genes occurs in insulated neighborhoods in**
1156 **mammalian chromosomes.** *Cell* 2014, **159**:374-387.
- 1157 6. Sexton T, Schober H, Fraser P, Gasser SM: **Gene regulation through nuclear organization.** *Nat*
1158 *Struct Mol Biol* 2007, **14**:1049-1055.
- 1159 7. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragooczy T, Telling A, Amit I, Lajoie
1160 BR, Sabo PJ, Dorschner MO, et al: **Comprehensive mapping of long-range interactions reveals**
1161 **folding principles of the human genome.** *Science* 2009, **326**:289-293.
- 1162 8. Pope BD, Ryba T, Dileep V, Yue F, Wu W, Denas O, Vera DL, Wang Y, Hansen RS, Canfield TK, et
1163 al: **Topologically associating domains are stable units of replication-timing regulation.**
1164 *Nature* 2014, **515**:402-405.
- 1165 9. Nora EP, Dekker J, Heard E: **Segmental folding of chromosomes: a basis for structural and**
1166 **regulatory chromosomal neighborhoods?** *Bioessays* 2013, **35**:818-828.
- 1167 10. Le Dily F, Bau D, Pohl A, Vicent GP, Serra F, Soronellas D, Castellano G, Wright RH, Ballare C,
1168 Filion G, et al: **Distinct structural transitions of chromatin topological domains correlate with**
1169 **coordinated hormone-induced gene regulation.** *Genes Dev* 2014, **28**:2151-2162.
- 1170 11. Ong CT, Corces VG: **CTCF: an architectural protein bridging genome topology and function.**
1171 *Nat Rev Genet* 2014, **15**:234-246.
- 1172 12. Sanborn AL, Rao SS, Huang SC, Durand NC, Huntley MH, Jewett AI, Bochkov ID, Chinnappan D,
1173 Cutkosky A, Li J, et al: **Chromatin extrusion explains key features of loop and domain**
1174 **formation in wild-type and engineered genomes.** *Proc Natl Acad Sci U S A* 2015, **112**:E6456-
1175 6465.
- 1176 13. Guo Y, Xu Q, Canzio D, Shou J, Li J, Gorkin DU, Jung I, Wu H, Zhai Y, Tang Y, et al: **CRISPR**
1177 **Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function.** *Cell*
1178 2015, **162**:900-910.
- 1179 14. Fudenberg G, Imakaev M, Lu C, Goloborodko A, Abdennur N, Mirny LA: **Formation of**
1180 **Chromosomal Domains by Loop Extrusion.** *Cell Rep* 2016, **15**:2038-2049.
- 1181 15. Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol
1182 I, Omer AD, Lander ES, Aiden EL: **A 3D map of the human genome at kilobase resolution**
1183 **reveals principles of chromatin looping.** *Cell* 2014, **159**:1665-1680.

- 1184 16. Phillips JE, Corces VG: **CTCF: master weaver of the genome.** *Cell* 2009, **137**:1194-1211.
- 1185 17. Zuin J, Dixon JR, van der Reijden MI, Ye Z, Kolovos P, Brouwer RW, van de Corput MP, van de
1186 Werken HJ, Knoch TA, van IWF, et al: **Cohesin and CTCF differentially affect chromatin**
1187 **architecture and gene expression in human cells.** *Proc Natl Acad Sci U S A* 2014, **111**:996-
1188 1001.
- 1189 18. Kagey MH, Newman JJ, Bilodeau S, Zhan Y, Orlando DA, van Berkum NL, Ebmeier CC, Goossens
1190 J, Rahl PB, Levine SS, et al: **Mediator and cohesin connect gene expression and chromatin**
1191 **architecture.** *Nature* 2010, **467**:430-435.
- 1192 19. Faure AJ, Schmidt D, Watt S, Schwalie PC, Wilson MD, Xu H, Ramsay RG, Odom DT, Flicek P:
1193 **Cohesin regulates tissue-specific expression by stabilizing highly occupied cis-regulatory**
1194 **modules.** *Genome Res* 2012, **22**:2163-2175.
- 1195 20. Uuskula-Reimand L, Hou H, Samavarchi-Tehrani P, Rudan MV, Liang M, Medina-Rivera A,
1196 Mohammed H, Schmidt D, Schwalie P, Young EJ, et al: **Topoisomerase II beta interacts with**
1197 **cohesin and CTCF at topological domain borders.** *Genome Biol* 2016, **17**:182.
- 1198 21. Heath HRdA, C; Sleutels, F; Dingjan, G; van de Nobelen, S; Jonkers, I; Ling, K; Gribnau, J;
1199 Renkawitz, R; Grosveld, F; Hendriks, R. W.; Galjart, N: **CTCF regulates cell cycle progression of**
1200 **aB T cells in the thymus.** *EMBO J* 2008, **27**:2839-2850.
- 1201 22. White JK, Gerdin AK, Karp NA, Ryder E, Buljan M, Bussell JN, Salisbury J, Clare S, Ingham NJ,
1202 Podrini C, et al: **Genome-wide generation and systematic phenotyping of knockout mice**
1203 **reveals new roles for many genes.** *Cell* 2013, **154**:452-464.
- 1204 23. Xu H, Balakrishnan K, Malaterre J, Beasley M, Yan Y, Essers J, Appeldoorn E, Tomaszewski JM,
1205 Vazquez M, Verschoor S, et al: **Rad21-cohesin haploinsufficiency impedes DNA repair and**
1206 **enhances gastrointestinal radiosensitivity in mice.** *PLoS One* 2010, **5**:e12112.
- 1207 24. Moore JM, Rabaia NA, Smith LE, Fagerlie S, Gurley K, Loukinov D, Disteche CM, Collins SJ,
1208 Kemp CJ, Lobanenkov VV, Filippova GN: **Loss of maternal CTCF is associated with peri-**
1209 **implantation lethality of Ctcf null embryos.** *PLoS One* 2012, **7**:e34915.
- 1210 25. Nora EP, Goloborodko A, Valton AL, Gibcus JH, Uebersohn A, Abdennur N, Dekker J, Mirny LA,
1211 Bruneau BG: **Targeted Degradation of CTCF Decouples Local Insulation of Chromosome**
1212 **Domains from Genomic Compartmentalization.** *Cell* 2017, **169**:930-944 e922.
- 1213 26. Rao SSP, Huang SC, Glenn St Hilaire B, Engreitz JM, Perez EM, Kieffer-Kwon KR, Sanborn AL,
1214 Johnstone SE, Bascom GD, Bochkov ID, et al: **Cohesin Loss Eliminates All Loop Domains.** *Cell*
1215 2017, **171**:305-320 e324.
- 1216 27. Schwarzer W, Abdennur N, Goloborodko A, Pekowska A, Fudenberg G, Loe-Mie Y, Fonseca NA,
1217 Huber W, C HH, Mirny L, Spitz F: **Two independent modes of chromatin organization**
1218 **revealed by cohesin removal.** *Nature* 2017, **551**:51-56.
- 1219 28. Ji X, Dadon DB, Powell BE, Fan ZP, Borges-Rivera D, Shachar S, Weintraub AS, Hnisz D,
1220 Pegoraro G, Lee TI, et al: **3D Chromosome Regulatory Landscape of Human Pluripotent Cells.**
1221 *Cell Stem Cell* 2016, **18**:262-275.
- 1222 29. Katainen R, Dave K, Pitkanen E, Palin K, Kivioja T, Valimaki N, Gylfe AE, Ristolainen H, Hanninen
1223 UA, Cajuso T, et al: **CTCF/cohesin-binding sites are frequently mutated in cancer.** *Nat Genet*
1224 2015, **47**:818-821.
- 1225 30. Fujimoto A, Furuta M, Totoki Y, Tsunoda T, Kato M, Shiraishi Y, Tanaka H, Taniguchi H,
1226 Kawakami Y, Ueno M, et al: **Whole-genome mutational landscape and characterization of**
1227 **noncoding and structural mutations in liver cancer.** *Nat Genet* 2016, **48**:500-509.
- 1228 31. Oti M, Falck J, Huynen MA, Zhou H: **CTCF-mediated chromatin loops enclose inducible gene**
1229 **regulatory domains.** *BMC Genomics* 2016, **17**:252.

- 1230 32. Handoko L, Xu H, Li G, Ngan CY, Chew E, Schnapp M, Lee CW, Ye C, Ping JL, Mulawadi F, et al:
1231 **CTCF-mediated functional chromatin interactome in pluripotent cells.** *Nat Genet* 2011,
1232 **43:630-638.**
- 1233 33. Fullwood MJ, Liu MH, Pan YF, Liu J, Xu H, Mohamed YB, Orlov YL, Velkov S, Ho A, Mei PH, et al:
1234 **An oestrogen-receptor-alpha-bound human chromatin interactome.** *Nature* 2009, **462:58-**
1235 **64.**
- 1236 34. Jin F, Li Y, Dixon JR, Selvaraj S, Ye Z, Lee AY, Yen CA, Schmitt AD, Espinoza CA, Ren B: **A high-**
1237 **resolution map of the three-dimensional chromatin interactome in human cells.** *Nature*
1238 2013, **503:290-294.**
- 1239 35. Hnisz D, Day DS, Young RA: **Insulated Neighborhoods: Structural and Functional Units of**
1240 **Mammalian Gene Control.** *Cell* 2016, **167:1188-1200.**
- 1241 36. Weinreb C, Raphael BJ: **Identification of hierarchical chromatin domains.** *Bioinformatics*
1242 2016, **32:1601-1609.**
- 1243 37. Sofueva S, Yaffe E, Chan WC, Georgopoulou D, Vietri Rudan M, Mira-Bontenbal H, Pollard SM,
1244 Schroth GP, Tanay A, Hadjir S: **Cohesin-mediated interactions organize chromosomal**
1245 **domain architecture.** *EMBO J* 2013, **32:3119-3129.**
- 1246 38. Tang Z, Luo OJ, Li X, Zheng M, Zhu JJ, Szalaj P, Trzaskoma P, Magalska A, Wlodarczyk J,
1247 Ruszczyccki B, et al: **CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin**
1248 **Topology for Transcription.** *Cell* 2015, **163:1611-1627.**
- 1249 39. Rowley MJ, Nichols MH, Lyu X, Ando-Kuri M, Rivera ISM, Hermetz K, Wang P, Ruan Y, Corces
1250 VG: **Evolutionarily Conserved Principles Predict 3D Chromatin Organization.** *Mol Cell* 2017,
1251 **67:837-852 e837.**
- 1252 40. Stevens TJ, Lando D, Basu S, Atkinson LP, Cao Y, Lee SF, Leeb M, Wohlfahrt KJ, Boucher W,
1253 O'Shaughnessy-Kirwan A, et al: **3D structures of individual mammalian genomes studied by**
1254 **single-cell Hi-C.** *Nature* 2017, **544:59-64.**
- 1255 41. Seitan VC, Faure AJ, Zhan Y, McCord RP, Lajoie BR, Ing-Simmons E, Lenhard B, Giorgetti L,
1256 Heard E, Fisher AG, et al: **Cohesin-based chromatin interactions enable regulated gene**
1257 **expression within preexisting architectural compartments.** *Genome Res* 2013, **23:2066-2077.**
- 1258 42. Shen Y, Yue F, McCleary DF, Ye Z, Edsall L, Kuan S, Wagner U, Dixon J, Lee L, Lobanenkov VV,
1259 Ren B: **A map of the cis-regulatory sequences in the mouse genome.** *Nature* 2012, **488:116-**
1260 **120.**
- 1261 43. Blanchette M, Bataille AR, Chen X, Poitras C, Laganriere J, Lefebvre C, Deblois G, Giguere V,
1262 Ferretti V, Bergeron D, et al: **Genome-wide computational prediction of transcriptional**
1263 **regulatory modules reveals new insights into human gene expression.** *Genome Res* 2006,
1264 **16:656-668.**
- 1265 44. Kim THB, L. O.; Zheng, M; Qu, C; Singer, M. A.; Richmond, T.A.; Wu, Y; Green, R.D.; Ren, B: **A**
1266 **high-resolution map of active promoters in the human genome.** *Nature* 2005, **436:876-880.**
- 1267 45. Yanai I, Benjamin H, Shmoish M, Chalifa-Caspi V, Shklar M, Ophir R, Bar-Even A, Horn-Saban S,
1268 Safran M, Domany E, et al: **Genome-wide midrange transcription profiles reveal expression**
1269 **level relationships in human tissue specification.** *Bioinformatics* 2005, **21:650-659.**
- 1270 46. Kryuchkova-Mostacci N, Robinson-Rechavi M: **A benchmark of gene expression tissue-**
1271 **specificity metrics.** *Brief Bioinform* 2017, **18:205-214.**
- 1272 47. Busslinger GA, Stocsits RR, van der Lelij P, Axelsson E, Tedeschi A, Galjart N, Peters JM:
1273 **Cohesin is positioned in mammalian genomes by transcription, CTCF and Wapl.** *Nature* 2017,
1274 **544:503-507.**

- 1275 48. Davidson IF, Goetz D, Zaczek MP, Molodtsov MI, Huis In 't Veld PJ, Weissmann F, Litos G,
1276 Cisneros DA, Ocampo-Hafalla M, Ladurner R, et al: **Rapid movement and transcriptional re-**
1277 **localization of human cohesin on DNA.** *EMBO J* 2016, **35**:2671-2685.
- 1278 49. Nakahashi H, Kwon KR, Resch W, Vian L, Dose M, Stavreva D, Hakim O, Pruett N, Nelson S,
1279 Yamane A, et al: **A genome-wide map of CTCF multivalency redefines the CTCF code.** *Cell Rep*
1280 2013, **3**:1678-1689.
- 1281 50. Xiao T, Wallace J, Felsenfeld G: **Specific Sites in the C Terminus of CTCF Interact with the SA2**
1282 **Subunit of the Cohesin Complex and Are Required for Cohesin-Dependent Insulation**
1283 **Activity.** *Molecular and Cellular Biology* 2011, **31**:2174-2183.
- 1284 51. Bonev B, Mendelson Cohen N, Szabo Q, Fritsch L, Papadopoulos GL, Lubling Y, Xu X, Lv X,
1285 Hugnot J-P, Tanay A, Cavalli G: **Multiscale 3D Genome Rewiring during Mouse Neural**
1286 **Development.** *Cell* 2017, **171**:557-572.e524.
- 1287 52. Mumbach MR, Rubin AJ, Flynn RA, Dai C, Khavari PA, Greenleaf WJ, Chang HY: **HiChIP:**
1288 **efficient and sensitive analysis of protein-directed genome architecture.** *Nat Methods* 2016,
1289 **13**:919-922.
- 1290 53. Kai Y, Andricovich J, Zeng Z, Zhu J, Tzatsos A, Peng W: **Predicting CTCF-mediated chromatin**
1291 **interactions by integrating genomic and epigenomic features.** *bioRxiv* 2017.
- 1292 54. Gómez-Marín C, Tena JJ, Acemel RD, López-Mayorga M, Naranjo S, de la Calle-Mustienes E,
1293 Maeso I, Beccari L, Aneas I, Vielmas E, et al: **Evolutionary comparison reveals that diverging**
1294 **CTCF sites are signatures of ancestral topological associating domains borders.** *Proceedings*
1295 *of the National Academy of Sciences* 2015, **112**:7542-7547.
- 1296 55. Fuglede B, Topsoe F: **Jensen-Shannon Divergence and Hilbert space embedding.** *IEEE*
1297 *International Symposium on Information Theory* 2004, **31**.
- 1298 56. Heidari N, Phanstiel DH, He C, Grubert F, Jahanbani F, Kasowski M, Zhang MQ, Snyder MP:
1299 **Genome-wide map of regulatory interactions in the human genome.** *Genome Res* 2014,
1300 **24**:1905-1917.
- 1301 57. Sugathan A, Waxman DJ: **Genome-Wide Analysis of Chromatin States Reveals Distinct**
1302 **Mechanisms of Sex-Dependent Gene Regulation in Male and Female Mouse Liver.** *Molecular*
1303 *and Cellular Biology* 2013, **33**:3594-3610.
- 1304 58. Ling G, Sugathan A, Mazor T, Fraenkel E, Waxman DJ: **Unbiased, genome-wide in vivo**
1305 **mapping of transcriptional regulatory elements reveals sex differences in chromatin**
1306 **structure associated with sex-specific liver gene expression.** *Mol Cell Biol* 2010, **30**:5531-
1307 5544.
- 1308 59. Wang Z, Zang C, Rosenfeld JA, Schones DE, Barski A, Cuddapah S, Cui K, Roh TY, Peng W, Zhang
1309 MQ, Zhao K: **Combinatorial patterns of histone acetylations and methylations in the human**
1310 **genome.** *Nat Genet* 2008, **40**:897-903.
- 1311 60. Hay D, Hughes JR, Babbs C, Davies JOJ, Graham BJ, Hanssen L, Kassouf MT, Marieke Oudelaar
1312 AM, Sharpe JA, Suci MC, et al: **Genetic dissection of the alpha-globin super-enhancer in**
1313 **vivo.** *Nat Genet* 2016, **48**:895-903.
- 1314 61. Whyte WA, Orlando DA, Hnisz D, Abraham BJ, Lin CY, Kagey MH, Rahl PB, Lee TI, Young RA:
1315 **Master transcription factors and mediator establish super-enhancers at key cell identity**
1316 **genes.** *Cell* 2013, **153**:307-319.
- 1317 62. Fang B, Everett LJ, Jager J, Briggs E, Armour SM, Feng D, Roy A, Gerhart-Hines Z, Sun Z, Lazar
1318 MA: **Circadian enhancers coordinate multiple phases of rhythmic gene transcription in vivo.**
1319 *Cell* 2014, **159**:1140-1152.

- 1320 63. Iborra FJ, Pombo A, Jackson DA, Cook PR: **Active RNA polymerases are localized within**
1321 **discrete transcription 'factories' in human nuclei.** *Journal of Cell Science* 1996, **109**:1427-
1322 1436.
- 1323 64. Willi M, Yoo KH, Reinisch F, Kuhns TM, Lee HK, Wang C, Hennighausen L: **Facultative CTCF**
1324 **sites moderate mammary super-enhancer activity and regulate juxtaposed gene in non-**
1325 **mammary cells.** *Nat Commun* 2017, **8**:16069.
- 1326 65. Hanssen LLP, Kassouf MT, Oudelaar AM, Biggs D, Preece C, Downes DJ, Gosden M, Sharpe JA,
1327 Sloane-Stanley JA, Hughes JR, et al: **Tissue-specific CTCF-cohesin-mediated chromatin**
1328 **architecture delimits enhancer interactions and function in vivo.** *Nat Cell Biol* 2017, **19**:952-
1329 961.
- 1330 66. Osborne CS, Chakalova L, Brown KE, Carter D, Horton A, Debrand E, Goyenechea B, Mitchell
1331 JA, Lopes S, Reik W, Fraser P: **Active genes dynamically colocalize to shared sites of ongoing**
1332 **transcription.** *Nat Genet* 2004, **36**:1065-1071.
- 1333 67. Martin P, McGovern A, Orozco G, Duffus K, Yarwood A, Schoenfelder S, Cooper NJ, Barton A,
1334 Wallace C, Fraser P, et al: **Capture Hi-C reveals novel candidate genes and complex long-**
1335 **range interactions with related autoimmune risk loci.** *Nat Commun* 2015, **6**:10069.
- 1336 68. Zhang Y, An L, Hu M, Tang J, Yue F: **HiCPlus: Resolution Enhancement of Hi-C interaction**
1337 **heatmap.** *bioRxiv* 2017.
- 1338 69. Lutz M, Burke LJ, Barreto G, Goeman F, Greb H, Arnold R, Schultheiß H, Brehm A, Kouzarides T,
1339 Lobanenkov V, Renkawitz R: **Transcriptional repression by the insulator protein CTCF**
1340 **involves histone deacetylases.** *Nucleic Acids Research* 2000, **28**:1707-1713.
- 1341 70. Ross-Innes CS, Brown GD, Carroll JS: **A co-ordinated interaction between CTCF and ER in**
1342 **breast cancer cells.** *BMC Genomics* 2011, **12**:593.
- 1343 71. Sun S, Del Rosario BC, Szanto A, Ogawa Y, Jeon Y, Lee JT: **Jpx RNA activates Xist by evicting**
1344 **CTCF.** *Cell* 2013, **153**:1537-1551.
- 1345 72. Saldana-Meyer R, Gonzalez-Buendia E, Guerrero G, Narendra V, Bonasio R, Recillas-Targa F,
1346 Reinberg D: **CTCF regulates the human p53 gene through direct interaction with its natural**
1347 **antisense transcript, Wrap53.** *Genes Dev* 2014, **28**:723-734.
- 1348 73. Shukla S, Kavak E, Gregory M, Imashimizu M, Shutinoski B, Kashlev M, Oberdoerffer P,
1349 Sandberg R, Oberdoerffer S: **CTCF-promoted RNA polymerase II pausing links DNA**
1350 **methylation to splicing.** *Nature* 2011, **479**:74-79.
- 1351 74. Nagano T, Lubling Y, Stevens TJ, Schoenfelder S, Yaffe E, Dean W, Laue ED, Tanay A, Fraser P:
1352 **Single-cell Hi-C reveals cell-to-cell variability in chromosome structure.** *Nature* 2013, **502**:59-
1353 64.
- 1354 75. Wang S, Su JH, Beliveau BJ, Bintu B, Moffitt JR, Wu CT, Zhuang X: **Spatial organization of**
1355 **chromatin domains and compartments in single chromosomes.** *Science* 2016, **353**:598-602.
- 1356 76. Vanhille L, Griffon A, Maqbool MA, Zacarias-Cabeza J, Dao LT, Fernandez N, Ballester B,
1357 Andrau JC, Spicuglia S: **High-throughput and quantitative assessment of enhancer activity in**
1358 **mammals by CapStarr-seq.** *Nat Commun* 2015, **6**:6905.
- 1359 77. Lupianez DG, Kraft K, Heinrich V, Krawitz P, Brancati F, Klopocki E, Horn D, Kayserili H, Opitz
1360 JM, Laxova R, et al: **Disruptions of topological chromatin domains cause pathogenic rewiring**
1361 **of gene-enhancer interactions.** *Cell* 2015, **161**:1012-1025.
- 1362 78. Raviram R, Rocha PP, Muller CL, Miraldi ER, Badri S, Fu Y, Swanzey E, Proudhon C, Snetkova V,
1363 Bonneau R, Skok JA: **4C-ker: A Method to Reproducibly Identify Genome-Wide Interactions**
1364 **Captured by 4C-Seq Experiments.** *PLoS Comput Biol* 2016, **12**:e1004780.

- 1365 79. van de Werken HJ, Landan G, Holwerda SJ, Hoichman M, Klous P, Chachik R, Splinter E, Valdes-
1366 Quezada C, Oz Y, Bouwman BA, et al: **Robust 4C-seq data analysis to screen for regulatory**
1367 **DNA interactions.** *Nat Methods* 2012, **9**:969-972.
- 1368 80. Thongjuea S, Stadhouders R, Grosveld FG, Soler E, Lenhard B: **r3Cseq: an R/Bioconductor**
1369 **package for the discovery of long-range genomic interactions from chromosome**
1370 **conformation capture and next-generation sequencing data.** *Nucleic Acids Res* 2013,
1371 **41**:e132.
- 1372 81. Melia T, Hao P, Yilmaz F, Waxman DJ: **Hepatic Long Intergenic Noncoding RNAs: High**
1373 **Promoter Conservation and Dynamic, Sex-Dependent Transcriptional Regulation by Growth**
1374 **Hormone.** *Mol Cell Biol* 2016, **36**:50-69.
- 1375 82. Servant N, Varoquaux N, Lajoie BR, Viara E, Chen CJ, Vert JP, Heard E, Dekker J, Barillot E: **HiC-**
1376 **Pro: an optimized and flexible pipeline for Hi-C data processing.** *Genome Biol* 2015, **16**:259.
- 1377 83. Remeseiro S, Cuadrado A, Carretero M, Martinez P, Drosopoulos WC, Canamero M,
1378 Schildkraut CL, Blasco MA, Losada A: **Cohesin-SA1 deficiency drives aneuploidy and**
1379 **tumorigenesis in mice due to impaired replication of telomeres.** *EMBO J* 2012, **31**:2076-
1380 2089.
- 1381 84. Koike N, Yoo SH, Huang HC, Kumar V, Lee C, Kim TK, Takahashi JS: **Transcriptional architecture**
1382 **and chromatin landscape of the core circadian clock in mammals.** *Science* 2012, **338**:349-
1383 354.

1385 **Figure Legend:**

1386 **Figure 1: Features of TAD boundaries and TAD insulator function**

- 1387 Profiles in A-E represent a normalized aggregate count of peaks or features along the length of all TADs, sub-
1388 divided into 100 equally-sized bins per TAD, where bin #1 is the 5' start of the TAD and bin #100 is at the
1389 TAD 3' end. Normalization was performed to allow comparison of multiple groups with variable peak
1390 numbers in a single figure. The y-axis displays the enrichment within a given bin versus the average of the 5
1391 center bins (bins #48-52). In A-C, the number of binding sites in each group is shown in parenthesis.
- 1392 A. Cohesin-and-CTCF (CAC) sites are enriched at TAD boundaries, while cohesin-non-CTCF (CNC) sites are
1393 weakly depleted. As the cohesin (Coh) complex is a multi-protein complex, the darker color within each
1394 group represents a stricter overlap of three cohesin subunits (Rad21, Stag1 and Stag2).
- 1395 B. In both mouse liver and MEFs, cohesin binding sites that are resistant to knockdown (KD) or knockout (KO)
1396 of cohesin component subunits (~40% of cohesin binding sites) are strongly enriched for TAD boundaries.
1397 Cohesin sites that are sensitive to loss following KD or KO (~60% of sites) are not enriched at TAD
1398 boundaries.
- 1399 C. CTCF binding sites in liver that are deeply-shared across other ENCODE tissues (≥ 12 out of 15 other
1400 tissues examined) are strongly enriched at liver TAD boundaries, while those that are either unique to liver
1401 or shared in only one other tissue are not enriched at TAD boundaries.
- 1402 D. TAD boundaries show greater hypomethylation than the TAD interior. The most hypomethylated CpGs are
1403 enriched at TAD boundaries, which likely represents a combination of hypomethylation at gene promoters
1404 and hypomethylation at CTCF binding sites. CpG methylation states, determined by liver whole genome
1405 bisulfite sequence analysis were subdivided into 10 bins based on the degree of methylation (0-10%
1406 methylated, 10-20%, etc.) prior to TAD distribution analysis.

- 1407 E. 10 liver-expressed TFs are not enriched at TAD boundaries. These profiles are representative of the vast
1408 majority of the >50 publically available ChIP peak lists for liver-expressed TFs. Notable exceptions, related
1409 to promoter-associated features, marks, and transcription factors, are shown in Figure 1-figure supplement
1410 1B,D.
- 1411 F. Shown is a heat map of the distribution of the indicated activating and repressive marks and other features
1412 across a 1 Mb window around each TAD boundary. TAD clusters, numbered at the left, were defined using
1413 k means clustering (k=4). The boundaries between TADs transition from active to inactive chromatin
1414 compartments (or vice versa) for TAD clusters 2 and 3. In downstream analyses based on these results, a
1415 TAD was considered active if the boundary at the start of a TAD fell into clusters 1 or 2 and the boundary at
1416 the end of the same TAD fell into clusters 1 or 3; inactive TADs are those whose boundaries begin in
1417 clusters 3 or 4 and end in clusters 2 or 4 (see Methods). See Supplementary File 1A for a full listing of the
1418 3,538 autosomal TADs analyzed and their active/inactive status.
- 1419 G. UCSC browser screenshot for a transitional TAD boundary on chromosome 13 from TAD cluster 3 in Figure
1420 1F. Arrows at bottom indicate CTCF motif orientation.
- 1421 H. Box plots showing liver gene expression (RNA-seq) for 12,258 genes in 1,930 active TADs and 4,643 genes
1422 in 1,000 inactive TADs (Supplementary File 1A). 939 genes in 473 of the inactive TADs are expressed at >
1423 1 FPKM (Supplementary File 1E). Genes found in active compartment TADs are more highly expressed,
1424 with the majority of genes showing >1 FPKM, than genes found in inactive TAD compartments. Genes in
1425 weakly active and weakly inactive TADs were excluded from these analyses.
- 1426 I. Genes whose TSS are located in inactive TADs ('B compartments') are more tissue specific in their
1427 expression pattern than genes found in active TADs ('A compartments'). The top GO category for expressed
1428 genes in the A compartment is RNA binding, while the top category for expressed genes in the B
1429 compartment is monooxygenase activity (not shown).

1430
1431 **Figure 2: Predicted intra-TAD loop anchors share many properties of TAD anchors**

- 1432 A. Diagram illustrating intra-TAD loop prediction based on CTCF motif orientation and CTCF and cohesin
1433 (Rad21) ChIP-seq binding strength data. Iteration was conducted until 20,000 loops were predicted per
1434 sample, prior to filtering and intersection across samples, as detailed in Methods.
- 1435 B. Shown is a 2 Mb segment of mouse chromosome 2 indicating TAD loops (blue) and intra-TAD loops (pink)
1436 in relation to genes. Also shown are cohesin interaction loops identified experimentally in mouse ESC by
1437 Smc1 ChIA-PET [5].
- 1438 C. TAD and predicted intra-TAD loop anchors are more tissue ubiquitous than other categories of CTCF/CAC
1439 sites. Each of the four CTCF site subgroups was defined in mouse liver as detailed in Supplementary File
1440 1C. The x-axis indicates the number of ENCODE tissues out of 15 tissues examined that also have CTCF
1441 bound, where a higher value indicates more tissue-ubiquitous CTCF binding. These data are shown for
1442 "lone" CTCF binding sites (10,553), non-anchor cohesin-and-CTCF sites ("Other CAC"; 26,970), TAD
1443 anchors (5,861), and intra-TAD loop anchors (9,052, which excludes those at a TAD loop anchor). While
1444 "Other CAC" sites tend to be weaker (Figure 2F, below), 93% are bound by CTCF in at least one other
1445 mouse tissue, and 66% were verified in at least 6 other tissues. Similarly, for "Lone CTCF", 81% of sites

- 1446 were bound by CTCF in at least one other mouse tissue, and 39% were verified in at least 6 other tissues
1447 (*not shown*).
- 1448 D. TAD and intra-TAD loop anchors are more resistant to the knockdown effects of Rad21^{+/-} haploinsufficiency
1449 than other CAC sites or cohesin-bound regions. A larger fraction is also bound by the novel extrusion
1450 complex factor Top2b (Supplementary File 1C).
- 1451 E. Loop anchors show greater intra-motif conservation than other CTCF-bound regions. Shown are the
1452 aggregate PhastCons score for oriented core motifs within either TAD (dark blue) or intra-TAD (light blue)
1453 anchors as compared to other CTCF peaks with motifs (yellow).
- 1454 F. Cohesin interacts with the COOH terminus of CTCF [50], which resulting in a shift of ~20 nt in cohesin ChIP
1455 signal relative to the CTCF summit (c.f. shift to the right of vertical red line) regardless of category of CTCF
1456 binding site (anchor/non-anchor). Blue arrows indicate the CTCF motif orientation and red triangles and
1457 vertical lines indicate position of the CTCF signal summit.

1458

1459 **Figure 3: Intra-TAD loops show directional interactions and insulate chromatin marks**

- 1460 A. TADs and intra-TAD loops both show a stronger orientation of interactions downstream of the motif than
1461 other CTCF-bound regions. TAD anchors also show higher inward bias than intra-TAD loops ($p < 0.0001$,
1462 KS t-test for pairwise comparisons). Inward bias is a chi-square-based metric similar to directionality index
1463 but defined on a per peak basis and oriented relative to the motif within the anchor/non-anchor peak. For
1464 this and all other anchor comparisons, anchors that are shared between TADs and intra-TAD loops were
1465 excluded from the intra-TAD group to ensure a fair comparison. Anchors shared between TADs and intra-
1466 TAD loops were considered as TAD anchors only.
- 1467 B. Virtual 4C analysis in mESC for a genomic region nearby *Ccl22* on mouse chromosome 8. The data shows
1468 a shift in Hi-C read distribution around intra-TAD loop anchors, indicating insulation. mESC Hi-C data from
1469 [51] was plotted across a 500 kb window surrounding the virtual 4C viewpoint, which is marked by a red line.
1470 Viewpoints were selected to be at the start of an intra-TAD loop (“IN”) as well as an adjacent upstream
1471 control region that does not overlap an intra-TAD loop anchor. The percentages shown indicate the
1472 distribution of interaction reads upstream and downstream of the viewpoint, over the 500 kb region, as
1473 shown. Orange shading indicates the width of the intra-TAD loop region interrogated. Chromosomal
1474 coordinates are for mouse genome mm10.
- 1475 C. Shown are aggregate plots generated from mouse liver Hi-C data [3] for each set of TAD and intra-TAD loop
1476 anchors, for the set of non-anchor CTCF sites listed in Supplementary File 1C, and for the set of CNC sites
1477 (Supplementary File 1D), which serves as a control. In aggregate, TAD and intra-TAD loop anchors show
1478 stronger and more directionally-biased interactions (contact enrichment, red) than other non-anchor CTCF
1479 bound genomic regions. They also show a greater depletion of distal chromatin interactions (contact
1480 depletion, blue). TAD anchors also show greater distal contact enrichment with the anchor and more local
1481 contact depletion spanning the anchor than intra-TAD loops. Red triangles indicate locations of left and right
1482 loop anchors and blue arrows indicate CTCF motif orientation. Shading indicates an enrichment (red) or
1483 depletion (blue) of contact frequency relative to a genome-wide background model.

- 1484 D. Shown are JSD values for four classes of mutually exclusive CTCF binding sites (TAD anchors, intra-TAD
 1485 loop anchors, other CAC sites, and CTCF sites lacking cohesin) as well as CNC sites, which are primarily
 1486 found at enhancers. TAD and intra-TAD loop anchors show greater insulation of the repressive histone
 1487 marks as measured by Jensen Shannon divergence between H3K27me3 and H3K9me3 ChIP-seq signal
 1488 upstream and downstream of the anchor region.
- 1489 E. Shown are the top 500 active insulated intra-TAD loops, based on high H3K27me3 ChIP-seq signal outside
 1490 the intra-TAD loop (red), and low H3K27me3 signal within the intra-TAD loop (blue). Data are expressed as
 1491 a Z-score of the H3K27me3 signal per bin relative to all H3K27me3 signals within a 20 kb widow centered
 1492 on all CTCF-bound regions.
- 1493 F. Shown are the top 500 inactive insulated intra-TAD loops, based on high signal H3K27me3 signal inside the
 1494 intra-TAD loops (red) and low H3K27me3 signal in neighboring regions (blue). Signal is shown as a Z-score
 1495 of H3K27me3 signal, as in E. At right is shown the IgG signal distribution as a negative control for the
 1496 upstream anchors of inactive intra-TAD loops (see Figure 3-figure supplement 2D for all panels).

1497

1498 **Figure 4: Categorization of DHS-based regulatory elements in mouse liver**

- 1499 A. Classification of open chromatin regions (DHS) in mouse liver, based on relative intensities for a
 1500 combination of H3K4me1 and H3K4me3 marks, and CTCF ChIP-seq data Based on the combinatorial
 1501 signal from these three datasets, five groups of DHS were identified: promoter-DHS, weak promoter-DHS,
 1502 enhancer-DHS, weak enhancer-DHS, and insulator-DHS, as described in Methods and in Figure 4-figure
 1503 supplement 1A.
- 1504 B. Shown is a heatmap representation of the simplified five-class DHS model shown in panel A, which
 1505 captures features such as CNC enrichment at enhancers and K27ac enrichment at enhancers and
 1506 promoters, with additional features described in Figure 4-figure supplement 1A.
- 1507 C. Scheme for using 19 published mouse liver H3K27ac ChIP-seq datasets to identify a core set of 503 use
 1508 liver super-enhancers using the ROSE software package (Supplementary File 2B). These 503 super-
 1509 enhancers were identified in all 19 samples, indicating they are active in both male and female liver, and
 1510 across multiple circadian time points.
- 1511 D. Genes associated with super-enhancers (SE) are more highly expressed ($\log_2(\text{FPKM}+1)$ values) than
 1512 genes associated with typical enhancers (TE), for both protein coding genes and liver-expressed multi-
 1513 exonic lncRNA genes. The super-enhancer-adjacent genes are also more tissue specific (higher Tau score)
 1514 than typical enhancer-adjacent genes. ****, KS t-test, $p < 0.0001$ for pairwise comparisons of SE-adjacent
 1515 genes vs. TE-adjacent genes.
- 1516 E. Venn diagrams show substantial overlap between typical enhancer gene targets across tissues (liver, ESCs,
 1517 ProB cells), but limited overlap between super-enhancer adjacent genes (within 10 kb of the super-
 1518 enhancer) for the same tissues. The numbers represent the percent of genes targeted in a given tissue by
 1519 the indicated class of enhancer (typical enhancers or super-enhancers) that are not targets of the
 1520 corresponding class of enhancers in the other two tissues. For example, 93.2% of genes targeted by liver
 1521 super-enhancers are not targeted by the set of super-enhancers identified in either Pro-B or mouse ESCs.

- 1522 Gene targets of each enhancer class were identified by GREAT using default parameters, then filtered to
1523 keep only those ≤ 10 kb from the enhancer.
- 1524 F. ChIP and DNase-seq signal at typical enhancers and super-enhancers, scaled to their median length (1 kb
1525 and 44 kb respectively; indicated by hash marks) flanked by 10 kb up- and down-stream. Super-enhancers
1526 show much greater accumulation of RNA polymerase 2, despite little or no apparent enrichment for the
1527 promoter mark H3K4me3.
- 1528 G. Super-enhancers (SE) target distinct categories of genes than typical enhancers (TE) in mouse liver. Thus,
1529 while GO terms such as oxidoreductase activity are enriched in the set of gene targets for both classes of
1530 enhancers, only super-enhancers are enriched for transcription-regulated terms (e.g., Regulation of
1531 transcription, Steroid hormone receptor activity) (Supplementary File 2C,D). Numbers represent the overlap
1532 of GO terms (either Molecular Function or Biological Process) in any DAVID annotation cluster (with an
1533 enrichment score > 1.3) enriched for genes regulated by either typical enhancers or super-enhancers.

1534

1535 **Figure 5: Impact of intra-TAD loops on gene expression**

- 1536 A. Two possible gene targets were assigned for each super-enhancer within an intra-TAD loop, one target
1537 gene for which the TSS is within the intra-TAD loop and another target gene for which the TSS is outside of
1538 the intra-TAD loop but is within 25 kb of the intra-TAD loop anchor. Box plots show that gene targets within
1539 an intra-TAD loop are significantly more highly expressed than the alternative, linearly more proximal, gene
1540 target.
- 1541 B. Shown is the standard deviation in Tau values (tissue-specificity index) of genes whose TSS's are within
1542 TADs or intra-TAD loops that contain at least three TSS. Genes within intra-TAD loops tend to be more
1543 uniformly tissue-specific or tissue-ubiquitous when compared to all genes within TADs, or when compared
1544 to a shuffled set of random regions matched in size to intra-TADs. Thus, sets of three or more genes within
1545 intra-TAD loops are consistently either more or less tissue specific than random clusters of genes within the
1546 same sized genomic spans.
- 1547 C-D. TAD and intra-TAD loops insulate a subset of super-enhancers (black horizontal bars) with key liver genes,
1548 allowing high expression of genes such as the TFs *Cebpb* and *Hnf4a*, relative to their immediate neighbors.
1549 *Cebpb* is an example at the TAD scale, while *Hnf4a* shows an intra-TAD loop. In both cases, the most
1550 linearly proximal gene is outside the TAD or intra-TAD loop and is expressed at a lower level than the loop-
1551 internal genes (and presumptive gene target).
- 1552 E. Shown is a UCSC genome browser screenshot of a transition from an active to a repressed TAD, with the
1553 expression of genes within the region shown in a bar graph, below. Insulated intra-TAD loops allow for
1554 expression of select gene targets within otherwise repressed genomic compartments. The obesity-related
1555 gene *Scd1* is insulated in an intra-TAD loop and is the only expressed gene in its TAD (FPKM > 100).
1556 H3K27me3 marks are shown both as reads per million signal track (below) and as signal over an IgG input
1557 control (above), expressed as $\log_2[(\text{H3K27me3 signal}) / (\text{Input signal})]$.

1558

1559 **Figure 6: *Alb* 4C-seq exemplifies intra-TAD insulation and super-enhancer interaction**

- 1560 A. The *Alb* promoter makes multiple directional contacts with the adjacent super-enhancer region in both male
1561 (M) and female (F) mouse liver, as determined by 4C-seq with a viewpoint at the *Alb* promoter. All
1562 reproducible interactions occur within the TAD loop containing the *Alb* TSS and its super-enhancer (red bar
1563 beneath H3K27ac track), and all but two contacts in male liver occur within the predicted intra-TAD loops
1564 (pink). 4C-seq interaction scores are shown as $-\log_{10}(pval)$ values across replicates, as calculated by R3C-
1565 seq (see Methods). Also see Figure 6-figure supplement 1.
- 1566 B. The 4C-seq interaction signal within the *Alb* TAD is orders of magnitude above the background signal and
1567 generally decays with distance. Far-*cis* and *trans* interactions are represented on a per TAD basis,
1568 expressed as RPKM per TAD, to control for sequencing depth and TAD length. The overall background
1569 within mouse chromosome 5 is significantly higher than all *trans* chromosomes; immediately adjacent TADs
1570 also show higher 4C-seq signal than the overall *cis* background. The 4C-seq signal decayed to background
1571 levels within ~ 3 TADs of the *Alb* viewpoint TAD. Each data point represents a single TAD and each color
1572 represents a 4C-seq replicate.
- 1573 C. Background model used for distal *cis* interactions, showing a rapid decay in per TAD signal intensity. Each
1574 data point represents a single TAD along chromosome 5.
- 1575 D. Distal *cis* and *trans* TADs that highly interact with the *Alb* promoter tend to be active TADs, while a majority
1576 of the TADs that interact less than the background model are predicted to be inactive. A simple inverse
1577 logarithmic decay of signal per TAD was used to determine the background signal along the *cis*
1578 chromosome, while the 4Cker package was used to determine high, medium, low, and non-interacting TADs
1579 in *trans* based on a hidden markov model with adaptive windows better suited for low signal regions.

1580

1581 **Supplemental Figure Legend:**

1582 **Figure 1-figure supplement 1: Additional Features of TAD boundaries**

- 1583 A. CTCF binding orientations are divergent at TAD boundaries. The top plot indicates bin enrichment
1584 relative to the TAD center, as in Figure 1A. The X-axis begins at the midpoint of one TAD, crosses the
1585 TAD boundary, and then extends to the midpoint of the subsequent TAD. The lower plot shows the in
1586 peak ChIP strength relative to the TAD boundary across a 1 Mb window. Blue shows the signal for
1587 CTCF ChIP-seq reads that overlap a (+) CTCF peak, while orange shows the CTCF ChIP-seq reads
1588 that overlap a (-) CTCF peak. Reads that either do not overlap a peak or fall in a peak with no CTCF
1589 motif are excluded.
- 1590 B-F. Shown are normalized aggregate count for the indicated peaks or features along the length of all TADs,
1591 sub-divided into 100 equally-sized bins per TAD, where bin #1 is at the 5' end of the TAD and bin #100
1592 is at the TAD 3' end. Normalization was performed to allow comparison of multiple groups with variable
1593 peak numbers and different TAD lengths in a single figure. The y-axis displays the enrichment within a
1594 given bin versus the average of the 5 center bins (bins #48-52).
- 1595 B. Two TFs, TP and E2F4, show enrichment for TAD boundaries, likely due to their strong bias for binding
1596 at gene promoters.
- 1597 C. The TSS's of protein-coding genes are enriched at TAD boundaries, while liver-expressed lncRNA TSS
1598 show little or no TAD boundary enrichment.

- 1599 D. Three promoter-associated histone marks, but not the enhancer mark H3K4me1, are enriched at TAD
1600 boundaries. This provides further evidence that TAD boundary regions are actively transcribed.
- 1601 E. Promoters and weak promoters that lack CAC binding are enriched at TAD boundaries. Similarly, CAC
1602 that do not overlap promoters or weak promoters are enriched at TAD boundaries. Thus, we expect that
1603 promoter and CAC enrichment at TAD boundaries are distinct phenomena and not simply an artifact of
1604 CAC binding at some promoters. Promoters and weak promoters were defined by the presence of a
1605 DHS and the ratio of H3K4me3 to H3K4me1 signal
1606 (see Methods).
- 1607 F. Intra-CTCF-motif CpGs are highly hypomethylated compared to the genome-wide average and to
1608 neighboring CpGs within 10 kb (“CpG <10 kb Away”). This is most pronounced for TAD and intra-TAD
1609 anchors. The strength of CTCF binding is more strongly associated with methylation level than with
1610 motif score (data not shown).
- 1611 G. Genes immediately adjacent to TAD boundaries are more tissue ubiquitous in expression (i.e.,
1612 housekeeping genes). Each bar represents the indicated percentile bin into the adjacent TADs, and by
1613 3 bins into the TAD from each end (or 6% of the total TAD length), the genes are not significantly (ns)
1614 more tissue ubiquitous in expression than the overall genome-wide level (horizontal dashed line). Bins
1615 were defined as above, with each TAD divided into 100 bins of equal length per TAD.

1616

1617 **Figure 2-figure supplement 1: Comparison of CTCF Features within TADs and Loop Prediction**

1618 **Improvements**

- 1619 A. Fewer than 25% of CTCF, CAC, and tissue ubiquitous CTCF sites are within 25 kb of a TAD boundary.
1620 While these features are strongly enriched at TAD boundaries (Figure 1A), the vast majority of any
1621 subgroup of CTCF sites is still TAD-internal.
- 1622 B. Comparison of features between the set of 9,543 mouse liver intra-TAD loops predicted in this study
1623 and an alternative set of 60,677 mouse liver loops predicted using the method described previously [31],
1624 without any additional filtering. These “60k loops” tend to be shorter, show much less overlap with
1625 mESC ChIA-PET loops, and only capture 59% of intra-TAD loops, as shown. To determine properties of
1626 the anchors for the 60k loop set, we considered a subset comprised of 25,983 unique alternative loop
1627 anchors (i.e. loop anchors that are not also anchors of intra-TAD loops or TADs; see Methods). This
1628 “26k loop anchor” subset shows many fewer directional interactions and less insulation (median IBI and
1629 JSD, respectively; also see Figure 2-figure supplement 1C,D).
- 1630 C. The set of 9,052 intra-TAD loop anchors (Supplementary File 1B) shows greater insulation of repressive
1631 histone marks than the set of 25,983 other putative CTCF-mediated loops (“26k anchors”) using a prior
1632 iterative loop prediction method [31]. CTCF peaks identified in the merged sample (combination of all
1633 biological replicates) were used as input for computational loop prediction exactly as described in [31],
1634 without consideration of cohesin strength/binding and without applying any additional filters. This loop
1635 list was then filtered to remove any loop that shares an anchor with an intra-TAD loop or whose anchor
1636 is within 50 kb of a TAD boundary. Shown here are the insulation scores (JSD) around intra-TAD loop

1637 anchors and other putative loop anchors (as defined in panel B) for H3K27me3 and H3K9me3 ChIP-seq
 1638 read distribution, both of which show greater insulation around intra-TAD loop anchors.

1639 D. Intra-TAD loop anchors show greater insulation of Hi-C data-based interactions as well as stronger
 1640 directionality of interactions than other putative CTCF loops. The graphic for intra-TAD loop anchors is
 1641 reproduced from Figure 3C. The 26k loop anchors defined in panels B were split into left (upstream) and
 1642 right (downstream) anchors based on CTCF motif orientation. See Figure 3C for further details.

1643 E. 91% of predicted intra-TAD loops are wholly contained within a single TAD, even without filtering for
 1644 TAD or TSS overlap. This compares to 67% for a random shuffled set of 9,543 regions of equal
 1645 length and number to the set of intra-TAD loops.

1646 F-H. Many of the intra-TAD loop structures that we predicted for mouse liver can be seen in the high
 1647 resolution Hi-C data from the mouse B-cell lymphoma cell line CH12-LX (12). TADs are marked in each
 1648 panel as horizontal red lines. Shown beneath each red line are the liver CTCF and cohesin ChIP-seq
 1649 data used to predict the liver intra-TAD loops indicated. Panel F shows two examples of single intra-
 1650 TAD loops within TADs. Panel G shows examples of nested intra-TAD loops, where one intra-TAD loop
 1651 anchor is predicted to interact with more than one CAC anchor. Finally, Panel H shows more complex
 1652 subdivision of TADs into multiple intra-TAD loops. The top section of each panel shows the Hi-C data
 1653 from CH12-LX cells, while the lower section of each panel presents our data from mouse liver. Red
 1654 arrowheads mark focal peaks in the contact matrix, which correspond to the midpoints of the predicted
 1655 intra-TAD loops.

1656
 1657 **Figure 2-figure supplement 2: Subclasses of CTCF binding events in relation to predicted loops**

1658 A. Summary of CTCF sites, CAC sites, TAD loop anchors, and intra-TAD loop anchors in mouse liver
 1659 based on lists in Supplementary File 1C. (1) CTCF peaks found in at least two biological replicate
 1660 samples (n=52,436). (2) The subset of the above CTCF sites that also overlap a cohesin (Rad21)
 1661 binding site (n=42,801). (3) CTCF sites predicted to be involved in an intra-TAD loop (9,052) or a TAD
 1662 loop (5,861). Due to some ambiguity and nesting of many loop structures, the number of intra-TAD loop
 1663 anchors shown is substantially less than the total number of intra-TAD loops (9,543) multiplied by a
 1664 factor of 2.

1665 B. Cohesin appears to be the primary contributing factor for topoisomerase-II β (Top2b) interaction with
 1666 CTCF, as Top2b is only present at 8.5% of CTCF sites lacking cohesin, but is found at 56% of cohesin
 1667 sites lacking CTCF (i.e., CNC sites).

1668 C. De novo motif discovery for loop anchors did not reveal any specific motifs that differentiate loop
 1669 anchors from other CTCF-bound regions. *De novo* motif discovery was performed using Homer. In
 1670 some cases, evidence of expanded CTCF motif usage was observed downstream of the core motif
 1671 (region 3, loop interior), however, fewer than 5% of the genomic regions analyzed contained any of the
 1672 specific motifs discovered. We found evidence for the M2 motif (region 2, loop exterior) in all groups,
 1673 fitting with the small secondary peak of conservation just upstream of the core motif. The PhastCons
 1674 intra-motif conservation figure duplicates that shown in Figure 2E.

1675 D. Analysis of loop anchors for known motifs did not identify any specific motifs that differentiate loop
1676 anchors from other CTCF-bound regions that contain the core CTCF motif (MA0139.1). Some modest
1677 degree of sequence optimization may occur at loop anchors, as additional CTCF motifs were
1678 consistently present at a higher proportion of loop anchors than at other CTCF-bound regions. The motif
1679 for Znf143 and the specific “M2” CTCF motif showed no differences and were found in ~20% of the
1680 genomic regions in each group. JASPAR IDs are indicated when available. Position weight matrices for
1681 “additional *de novo* CTCF motifs” can be found at CTCFBSDDB 2.0
1682 (http://insulatordb.uthsc.edu/download/CTCFBSDDB_PWM.mat).
1683

1684 **Figure 2-figure supplement 3: Intra-TAD loop prediction in two other mouse cell types: mESC and NPC**

1685 A. Comparison of CAC-mediated intra-TAD loop predicted for mouse embryonic stem cells (mESCs) and
1686 neural progenitor cells (NPCs) with those predicted in mouse liver. The number of loops present after
1687 merging predictions across all replicates is shown (column 1), followed by the number of loops that do
1688 not substantially overlap TADs (i.e., loops that show <80% reciprocal overlap with TADs; column 2).
1689 Column 3 shows the number of loops that contain ≥ 1 TSS (either protein-coding gene or multi-exonic
1690 long noncoding RNA TSS). Overlaps are presented for this set of filtered loops (column 3) with the sets
1691 of intra-TAD loops predicted in liver, Insulated Neighborhoods in mESC [35], and “CTCF-CTCF”
1692 interactions from Smc1 ChIA-PET experiments in mESC [5]. As the size of each group is different,
1693 overlaps in columns 5 and 6 show the percent overlap relative to the group indicated in the column
1694 header (either liver intra-TAD loops or insulated neighborhoods) followed by parentheses showing the
1695 percent overlap relative to the row group (either liver, mESC, or NPC intra-TAD loops).
1696 B. Shown is the overlap between mESC and NPC intra-TAD loops (63%), which is similar to the overlap
1697 between mESC and NPC TADs. this indicates that intra-TAD loops show a similar, or even somewhat
1698 higher, level of tissue ubiquity as do TADs.
1699 C. Tissue-specific intra-TAD loops are weaker than those shared across liver, mESC, and NPCs. To
1700 compare the relative strength of loops predicted by our approach, we divided Smc1 ChIA-PET loops [5]
1701 into 5 groups: ‘+++’, meaning “CTCF-CTCF” interactions that overlap intra-TAD loops predicted in all
1702 three cell types; ‘++-’, meaning “CTCF-CTCF” interactions that overlap intra-TAD loops predicted in
1703 mESCs and only one other cell type; ‘+--’, meaning “CTCF-CTCF” interactions that overlap intra-TAD
1704 loops predicted in mESCs and no other cell type; ‘---’, meaning “CTCF-CTCF” interactions that do not
1705 overlap any predicted intra-TAD loop in mESC; or ‘CNC’, meaning cohesin-non-CTCF interactions or
1706 cohesin-mediated interactions that are not anchored by mESC CTCF binding (primarily enhancer-
1707 promoter interactions). There is no significant difference in strength of interaction between mESC-
1708 unique intra-TAD loops (+-) and those not predicted in our model (---). Both of these groups are still
1709 stronger than CNC-mediated interactions (‘+--’ or ‘---’ vs CNC), as measured by the number of PETs
1710 supporting these interactions.
1711 D-F. Screenshots of intra-TAD loops predicted in mESC and NPC cells are shown below high-resolution Hi-
1712 C data for each cell type. These data provide support for both tissue-specific and tissue-ubiquitous intra-
1713 TAD loops. The same genomic region is shown on the left and on the right of each panel.

- 1714 D. Intra-TAD loops shared between mESC and NPC cells on mouse chromosome 17. Shown are four
 1715 shared intra-TADs loops, whose anchor-to-anchor interactions are apparent from the Hi-C data in both
 1716 mESC and NPC cells (blue arrowheads). Three of the upstream loops are contained within a weaker
 1717 TAD loop, which can be seen in both cell types (orange arrowhead).
- 1718 E. Intra-TAD loop on chromosome 1 that is predicted only in NPCs. This data supports the model that a
 1719 minority of intra-TAD loops are tissue-specific, as this genomic region shows an intra-TAD loop that was
 1720 only predicted in NPCs, with an interaction only seen in NPC cells, but not in mESC cells. Contact
 1721 matrix (green arrowhead).
- 1722 F. Intra-TAD loops on a segment of chromosome 1, that are predicted only in mESCs or only in NPCs.
 1723 Tissue-specific intra-TAD loops for mESCs and NPCs are observed, each of which is supported by a
 1724 corresponding enrichment in their respective Hi-C contact matrix (green arrowhead). Also shown is a
 1725 tissue-specific loop in NPCs that may represent some other type of looping event (i.e. enhancer-
 1726 promoter; purple arrowhead).

1727

1728 **Figure 2-figure supplement 4: Intra-TAD loop predictions in human cell lines GM12878 and K562**

- 1729 A. Comparison of CAC-mediated intra-TAD loop predicted in human lymphoblastoid-derived cells
 1730 (GM12878 cells) and in human chronic myelogenous leukemic cells (K562 cells). The number of loops
 1731 present after merging predictions across all replicates is shown, followed by filtering to ensure that only
 1732 loops containing a RefSeq TSS gene are retained. The overlap of each group with its respective sets of
 1733 loop domains (LD; column 4) and contact domains (CD; column 5) is shown; the percent of intra-TAD
 1734 loops that overlap loop domains or contact domains is listed first, followed by the percent of loop
 1735 domains and contact domains that overlap an intra-TAD loop (values in parentheses). The percentage
 1736 of intra-TAD loops that show CTCF ChIA-PET interactions in K562 cells is shown in the last column.
- 1737 B. The percent of loop domains (*left*) and contact domains (*middle*) that are shared between GM12878 and
 1738 K562 cells is consistently higher for intra-TAD loops predicted by our method (*right*) than for loop
 1739 domains or contact domains. It should be noted that the GM12878 cell dataset was sequenced more
 1740 deeply and with more replicates than the datasets for K562 cells, which is likely why GM12878 cells
 1741 show ~50% more loop domains and contact domains than K562 cells. Even comparing the percent
 1742 overlap relative to the smaller subset of loop domains/ contact domains in K562 cells, we observe
 1743 greater overlap between cell types for the intra-TAD loops predicted by our method.
- 1744 C. K562-specific loops are significantly weaker than intra-TAD loops predicted in both K562 and GM12878
 1745 cells. To compare the relative strength of loops predicted by our method, we divided CTCF ChIA-PET
 1746 loops from K562 cells (ENCODE: ENCSR436IAJ) into 3 groups: '++', meaning CTCF ChIA-PET
 1747 interactions that overlap with intra-TAD loops predicted in both K562 and GM12878 cells; '+-', meaning
 1748 CTCF ChIA-PET interactions that overlap intra-TAD loops predicted in K562 cells but not in GM12878
 1749 cells; or '--', meaning CTCF ChIA-PET interactions that do not overlap any intra-TAD loop in K562 cells.
 1750 Similar to the results for mouse (Figure 2-figure supplement 3C), the shared intra-TAD loops (++) show
 1751 significantly higher interaction strength than either the K562-specific or other CTCF loops. Further, the
 1752 K562-specific loops are stronger than other CTCF loops that do not overlap intra-TADs.

1753 D-F. Screenshots of intra-TAD loops predicted in K562 and GM12878 cells are shown below high-resolution
1754 Hi-C data for each cell line. The same genomic region is shown on the left and on the right of each panel.
1755 Shown below each gene track are stacked H3K27ac ChIP-seq tracks for the tier 1 ENCODE cell lines,
1756 including K562 (dark blue) and GM12878 cells (red).

1757 D. Intra-TAD loops shared between K562 and GM12878 cells on human chromosome 2. Shown are
1758 nested intra-TADs loops whose anchor-to-anchor interactions are apparent in both K562 and GM12878
1759 cells (blue arrowheads).

1760 E. Intra-TAD loop on human chromosome 1 that is predicted only in GM12878 cells. This supports the
1761 model that a minority of intra-TAD loops are tissue-specific, as this region shows an intra-TAD loop
1762 predicted only in GM12878 cells, and with an interaction only observed in the GM12878, but not the
1763 K562 cell contact matrix (green arrowhead). Other GM12878-specific interactions are observable within
1764 the intra-TAD loops between the promoter of MIR181A1HG and upstream GM12878-specific enhancers
1765 (purple arrowhead; red H3K27ac track).

1766 F. Intra-TAD loops on human chromosome 1 that are either predicted only in K562 cells, or are shared
1767 between K562 and GM12878 cells. Shown upstream is a K562-specific nesting (green arrowhead)
1768 within a larger shared intra-TAD loop (blue arrowhead). Shown downstream are additional examples of
1769 tissue-specific intra-TAD loops present in K562 cells, which are supported by corresponding
1770 enrichments in the K562 Hi-C contact matrices.

1771

1772 **Figure 3-figure supplement 1: Direct evidence of insulation from the asymmetric read distributions for**
1773 **virtual 4C viewpoints anchored at intra-TAD loop anchors or adjacent upstream regions**

1774 A. Shown is the distribution of Hi-C sequence reads for three intra-TADs referred to in the main text, which
1775 provides evidence of directional interactions and insulation in mESCs. For all plots, mESC Hi-C data
1776 from Bonev et. al 2017 [51] was used, plotted for a 500 kb window centered on the viewpoint (250 kb on
1777 either side; viewpoint shown as a red vertical line). Viewpoints were selected for the start of an intra-
1778 TAD loop ("IN") and for an adjacent upstream control region that does not overlap an intra-TAD loop
1779 anchor. The gene name in the title is followed by three '+' or '-' signs to indicate the presence or
1780 absence of a given intra-TAD loop in liver, mESC, or NPC, respectively. The percentages represent the
1781 proportion of reads either upstream or downstream of the viewpoint divided by the total Hi-C reads
1782 within the 500 kb region displayed. The orange box indicates the width of the intra-TAD loop
1783 interrogated. Chromosomal coordinates are shown for each "IN" viewpoint using the mm10 genome
1784 assembly, as used in [51].

1785 B. The distribution of Hi-C reads for two additional intra-TAD loops, formatted as above.

1786

1787 **Figure 3-figure supplement 2: Additional features of intra-TADs and their insulation**

1788 A. TAD anchors are more proximal to divergently-oriented upstream loop anchors, which may explain their
1789 elevated local insulation of chromatin contacts and the more diffuse bidirectional contact enrichment
1790 shown in Figure 3C. Numbers to the right of each boxplot indicate the median genomic distance to the
1791 nearest TAD/intra-TAD anchor oriented in the opposite direction.

- 1792 B. Genes whose TSS fall within inter-TAD regions are modestly enriched for housekeeping functions
1793 (ribosome, nucleosome, and mitosis) compared to a neighboring gene set whose TSS is found just on
1794 the inside (within 20 kb) of the TAD anchors. The distance cutoff of 20 kb was selected to match the test
1795 set of 1,081 inter-TAD genes to a control set of 1,132 intra-TAD genes. Functional annotation clustering
1796 was performed using DAVID with high stringency, but otherwise default parameters.
- 1797 C. Shown is an IgG control for the data presented in Figure 3D. The trend of high insulation of repressive
1798 histone marks for loop anchors and low insulation for CNC is not preserved when comparing the
1799 presumably random distribution of reads from a control IgG ChIP. This rules out other artefactual
1800 phenomena, such as mappability or random antibody pulldown, as contributing factors. The overall
1801 insulation scores are higher than indicated in Figure 3D for all groups, which indicates less insulation of
1802 IgG ChIP signal compared to H3K27me3.
- 1803 D. Shown are IgG controls for the data presented in Figure 3E and 3F. The genomic regions as analyzed
1804 in Figures 3E and 3F do not show an uneven distribution of IgG ChIP reads, which might be expected in
1805 the event of mapping differences or other alternative explanations.
- 1806

1807 **Figure 4-figure supplement 1: Characteristics of five classes of DHS in mouse liver**

- 1808 A. Schematic for the classification of open chromatin regions (DHS) in mouse liver based on their
1809 H3K4me1, H3K4me3, and CTCF ChIP-seq signals. DHS with >4 rpm for either H3K4me3 or H3K4me1
1810 were classified based on the ratio of H3K4me3 to H3K4me1 signal, where a high ratio indicates a
1811 promoter-like DHS, and a low ratio indicates an enhancer-like DHS regions. For the 39,410 liver DHS
1812 with low signal for both histone marks, the H3K4me1 signal was compared to the CTCF signal to
1813 classify each DHS. Regions with higher CTCF signal that overlapped a CTCF peak were classified as
1814 Insulator-DHS and the remaining group were classified as weak enhancer-DHS (also see
1815 Supplementary File 2A).
- 1816 B. The distance from each of the five DHS classes to the nearest RefSeq TSS. Regions with roughly equal
1817 H3K4me3 and H3K4me1 signal were classified as weak promoters based on their proximity to DHS.
1818 Similarly, the low signal group (in gray) were classified as weak enhancers based on distance from TSS
1819 and low signal for both DNase-seq and H3K4me1 ChIP.
- 1820 C. Protein coding and lncRNA genes associated with promoter-DHS are both more highly expressed than
1821 genes associated with weak promoter-DHS or other TSS without a nearby promoter or weak promoter
1822 DHS. All groups are significantly different from one another, (KS test; $p < 0.001$).
- 1823 D. Tissue specificity of DHS identified in mouse liver. Shown is the fraction of liver DHS that overlap DHS
1824 identified in the indicated number of other mouse tissues. Insulator DHS are unique among distal DHS
1825 for having a subset that is tissue ubiquitous. Both promoter classes show a significant fraction of sites
1826 that are open across all mouse tissues, while the two enhancer classes are much more tissue specific.
1827 A value of 0 indicates that the DHS is liver specific, while a value of 20 indicates that it is found in all 20
1828 mouse tissues.
- 1829

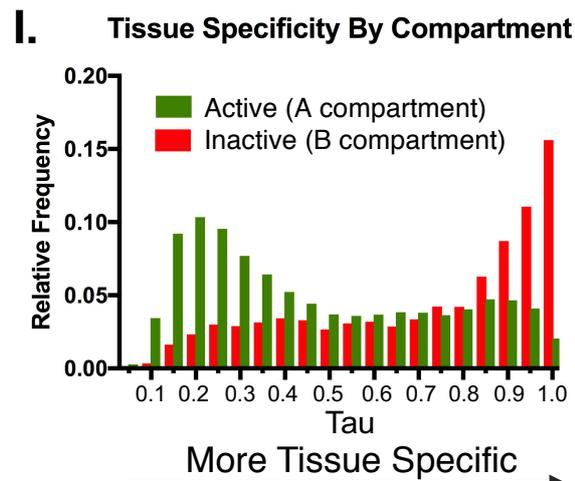
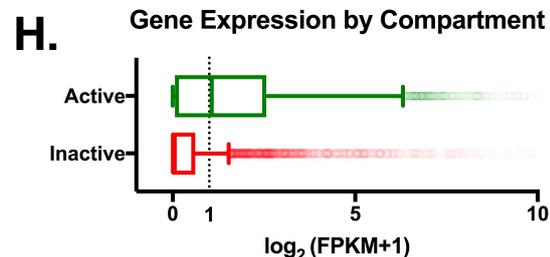
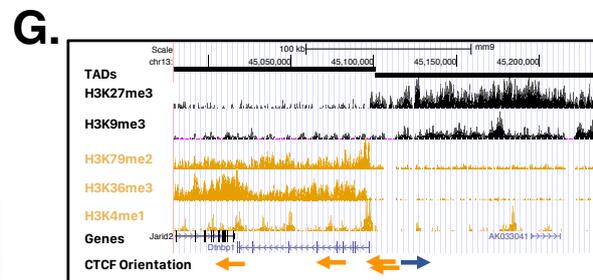
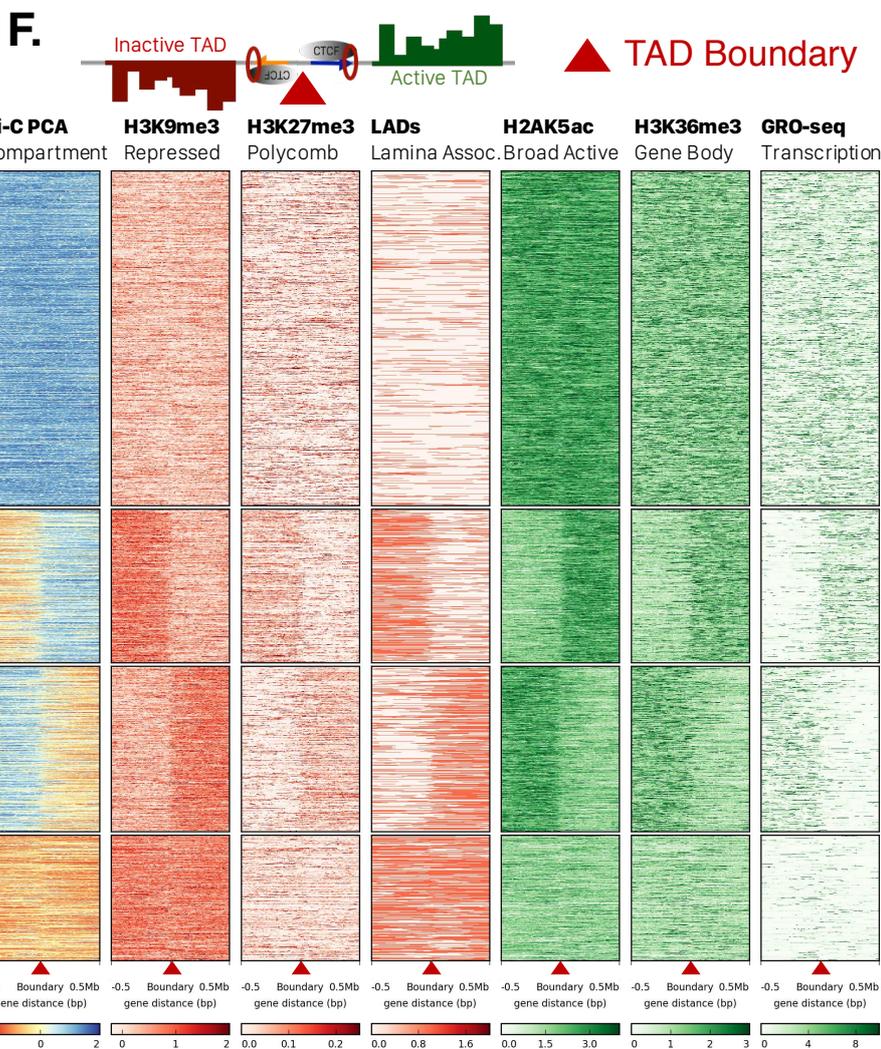
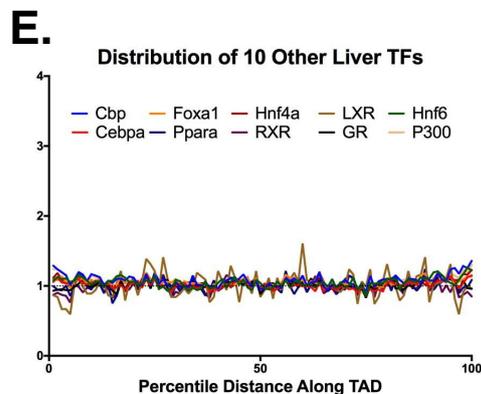
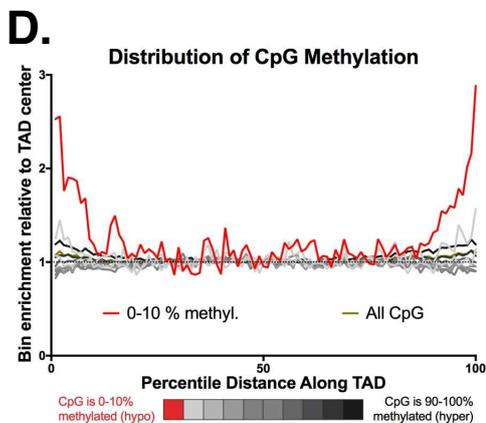
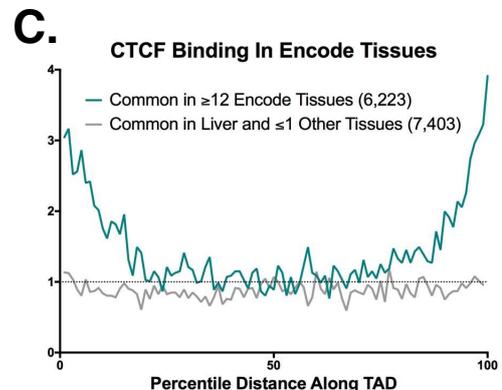
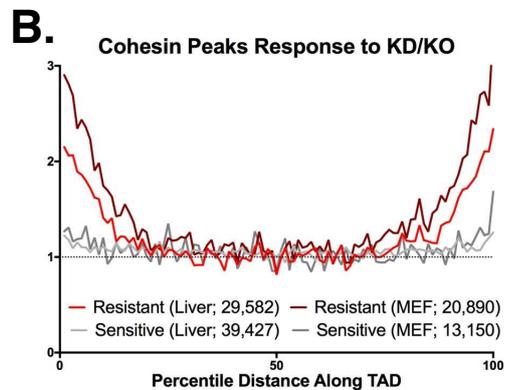
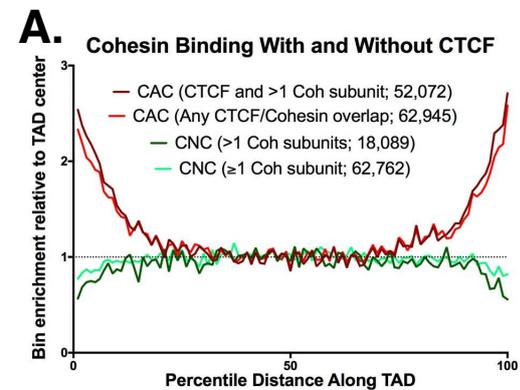
1830 **Figure 4-figure supplement 2: Features of super-enhancers and single-TSS intra-TAD loops**

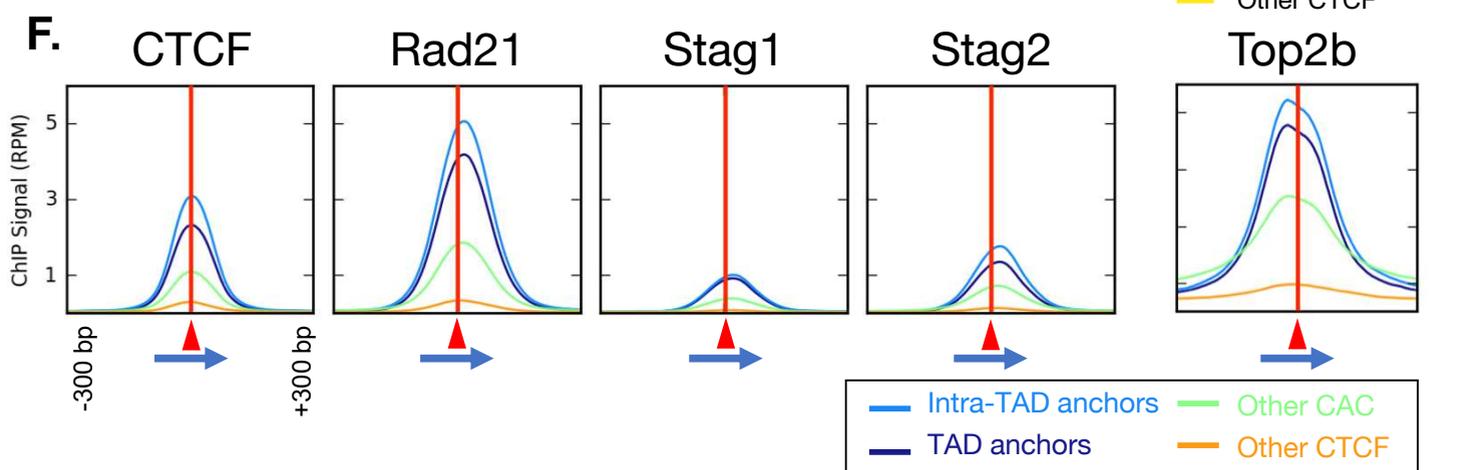
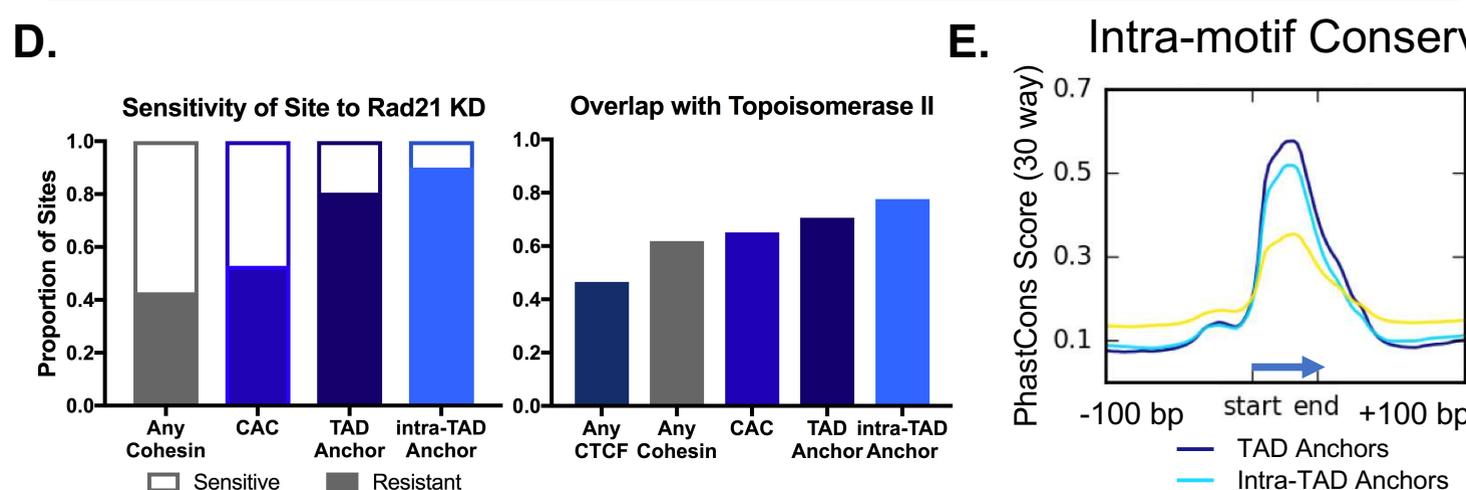
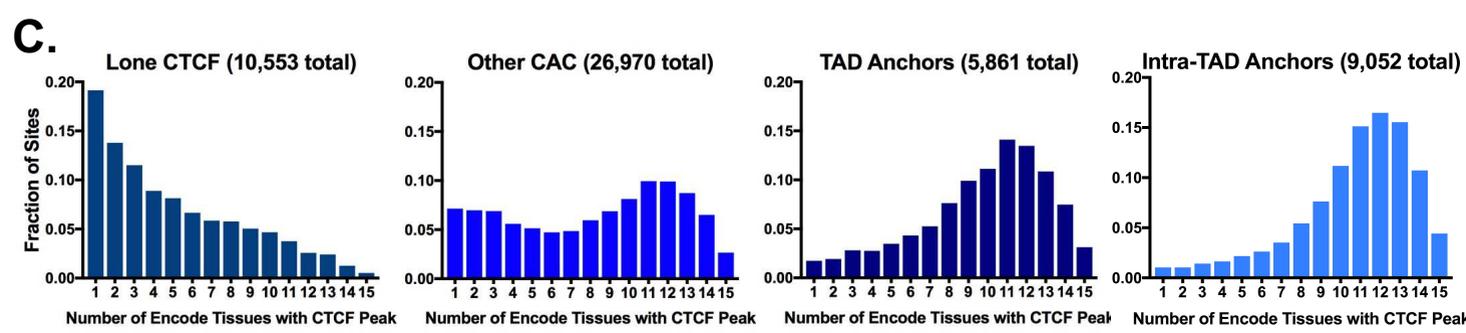
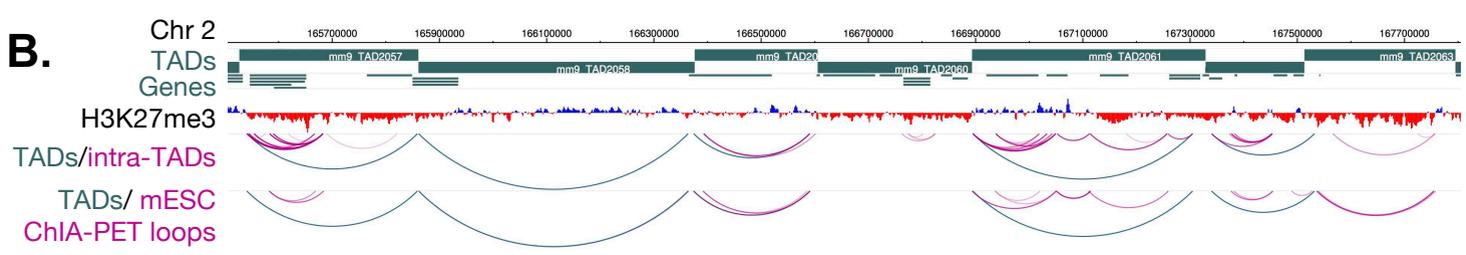
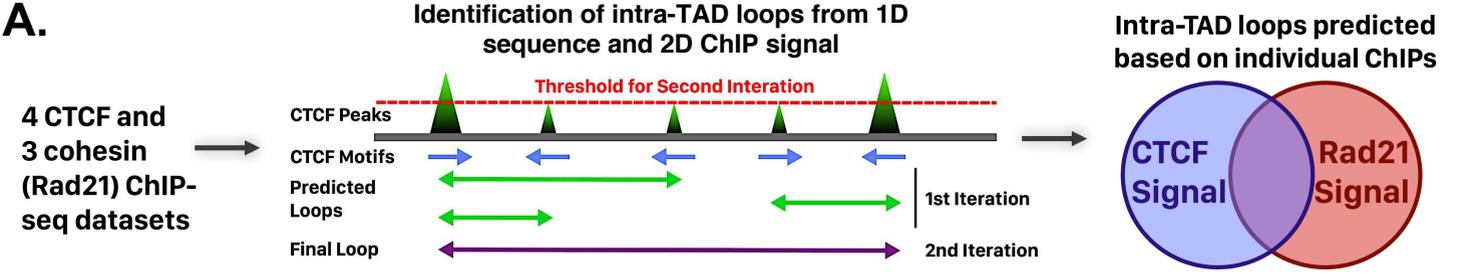
- 1831 A. Shown is the overlap of super-enhancers (SE) for all 19 liver H3K27ac ChIP-seq replicates (see
1832 Methods). 503 of the super-enhancers were identified in all 19 liver samples, and are termed core
1833 (robust) super-enhancers. The x-axis indicates the number of liver samples in which an enhancer
1834 cluster exceeded the threshold to be identified as a super-enhancer.
- 1835 B. Super-enhancer constituents show higher levels of eRNA production than typical enhancers. This is
1836 consistent with Figure 4F, which shows greater accumulation of Pol2 but not H3K4me3 than at typical
1837 enhancers, suggesting non-promoter transcription.
- 1838 C. Of the 9,543 intra-TAD loops, 3,142 (33%) contain a single TSS. This is a higher frequency of single
1839 TSS intra-TAD loops than is expected by random chance (expected (median) = 1,802 intra-TAD loops
1840 across 10,000 shuffled sets). The matched set represents a random list of genomic coordinates of the
1841 same size and number.
- 1842 D. For the set of genes that are in single TSS intra-TAD loops, gene ontology analysis followed by
1843 functional annotation clustering revealed an enrichment for transcriptional regulation and kinase activity.
1844 Functional annotation clustering was performed using DAVID with high stringency for cluster
1845 assignment, but otherwise default parameters.
- 1846

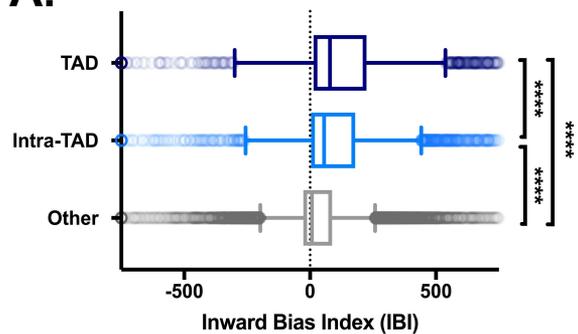
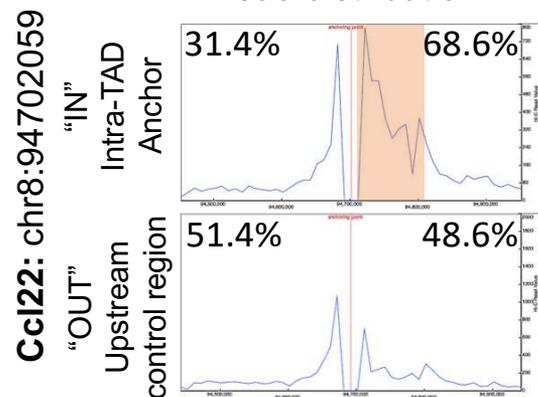
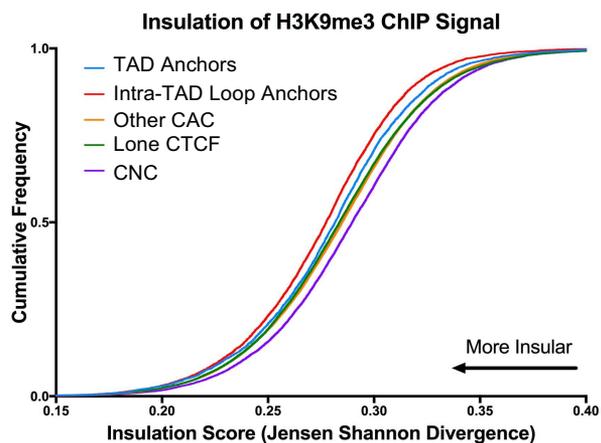
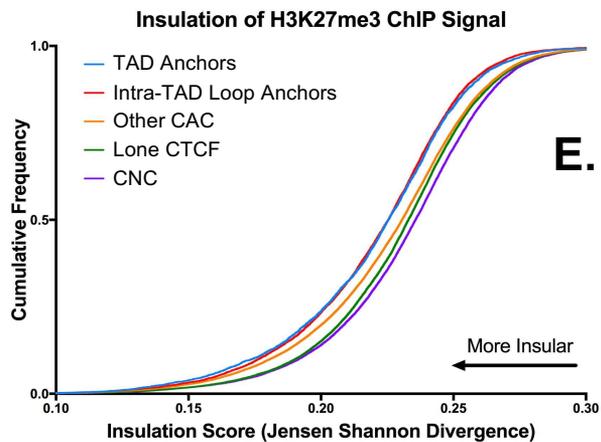
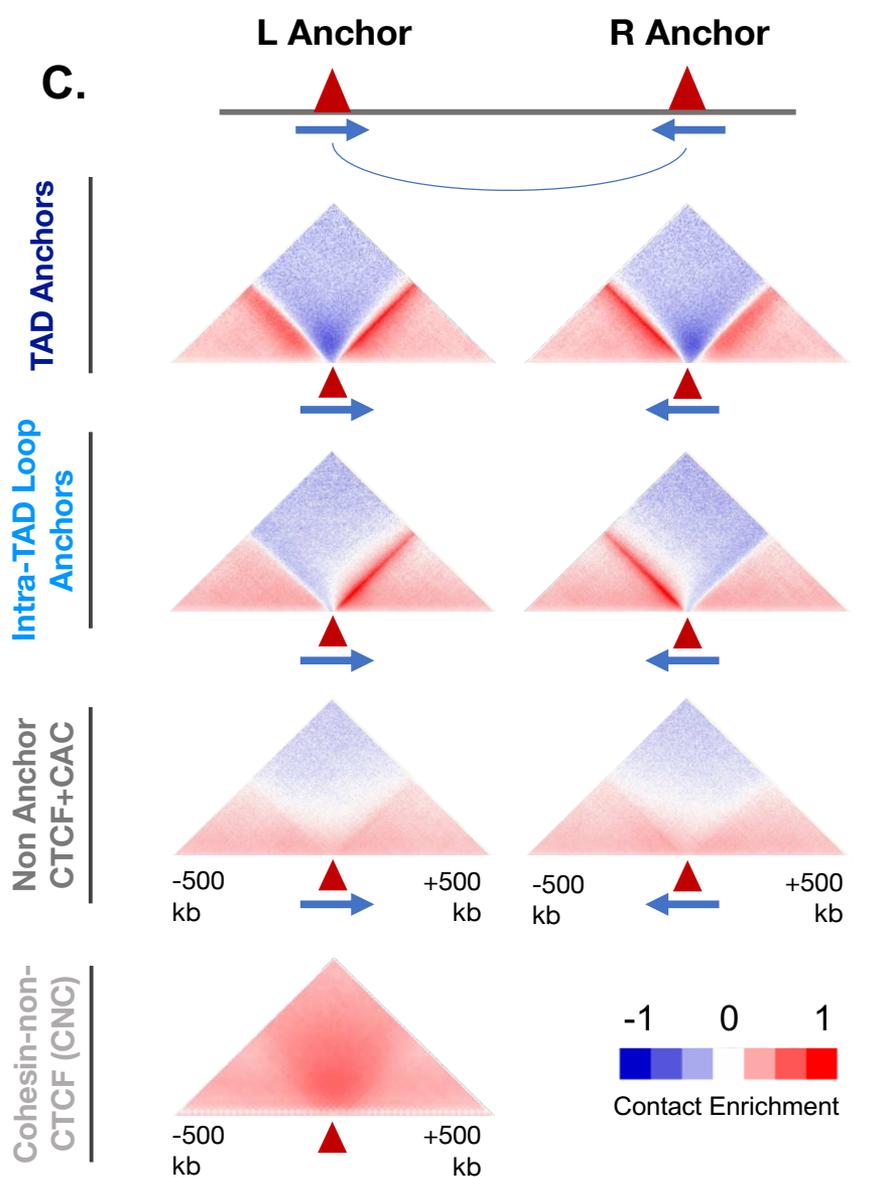
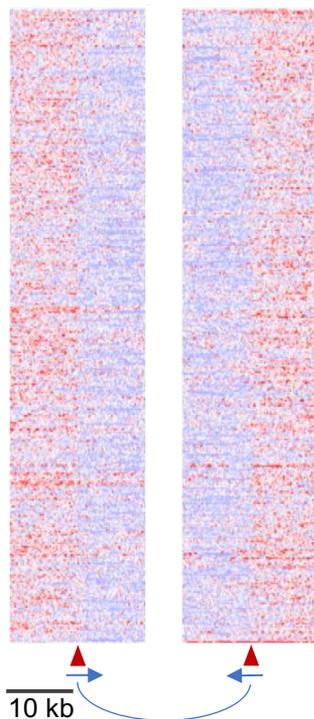
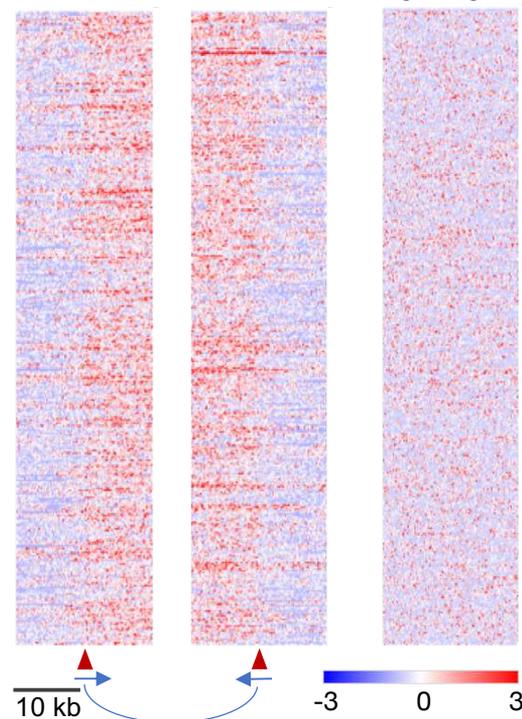
1847 **Figure 6-figure supplement 1: *Alb* 4C-seq replicates and cis/trans 4C-seq signal distribution**

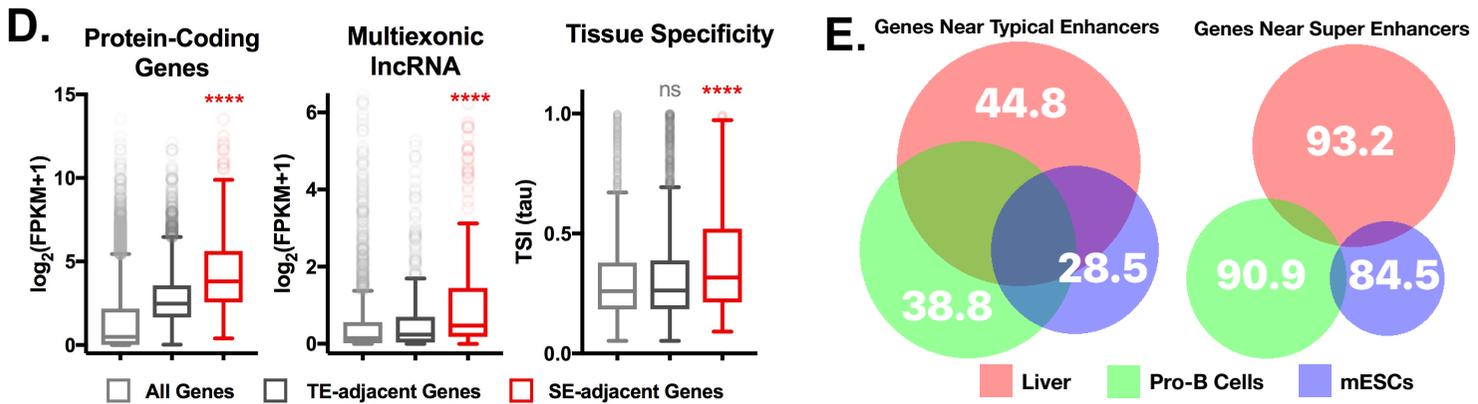
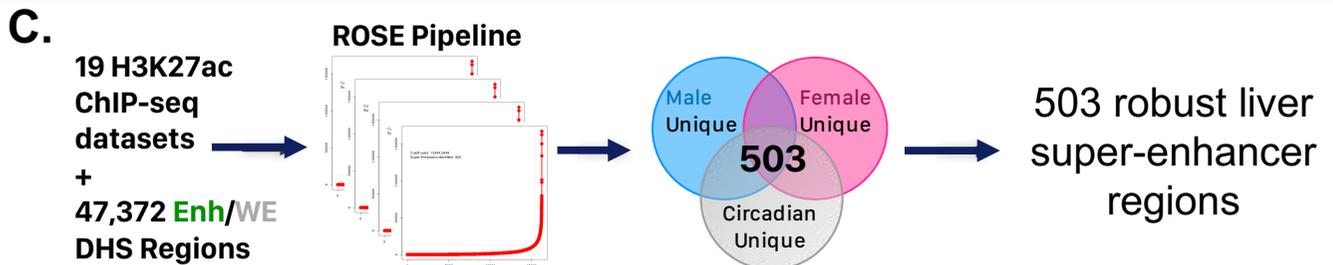
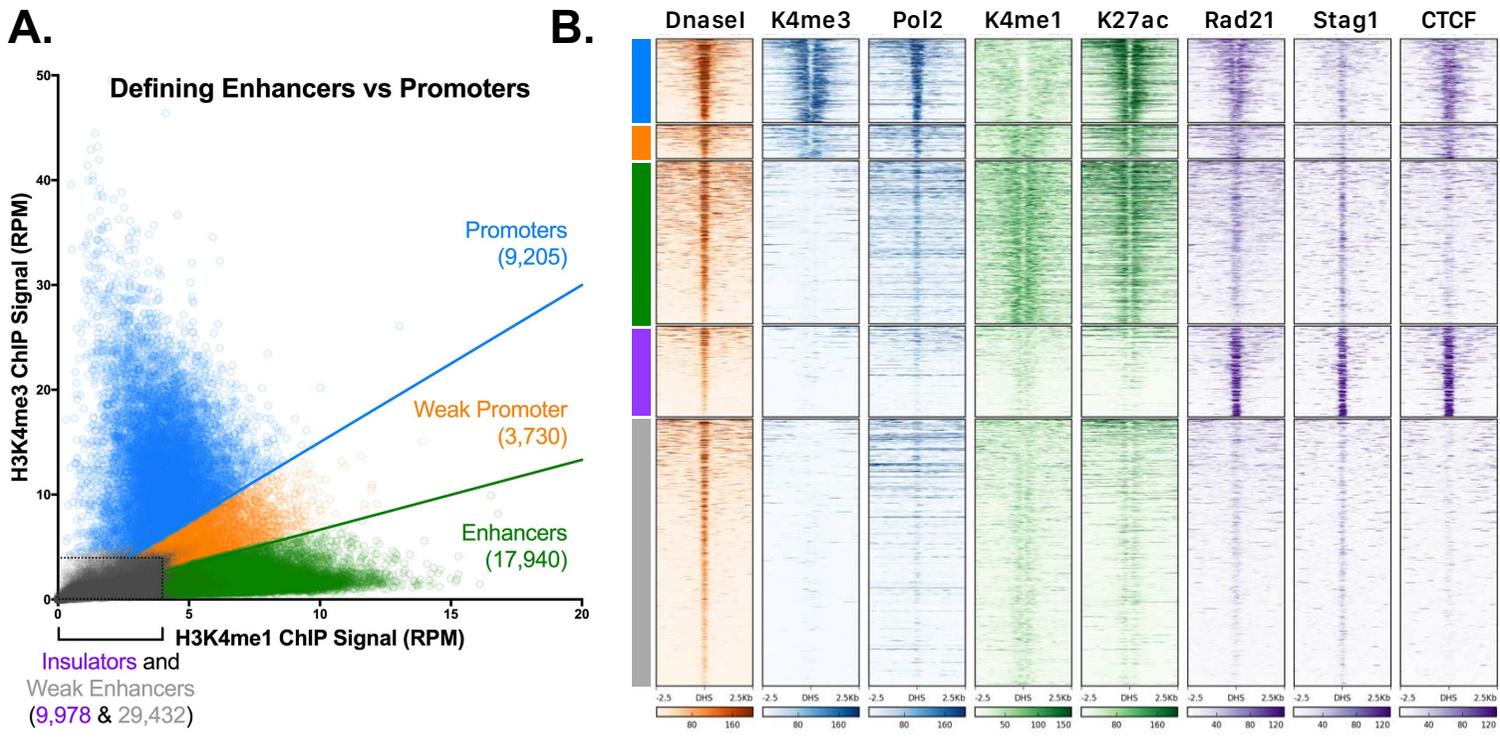
- 1848 A. UCSC genome browser screenshots for all biological replicates. The proximal view shows a 150 kb
1849 window around the *Alb* TSS, while the distal view shows a 400 kb view around the *Alb* TSS. For each
1850 sample, the colored track indicates normalized interactions for that region with the viewpoint (*Alb*
1851 promoter) in reads per million. The black bars below the colored signal track are interactions identified
1852 by the R3C-seq pipeline, expressed as a $-10\log$ of the p-value.
- 1853 B. *Alb* promoter interactions are highly reproducible, with an average R2 of 0.87 between pairs of
1854 biological replicates between individual male (M) and female (F) mouse livers, as indicated on the x-
1855 axis.
- 1856 C. Interactions originating from the *Alb* promoter viewpoint show no major apparent sex differences
1857 between male and female liver samples. This is visualized both in terms of strength of interaction ($-$
1858 $10\log(pval)$, as above) as well as normalized read depth (reads per million, below).
- 1859 D. The expression of protein-coding genes within non-, low, medium, or high interacting TADs in *cis* or
1860 *trans*, expressed as $\log_2(FPKM+1)$ values. Genes in highly interacting TADs are more highly expressed
1861 compared to non-interacting TADs or the genome wide/*cis* control ($p < 0.001$; KS t-test). This fits with the
1862 transcription factory model of compartment organization, as one would expect highly active regions to
1863 weakly associate in hubs of active transcription. Genes in TADs that are not interacting with the *Alb*
1864 promoter show the lowest expression of all groups. Shown for comparison is the expression of all genes
1865 on chromosome 5 (*cis*) and also genome-wide (*trans*).
- 1866 E. Median eigenvalues within *cis* or *trans* non-, low, medium, or high interacting TADs. A positive value
1867 indicates a more open/active chromatin compartment ("A type"), while a negative value indicates a more
1868 closed/inactive chromatin compartment ("B type"). The highest interacting TADs show higher

1869 eigenvalues than the background or non-interacting TADs ($p < 0.001$; KS test). Only *trans* non-interacting
1870 TADs show a consistent negative median eigenvalue, indicating a repressed compartment.
1871

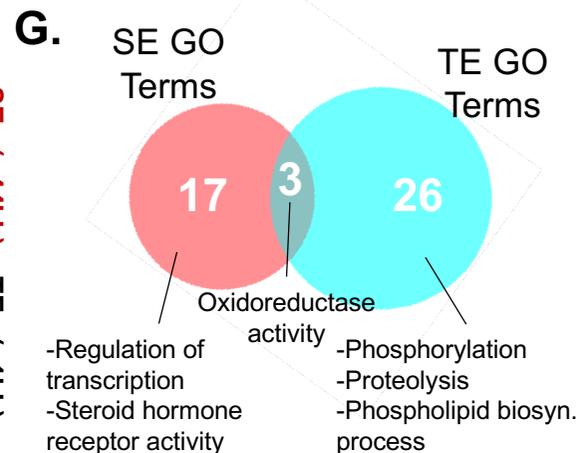
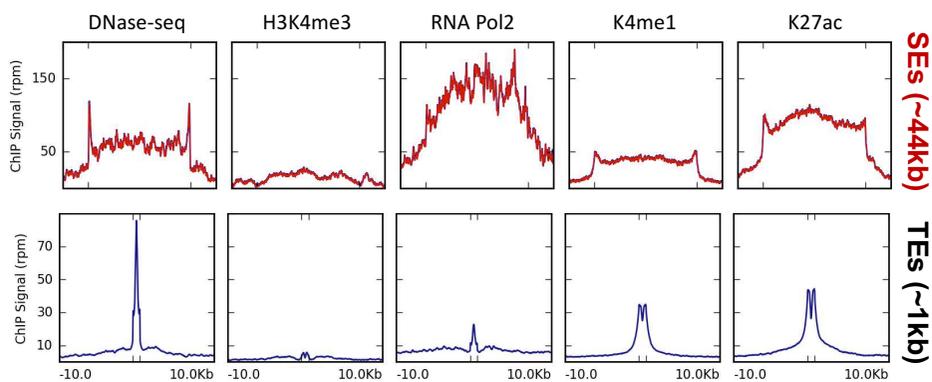




A. Directionality of Interactions from Anchors**B. Shift in Virtual 4C Read distribution****D. Insulation of H3K27me3 ChIP Signal****C.****E. Active intra-TAD Insulation (H3K27me3 signal)****F. Inactive intra-TAD Insulation (H3K27me3 signal)**

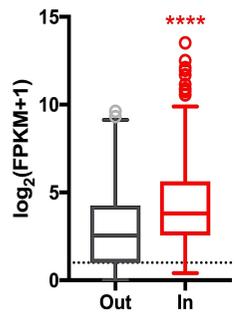


F. ChIP signal at SE/TE, scaled to median length +/- 10 kb:

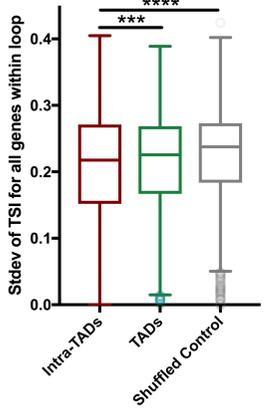




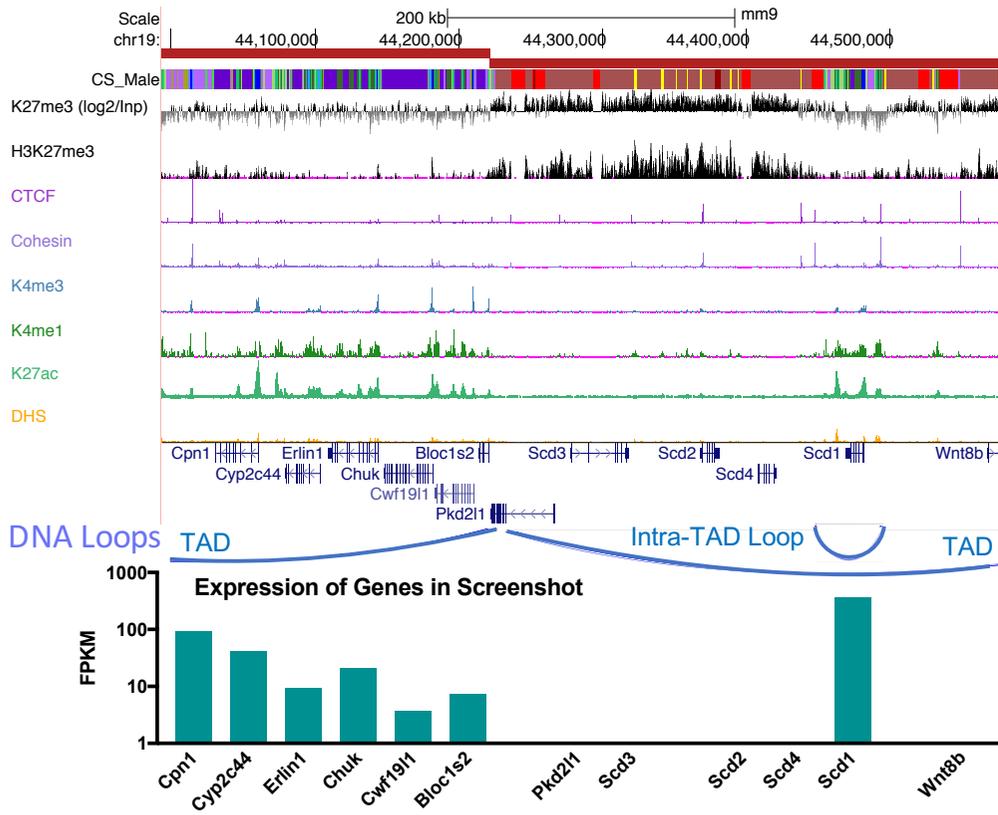
A. Nearest Gene Out of Loop vs In Loop



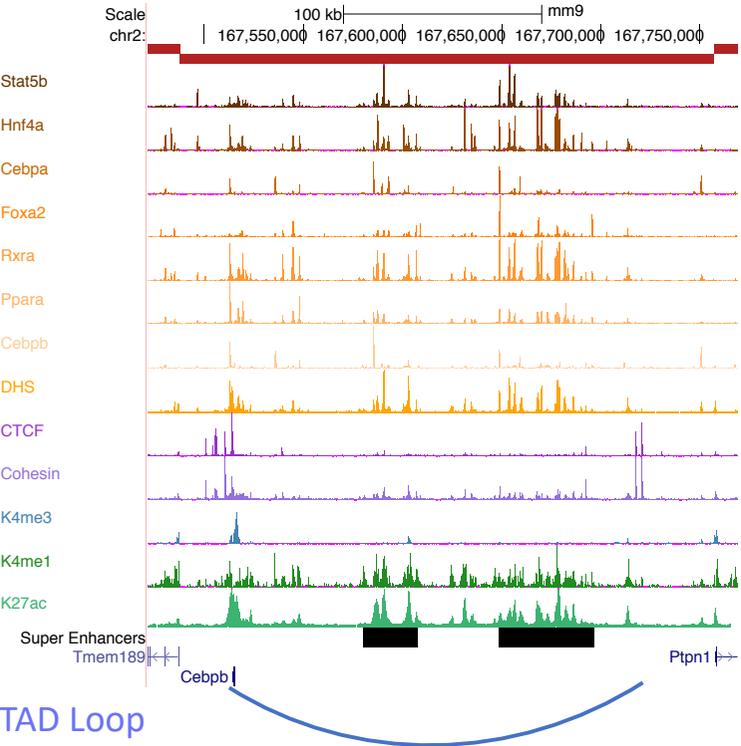
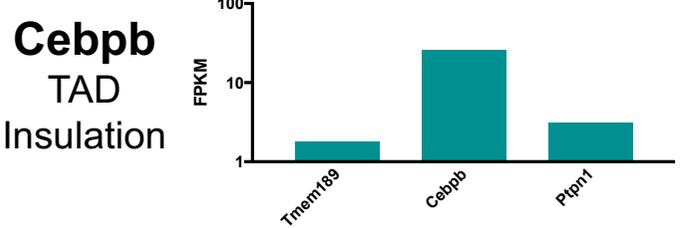
B. Genes within (intra-)TADs are more uniform in tissue-specificity



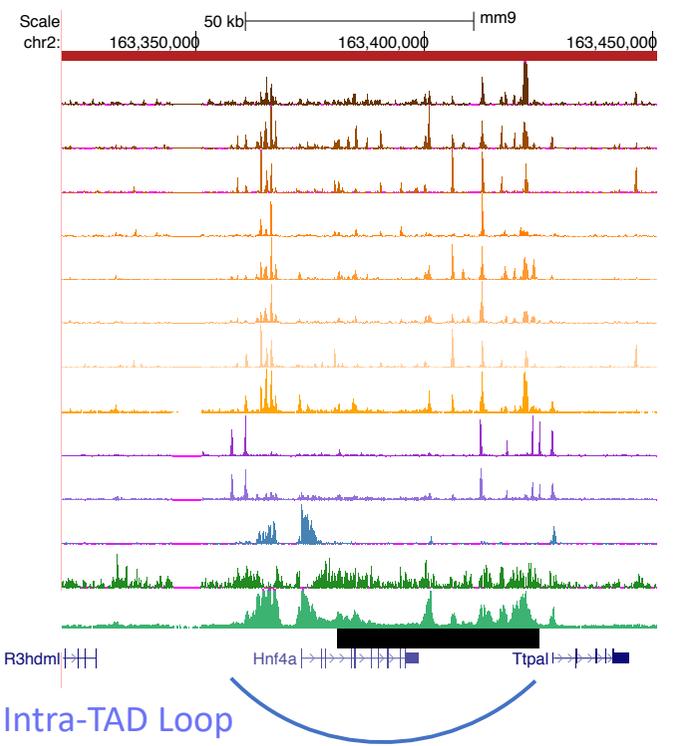
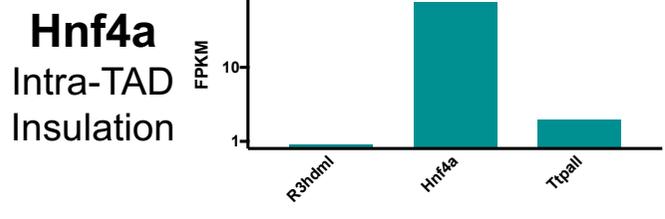
E. Scd1, High expression in B Compartment

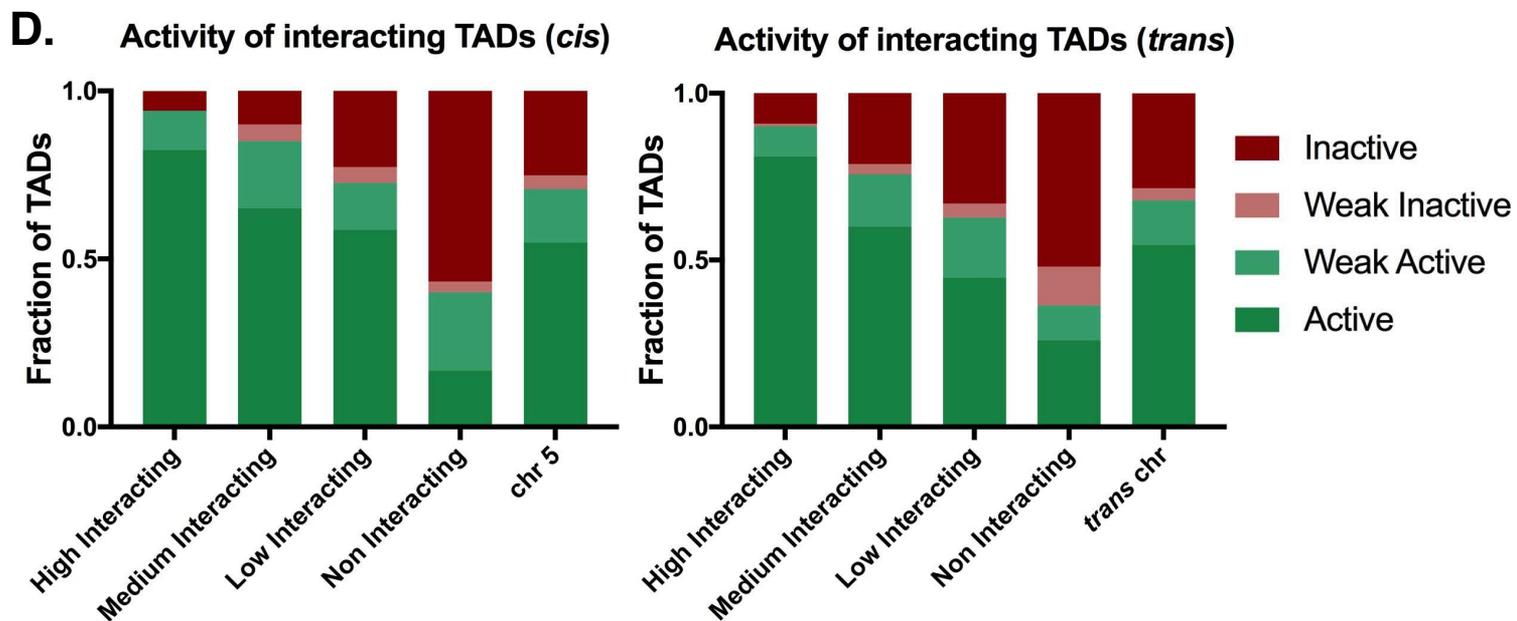
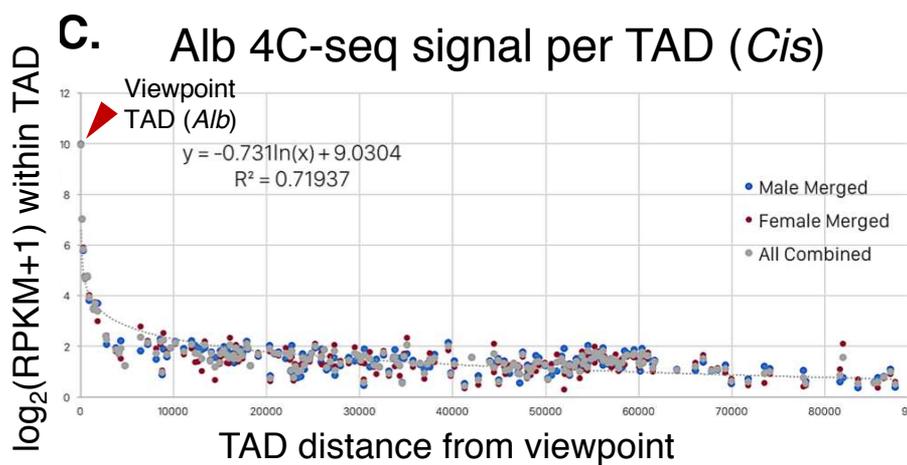
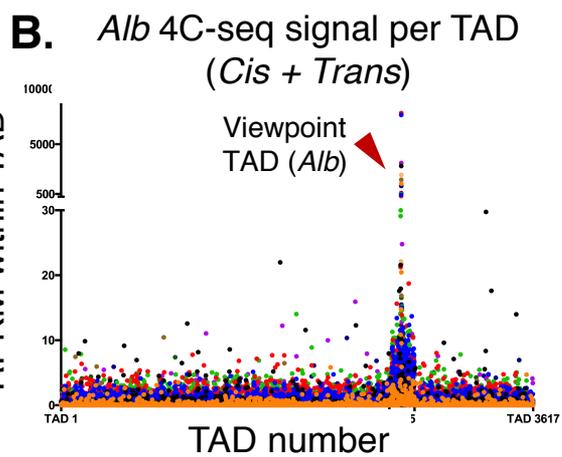
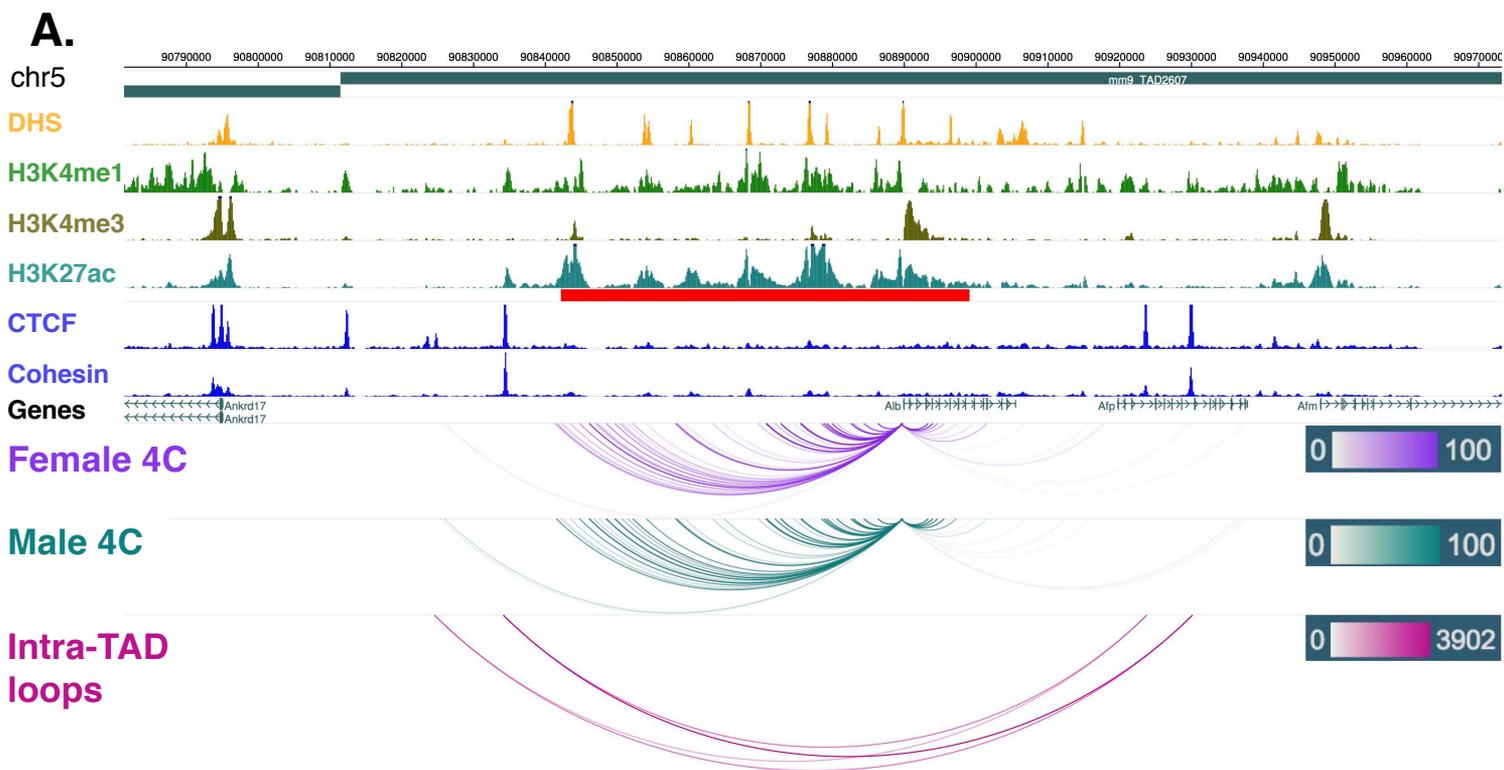


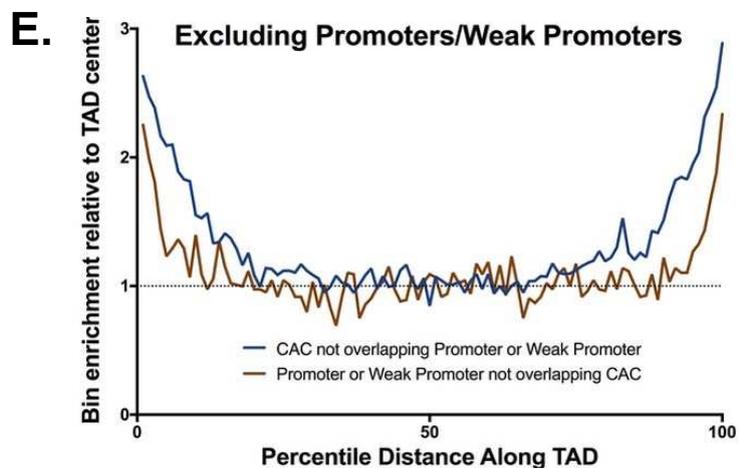
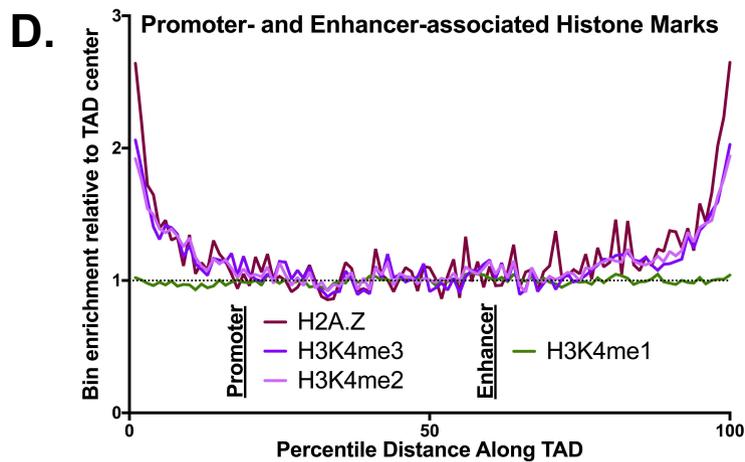
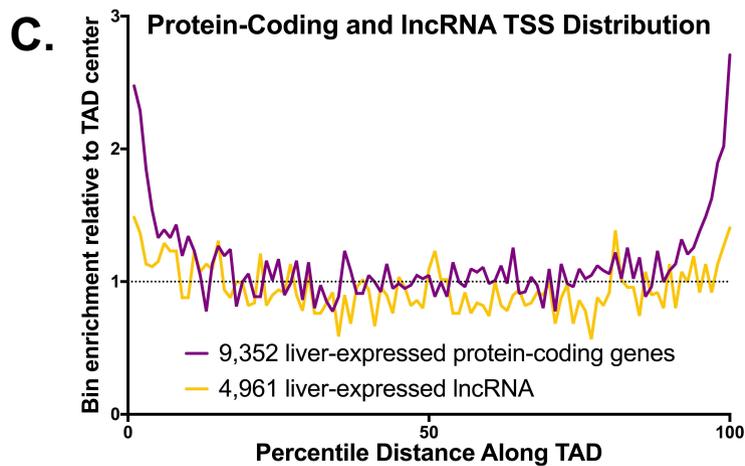
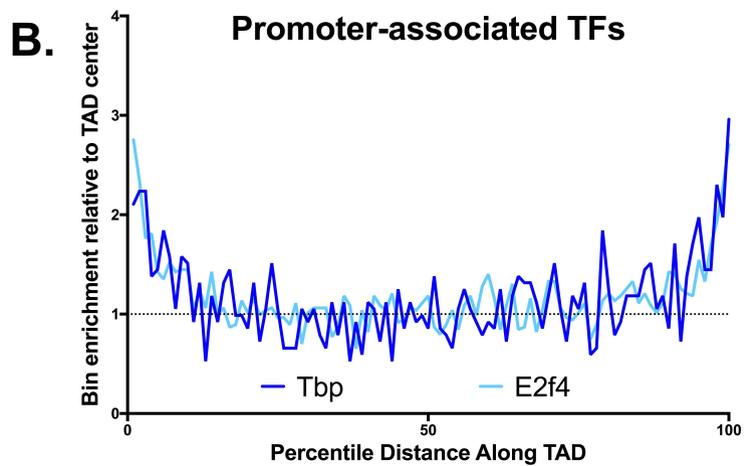
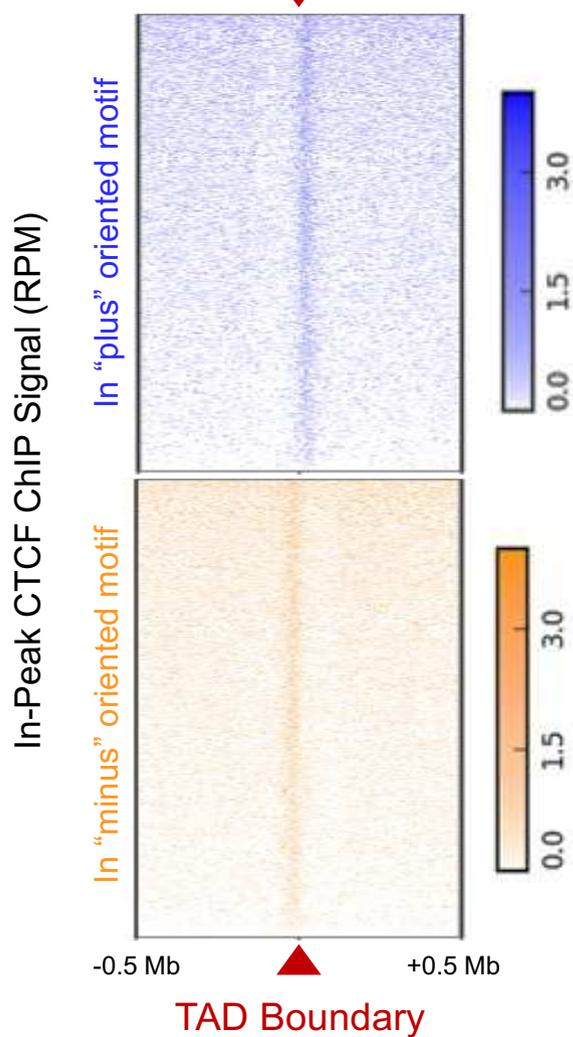
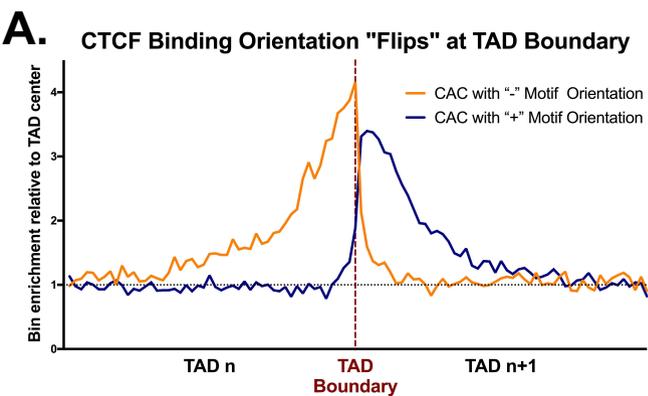
C. Cebpb TAD Insulation

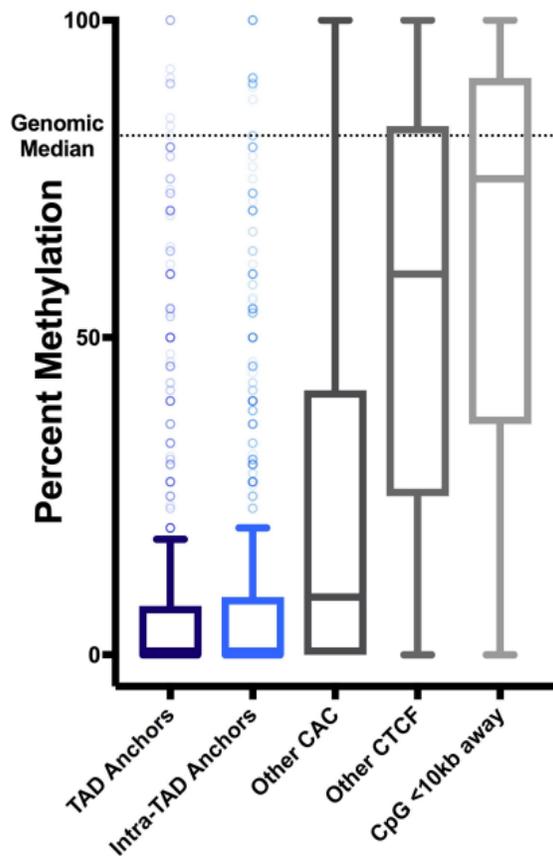
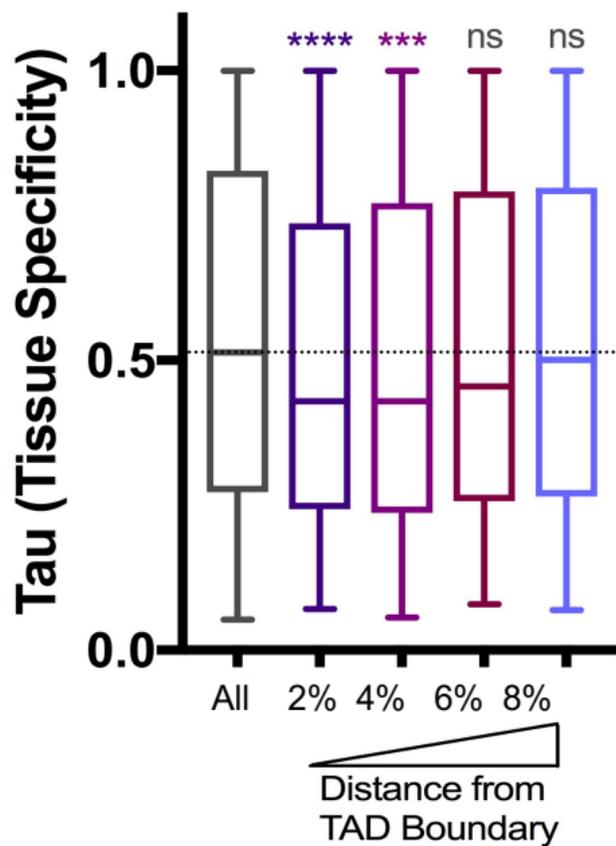


D. Hnf4a Intra-TAD Insulation

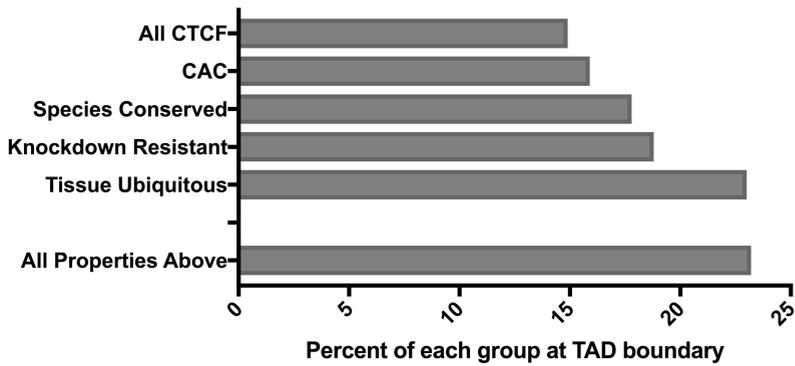






F.**CpG methylation
within CTCF motif****G.****Tissue Specificity
at TAD Boundaries**

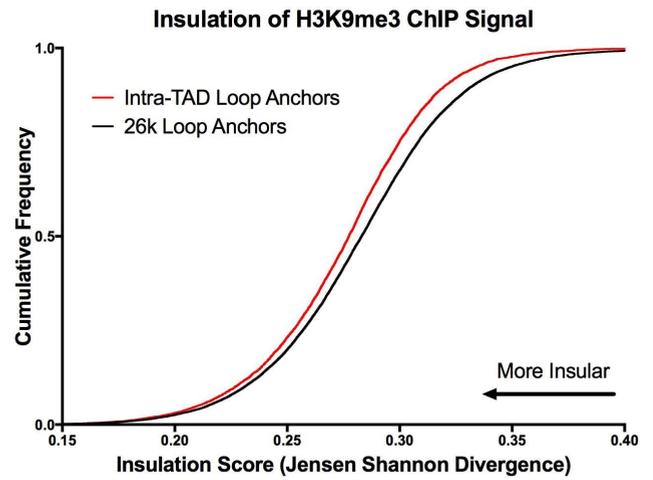
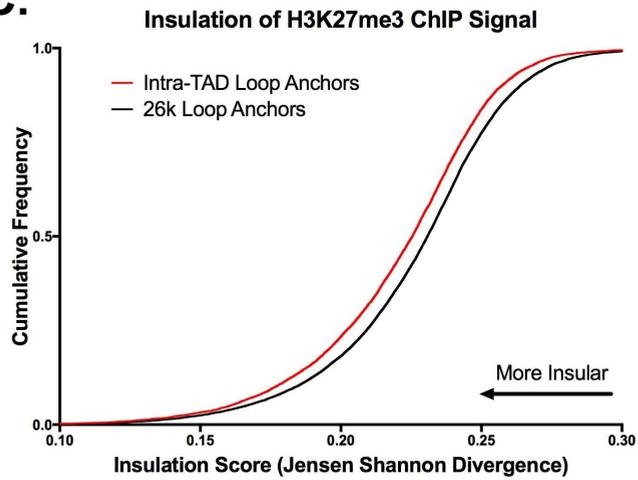
A. TAD Boundary Association as % of Group



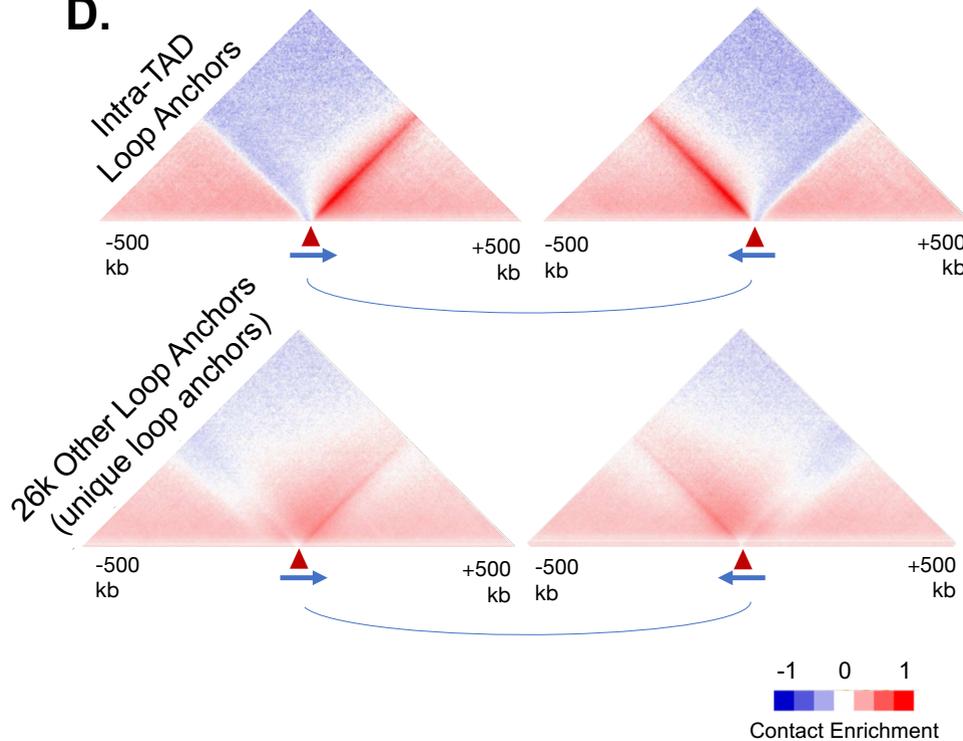
B.

	60k Loops	SubTADs
Number of Loops	60,677	9,543
Median Length (kb)	61	151
Overlap w/ intra-TAD loops (%)	59	100
Overlap w/ mESC ChIA-PET (%)	25.5	63.2
Median IBI (anchor directionality)	15.61	56.04
Median JSD K27me3 (anchor insulation)	0.2308	0.2254
Median JSD, K9me3 (anchor insulation)	0.2830	0.2774

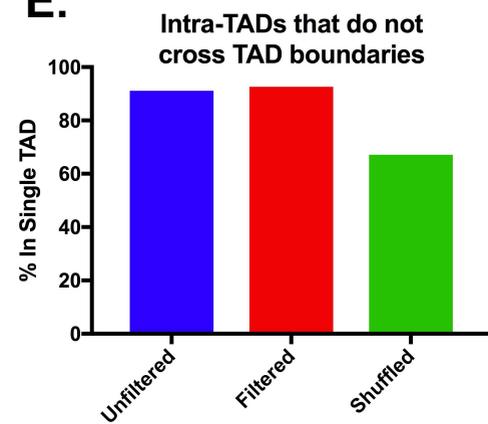
C.



D.



E.



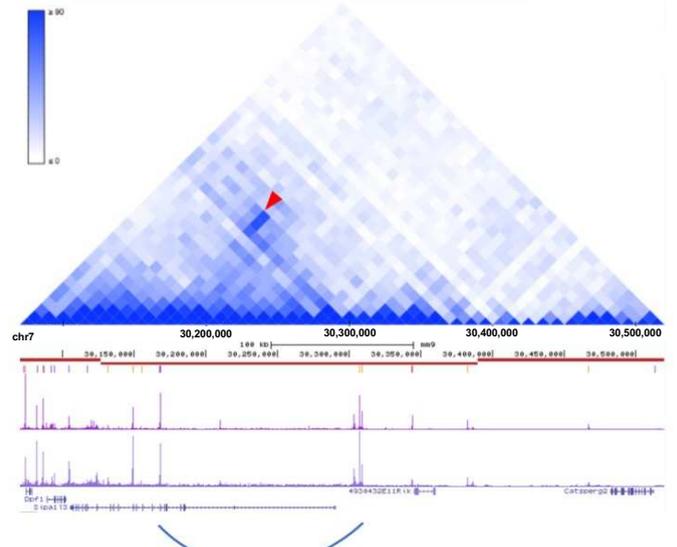
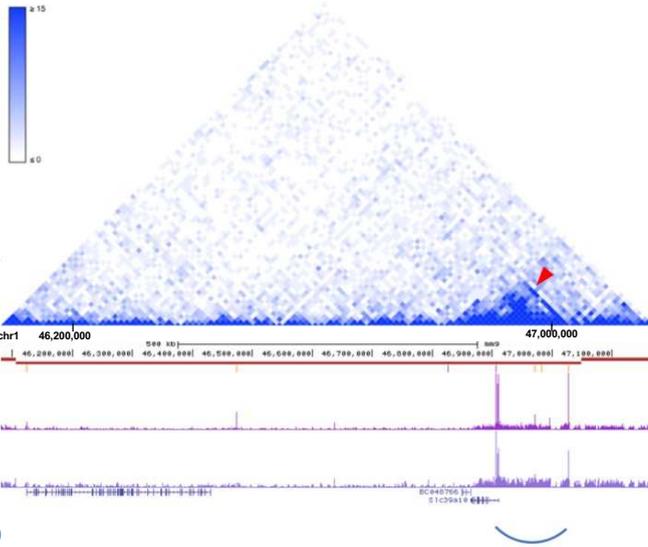
F.

CH12
High Res.
Hi-C

CTCF

Cohesin

Intra-TAD
Loops

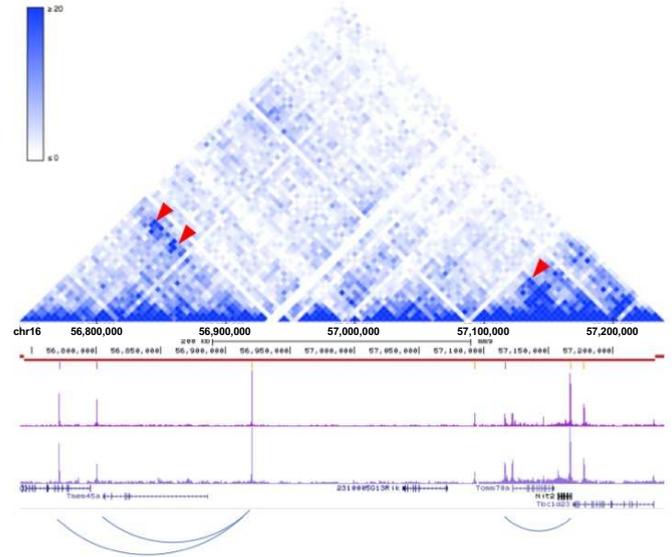
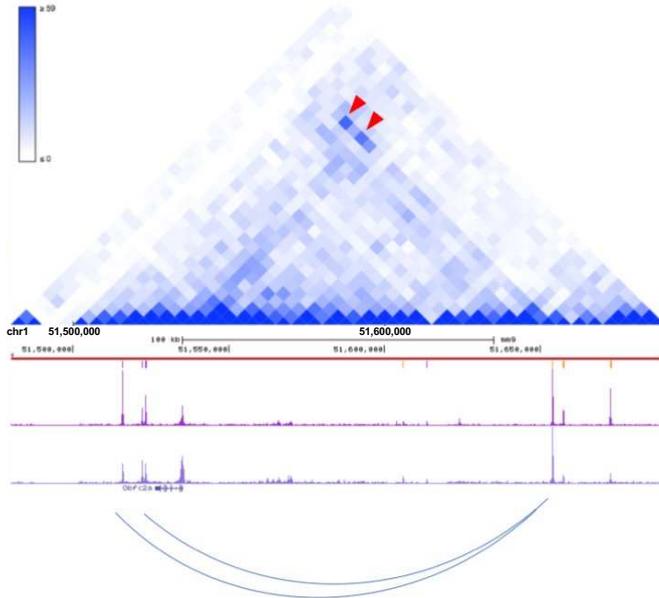
**G.**

CH12
High Res.
Hi-C

CTCF

Cohesin

Intra-TAD
Loops

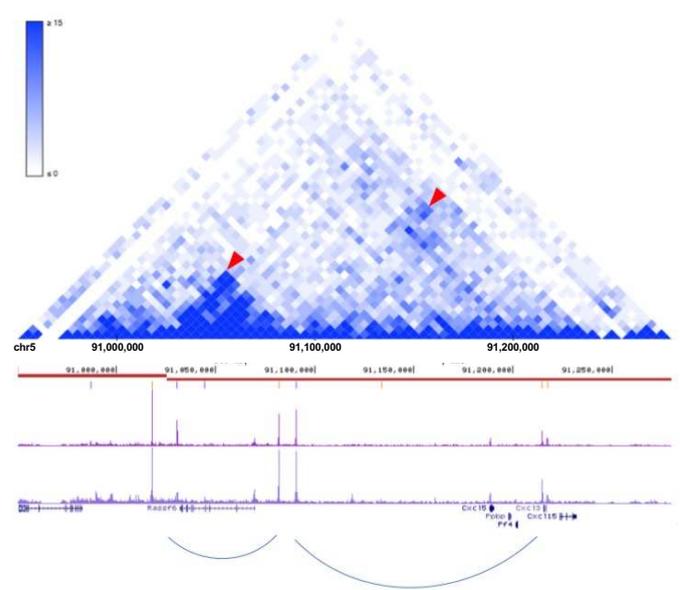
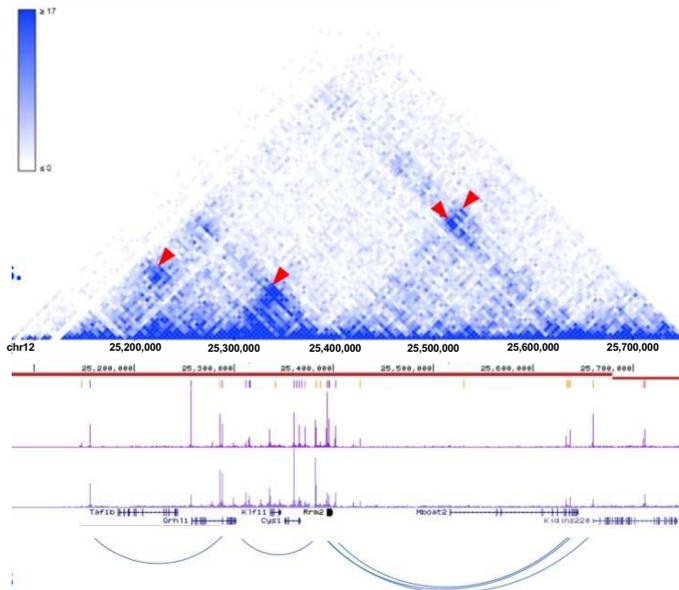
**H.**

CH12
High Res.
Hi-C

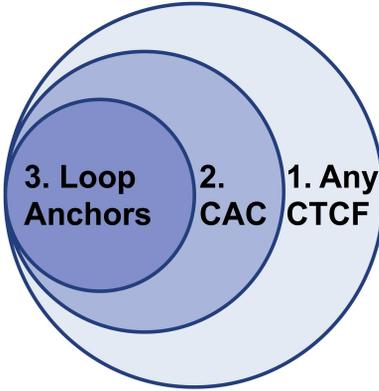
CTCF

Cohesin

Intra-TAD
Loops

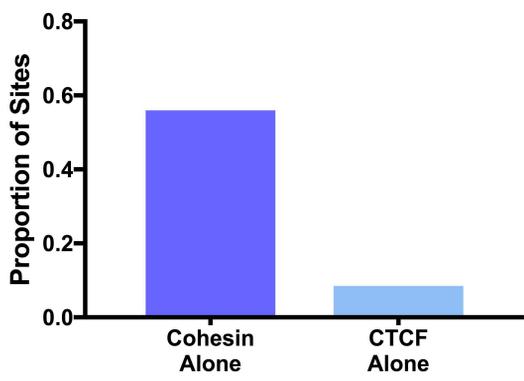


A. Subsets of CTCF Binding Sites:
*Not Size Proportional

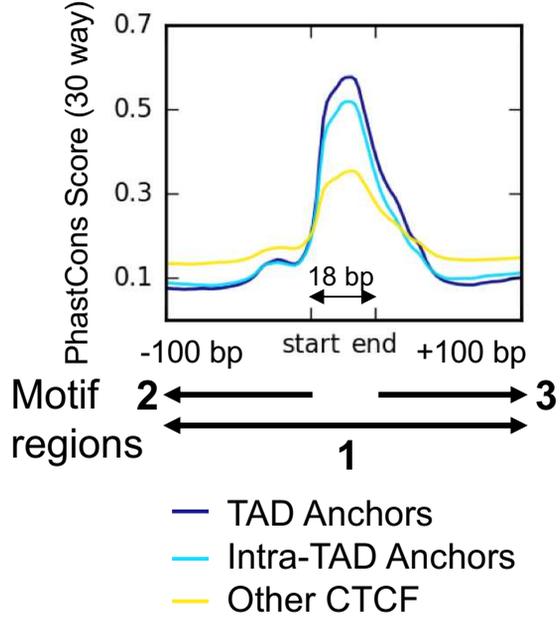


- (1) Any CTCF:
52,436
- (2) CAC:
42,801
- (3) Loop Anchors:
9,052 (intra-TAD)
and 5,861 (TAD)

B. Overlap with Topoisomerase II



C. Intra-motif Conserv.



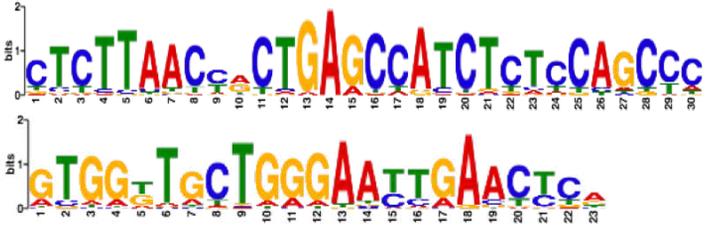
1: Summit +/- 150 bp



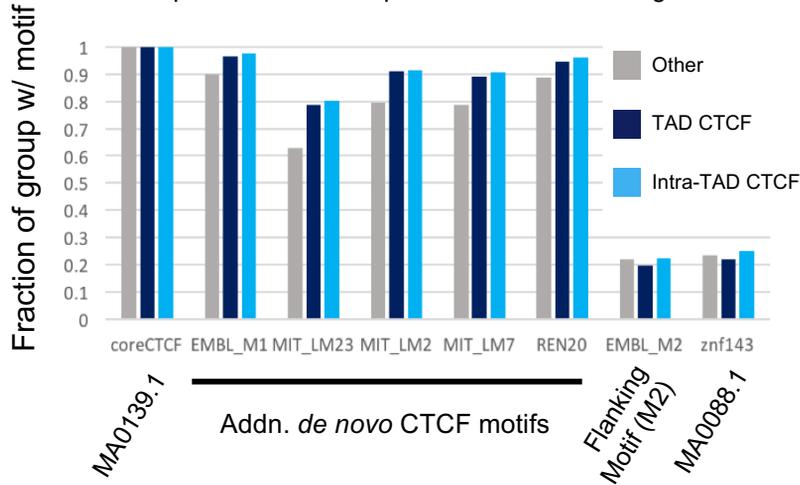
2: 150 bp outside of loop (excl. core motif)



3: 150 bp inside of loop (excl. core motif)

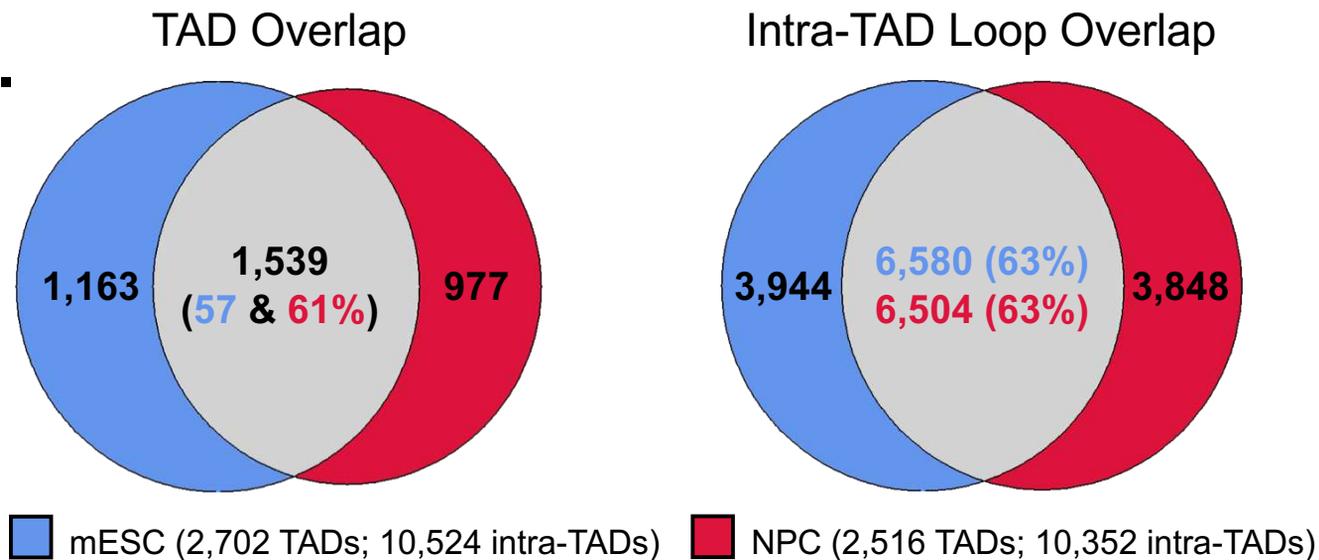


D. Loop anchors show expanded CTCF motif usage

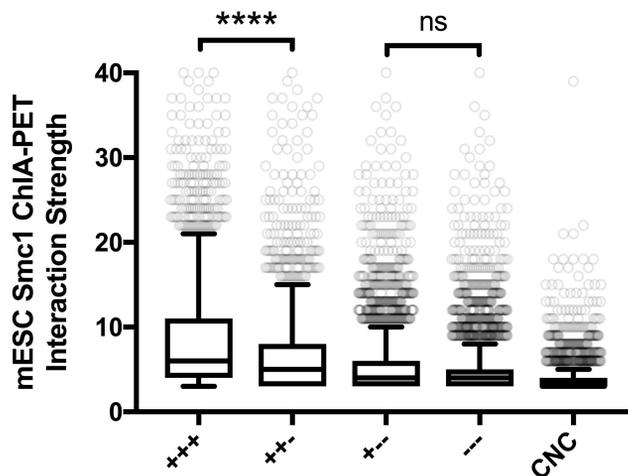


A.

Mouse (mm9)	# of merged loops	# of intra-TAD loops	# of intra-TAD loops with TSS	median size (kb)	% overlap with liver; as % of liver intraTADs (and % of NPC or mESC)	% overlap with mESC Insulated Neighborhoods (11,592 total)	% overlap with "CTCF-CTCF" mESC Smc1 ChIA-PET (13,987 total)
Liver	12,079	10,136	9,543	151	100 (100)	53.3 (34.3)	63.2
NPCs	14,025	11,276	10,352	197	62.6 (45.2)	48.6 (43.4)	57.0
mESCs	13,823	11,754	10,524	180	62.3 (45.2)	51.1 (46.3)	63.4

B.**C.**

Ubiquitous Intra-TAD Loops Are Stronger Interactions



A.

Human (hg19)	# of merged loops	# of loops with TSS	median size (kb)	% of overlap with loop domains in same cell line (and % of LDs) ¹	% of overlap with contact domains in same cell line (and % of CDs) ²	% overlap with K562 CTCF ChIA-PET Interactions (25,721 total)
GM12878	17,634	15,366	264	54.4 (51.9)	35.2 (32.7)	60.4
K562	17,452	15,130	214	39.8 (48.2)	26.1 (31.4)	64.8

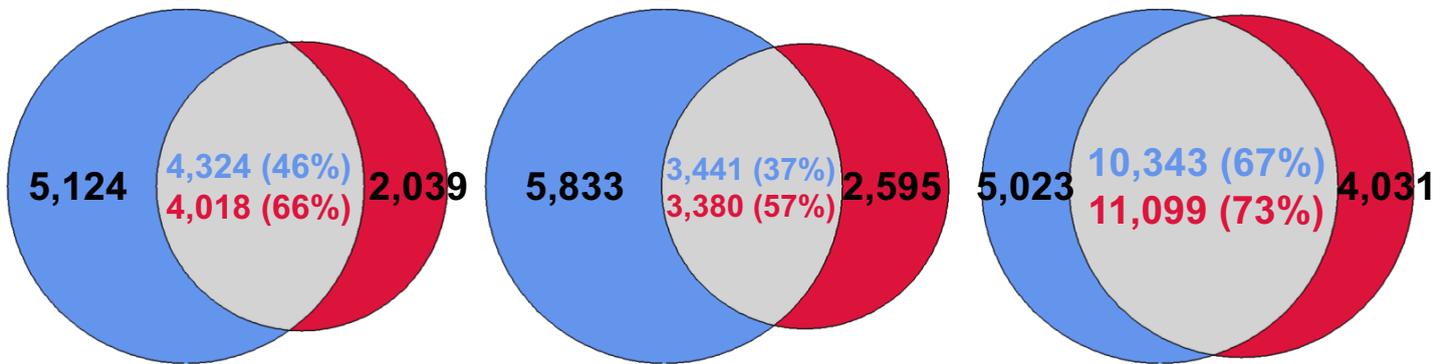
1. There are 9,448 and 6,057 loop domains in GM12878 and K562, respectively
2. There are 9,274 and 5,975 contact domains in GM12878 and K562, respectively

B.

Loop Domains

Contact Domains

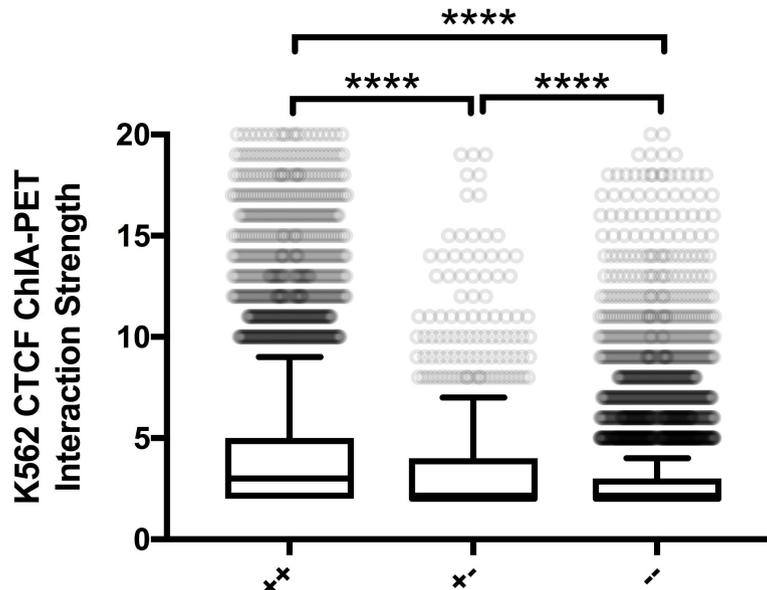
Intra-TAD Loops

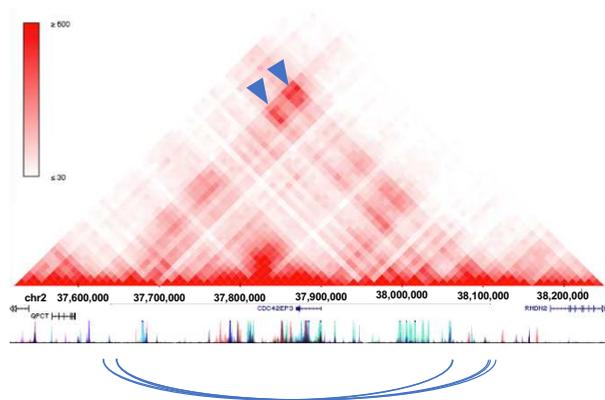
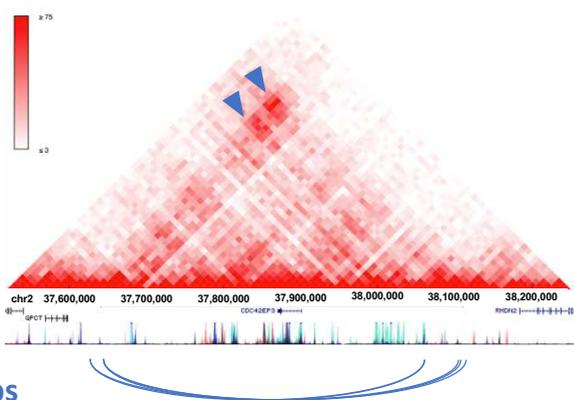
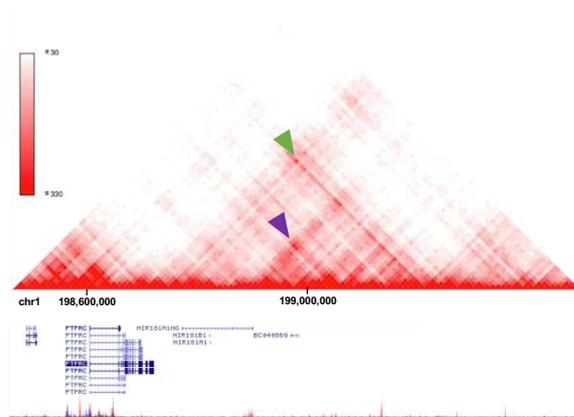
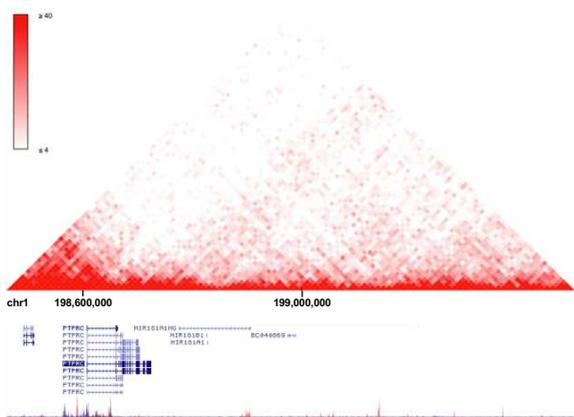
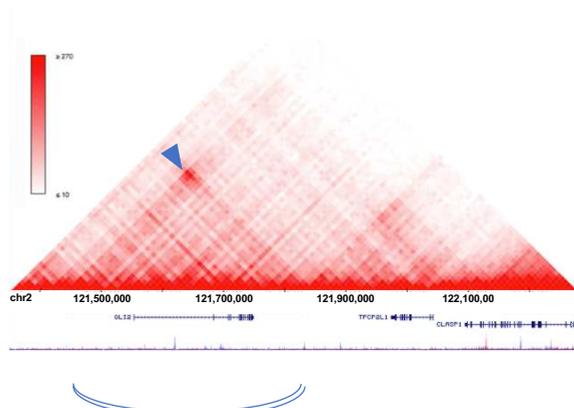
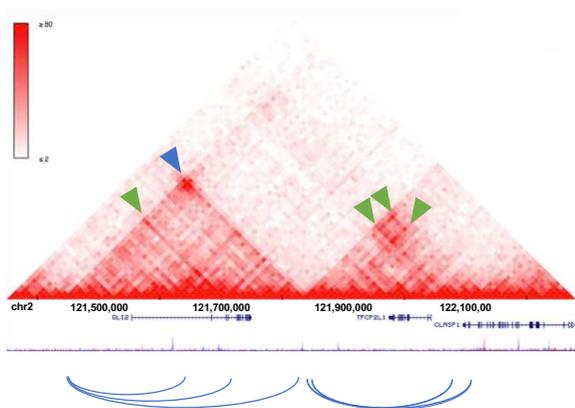


- GM12878 (9,448 Loop Domains; 9,274 Contact Domains; 15,366 intra-TAD loops)
- K562 (6,057 Loop Domains; 5,975 Contact Domains; 15,130 intra-TAD loops)

C.

Ubiquitous Intra-TAD Loops Are Stronger Interactions

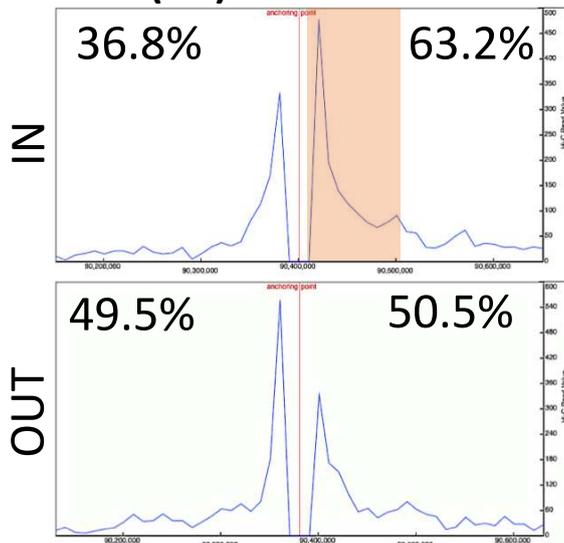
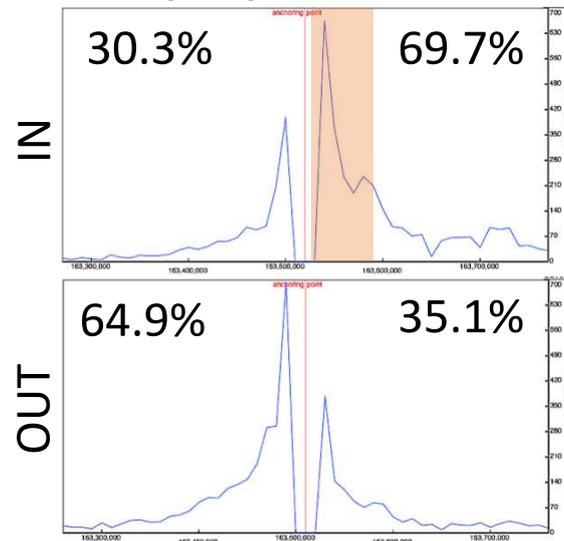
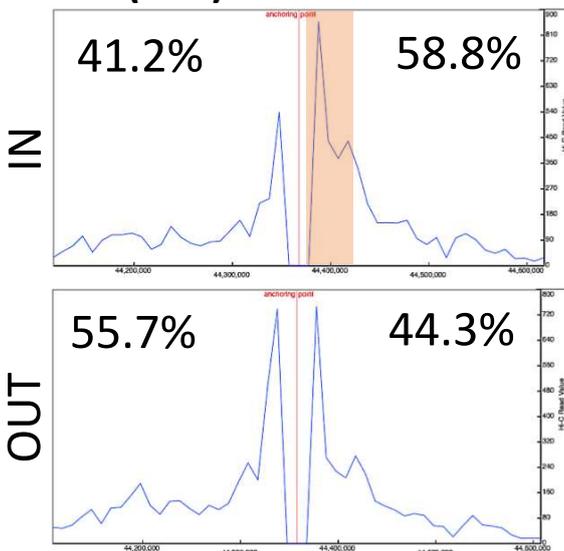
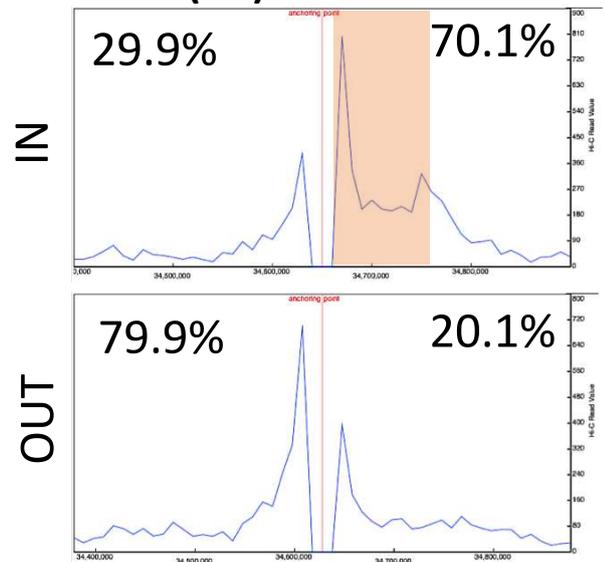
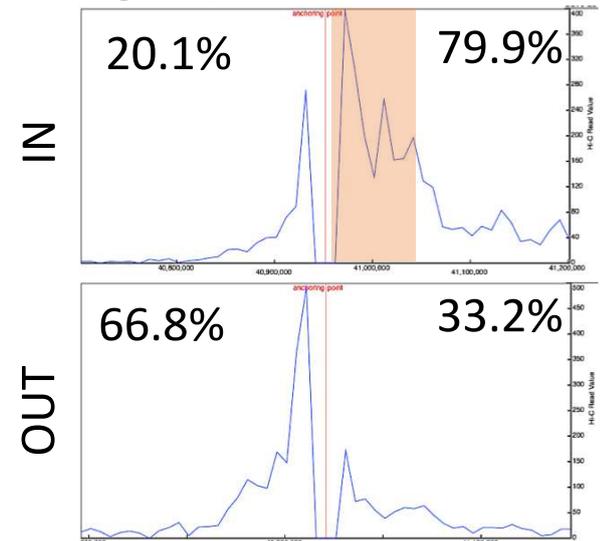


D.K562:GM12878:**High res.
Hi-C****Genes
Encode K27ac****Intra-TAD loops****E.**K562:GM12878:**High res.
Hi-C****Genes
Encode K27ac****Intra-TAD loops****F.**K562:GM12878:**High res.
Hi-C****Genes
Encode K27ac****Intra-TAD loops**

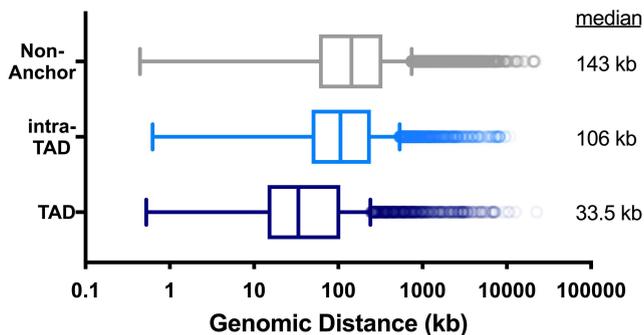
▲ Tissue Ubiquitous intra-TADs

▲ Tissue Specific intra-TADs

▲ Other tissue-specific interactions

A.**Alb (+--): chr5:90400869****Hnf4a (+++): chr2:163527154****Scd1 (+++): chr19:44367709****B.****Sox2 (-+-): chr3:34649994****Pdgfr1 (+++): chr8:40951860**

A. Distance to Nearest Divergent Loop Anchor



B.

1,081 Inter-TAD Genes (TSS within inter-TAD region)

Annotation Cluster 1 (Enrichment Score: 2.83)		Count	P Value
Up Keywords	Ribosomal Protein	21	4.9E-04
Kegg Pathway	Ribosome	18	7.3E-04
GO Term (CC)	Ribosome	20	1.1E-03
GO Term (MF)	Struct. Consist. Of Ribosome	22	1.2E-02

Annotation Cluster 2 (Enrichment Score: 2.7)		Count	P Value
GO Term (CC)	Nucleosome	15	4.3E-04
InterPro	Histone-fold	15	1.1E-03
Up Keywords	Nucleosome Core	13	1.2E-03
KEGG Pathway	Systemic Lupus Erythematosus	14	2.9E-02

Annotation Cluster 3 (Enrichment Score: 2.35)		Count	P Value
Up Keywords	Mitosis	24	8.3E-04
GO Term (BP)	Mitotic Nuclear Division	25	3.8E-03
Up Keywords	Cell Division	28	5.7E-03
GO Term (BP)	Cell Division	28	2.3E-02

Annotation Cluster 4 (Enrichment Score: 2.33)		Count	P Value
GO Term (BP)	DNA repl.-dep. Nucleosome Assembly	8	8.1E-04
GO Term (BP)	Protein Heterotetramerization	9	1.6E-03
GO Term (BP)	DNA Methylation on Cytosine	7	3.7E-03

1,132 Intra-TAD Genes (TSS < 20 kb from inter-TAD region)

Annotation Cluster 1 (Enrichment Score: 3.84)		Count	P Value
Up Keywords	Nucleotide-binding	115	4.1E-05
Up Keywords	ATP-binding	93	6.2E-05
GO Term (MF)	ATP-binding	101	1.2E-03

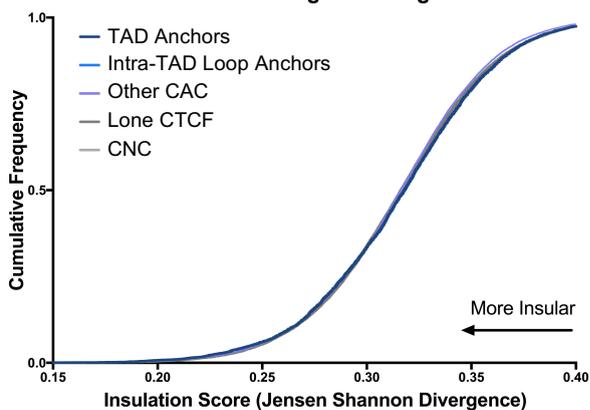
Annotation Cluster 2 (Enrichment Score: 2.35)		Count	P Value
Up Keywords	Kinase	50	1.9E-03
GO Term (MF)	Kinase Activity	50	4.0E-03
GO Term (BP)	Phosphorylation	44	1.2E-03

Annotation Cluster 3 (Enrichment Score: 1.54)		Count	P Value
Up Keywords	Oxidoreductase	43	9.4E-03
GO Term (MF)	Oxidoreductase Activity	40	5.0E-02
GO Term (BP)	Oxidation-reduction Process	44	5.1E-02

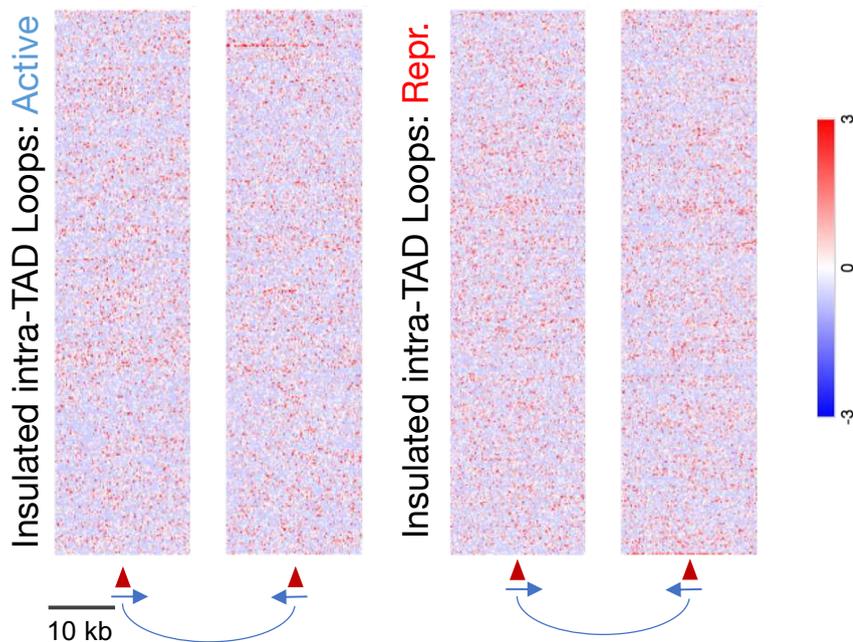
Annotation Cluster 4 (Enrichment Score: 1.5)		Count	P Value
Interpro	D-isomer Specific 2-hydroxyacid Dehydrogenase, NAD-binding	3	1.2E-02
Interpro	D-isomer Specific 2-hydroxyacid Dehydrogenase, Catalytic Domain	3	1.2E-02
GO Term (MF)	Oxidoreductase Activity, Acting on the CH-OH Group of Donors, NAD or NADP as Acceptor	4	2.1E-01

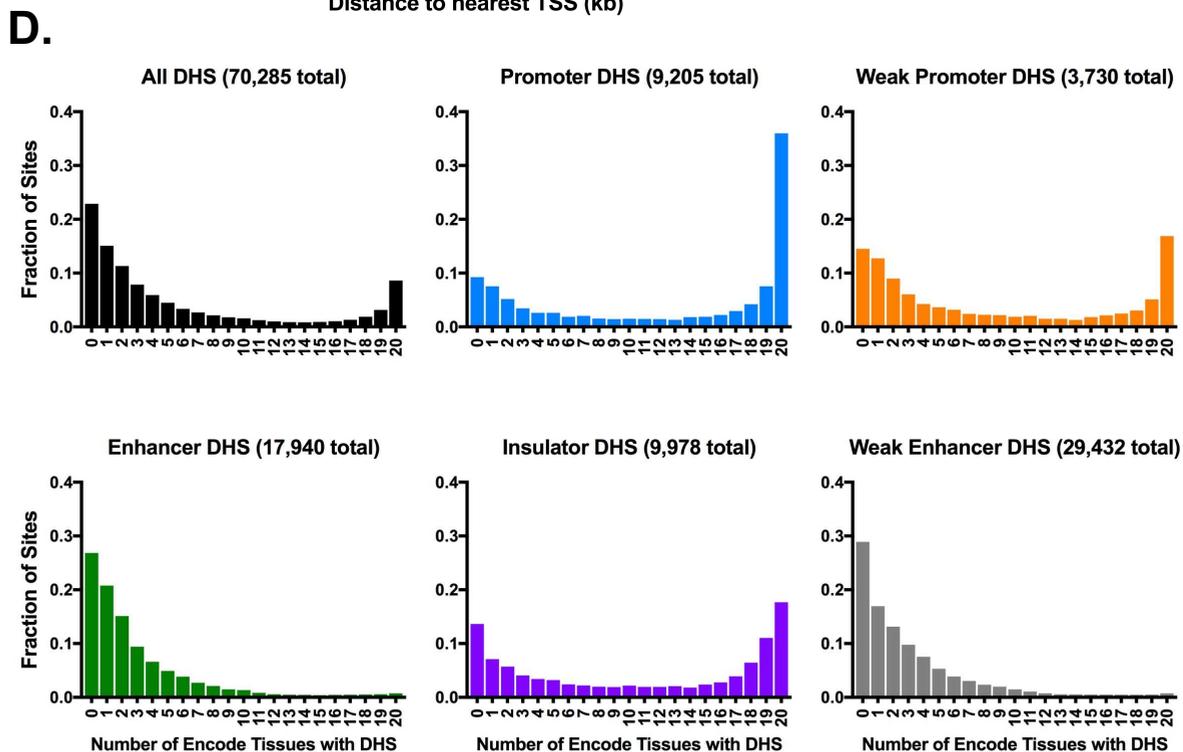
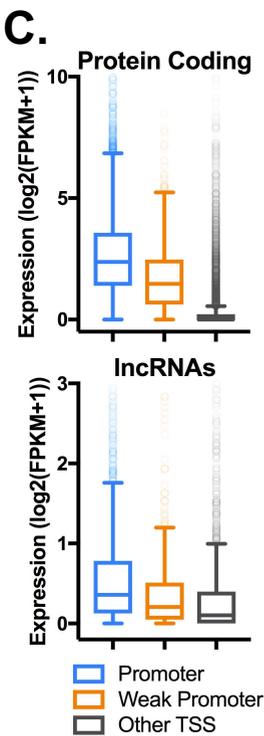
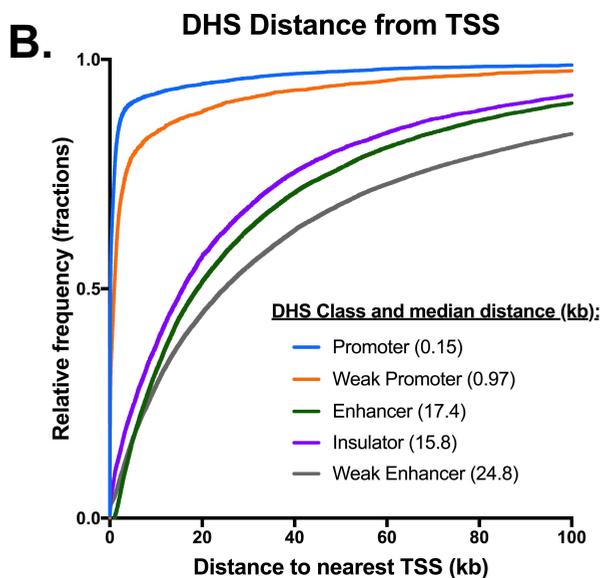
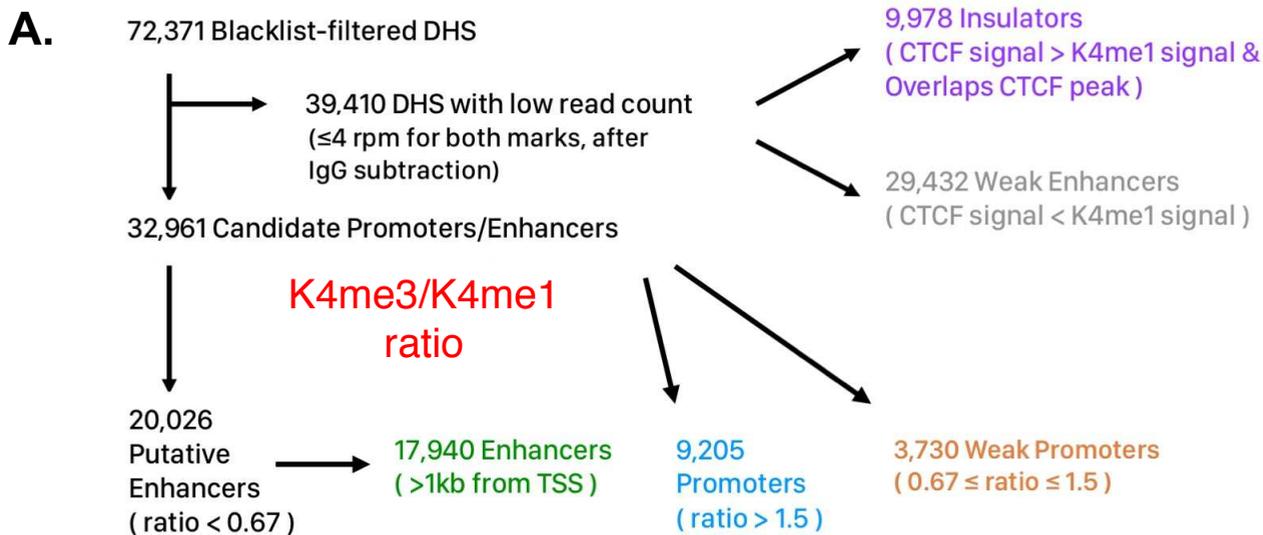
C.

Insulation of IgG ChIP Signal

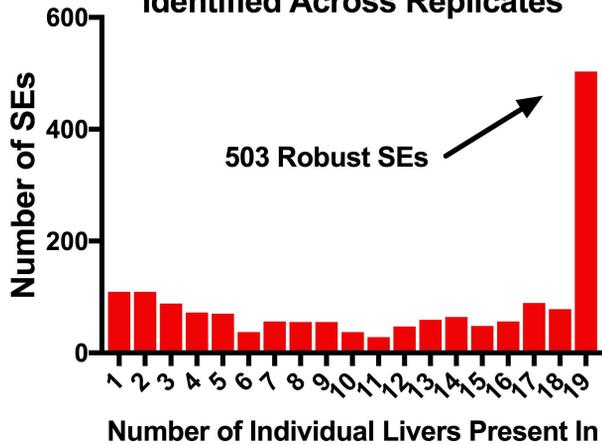


D. IgG Signal around insulated intra-TAD Loop Anchors

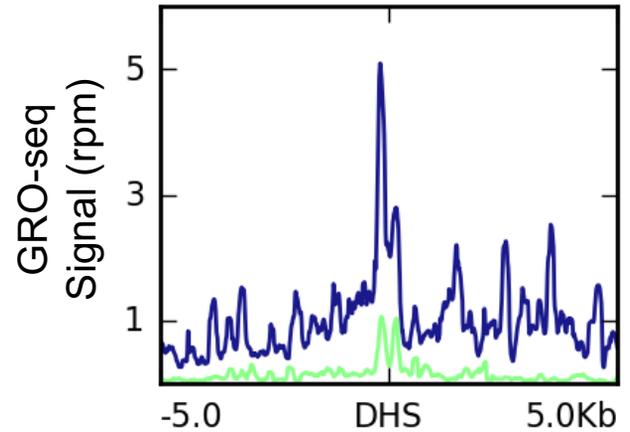




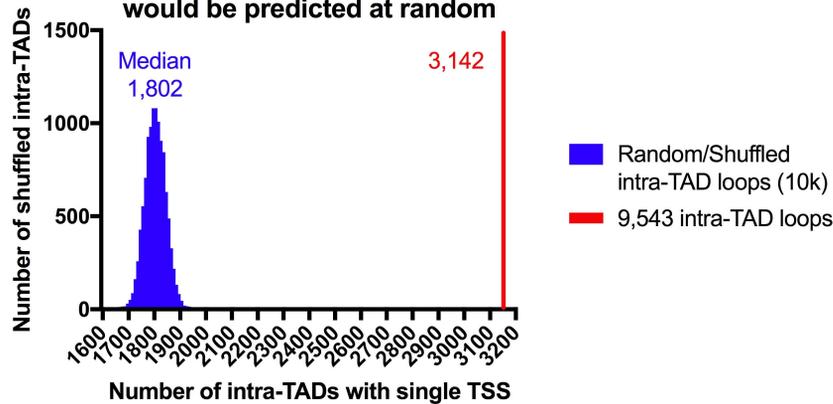
A. Overlap of Super Enhancers Identified Across Replicates



B. Enhancer RNA at SE/TE Constituents:



C. More single-gene intra-TAD loops than would be predicted at random



D.

GO Term Enrichment for Genes in Single-TSS intra-TAD Loops:

Annotation Cluster 1 (Enrichment Score: 3.75)		Count	P Value
Up Keywords	Transcription Regulation	149	8.7E-05
Up Keywords	Transcription	151	1.7E-04
GO Term (BP)	Transcription, DNA-templated	155	1.2E-01
GO Term (BP)	Regulation of Transcription, DNA-templated	172	4.7E-01

Annotation Cluster 2 (Enrichment Score: 3.74)		Count	P Value
InterPro	PDZ Domain	24	4.8E-05
Up Seq Feature	Domain:PDZ	18	3.1E-04
SMART	PDZ	24	4.1E-04

Annotation Cluster 3 (Enrichment Score: 3.6)		Count	P Value
GO Term (BP)	Protein Phosphorylation	65	2.0E-05
InterPro	Protein Kinase, Catalytic Domain	54	9.5E-05
InterPro	Protein Kinase, ATP Binding Site	44	1.2E-04
InterPro	Protein Kinase-like Domain	56	2.0E-04
GO Term (MF)	Protein Kinase Activity	56	4.5E-04
Up Seq Feature	Domain:Protein Kinase	54	4.7E-04
Up Seq Feature	Binding Site:ATP	55	7.2E-03

Annotation Cluster 4 (Enrichment Score: 3.37)		Count	P Value
Interpro	Transforming Growth Factor-beta, N-term.	9	3.6E-05
GO Term (BP)	Growth	11	7.1E-05
Interpro	Transforming Growth Factor-beta-related	9	3.0E-04
Interpro	Transforming Growth Factor-beta, Conserv. Site	9	4.8E-04
Interpro	Transforming Growth Factor-beta, C-term.	9	1.3E-03
GO Term (MF)	Transforming Growth Factor Beta Receptor Binding	10	1.8E-03
SMART	TGFB	9	2.8E-03

Annotation Cluster 4 (Enrichment Score: 3.36)		Count	P Value
Interpro	Transforming Growth Factor-beta, N-term.	68	3.2E-05
GO Term (BP)	Growth	66	9.6E-04
Interpro	Transforming Growth Factor-beta-related	59	2.7E-03

