

1 **High-resolution CTCF footprinting reveals impact of chromatin state on**
2 **cohesin extrusion dynamics**

3
4 Corriene E. Sept^{1,2,3}, Y. Esther Tak^{4,5}, Christian G. Cerdá-Smith⁶, Haley M. Hutchinson⁶, Viraat
5 Goel^{3,7,8}, Marco Blanchette⁹, Mital S. Bhakta⁹, Anders S. Hansen^{3,7,8}, J. Keith Joung^{4,5}, Sarah
6 Johnstone^{3,10}, Christine E. Eyler^{11,12}, & Martin J. Aryee^{1,2,3}

7
8 **Affiliations:**
9 ¹Department of Biostatistics, Harvard T.H. Chan School of Public Health; Boston, MA 02115,
10 USA
11 ²Department of Data Sciences, Dana-Farber Cancer Institute; Boston, MA 02115, USA
12 ³Broad Institute of MIT and Harvard; Cambridge, MA 02142, USA
13 ⁴Molecular Pathology Unit, Massachusetts General Hospital; Charlestown, MA 02129, USA
14 ⁵Department of Pathology, Harvard Medical School; Boston, MA 02115, USA
15 ⁶Department of Pharmacology and Cancer Biology, Duke University School of Medicine;
16 Durham, NC 27710, USA.
17 ⁷Department of Biological Engineering, Massachusetts Institute of Technology; Cambridge, MA
18 02139, USA
19 ⁸Koch Institute for Integrative Cancer Research; Cambridge, MA 02139, USA
20 ⁹Dovetail Genomics, Cantata Bio LLC, Scotts Valley, CA 95066, USA
21 ¹⁰Department of Pathology, Dana-Farber Cancer Institute; Boston, MA 02215, USA.
22 ¹¹Department of Radiation Oncology, Duke University School of Medicine; Durham, NC 27710,
23 USA.
24 ¹²Duke Cancer Institute, Duke University School of Medicine; Durham, NC 27710, USA.
25
26

27 **Abstract**

28
29 DNA looping is vital for establishing many enhancer-promoter interactions. While CTCF is
30 known to anchor many cohesin-mediated loops, the looped chromatin fiber appears to
31 predominantly exist in a poorly characterized actively extruding state. To better characterize
32 extruding chromatin loop structures, we used CTCF MNase HiChIP data to determine both
33 CTCF binding at high resolution and 3D contact information. Here we present *FactorFinder*, a
34 tool that identifies CTCF binding sites at near base-pair resolution. We leverage this substantial
35 advance in resolution to determine that the fully extruded (CTCF-CTCF) state is rare genome-
36 wide with locus-specific variation from ~1-10%. We further investigate the impact of chromatin
37 state on loop extrusion dynamics, and find that active enhancers and RNA Pol II impede cohesin
38 extrusion, facilitating an enrichment of enhancer-promoter contacts in the partially extruded loop
39 state. We propose a model of topological regulation whereby the transient, partially extruded
40 states play active roles in transcription.

41 **Background**

42
43 Topologically associated domains (TADs) and regulatory enhancer-promoter chromatin loops
44 are largely formed by the cohesin complex through the process of CTCF-mediated loop
45 extrusion^{1,2}. Topological alterations and subsequent changes in enhancer-promoter (EP) contacts
46 can modify gene expression^{3,4} and cause aberrant phenotypes⁵⁻⁸. CCCTC-binding factor (CTCF)
47 can act as an extrusion barrier through its ability to bind and stabilize cohesin on DNA, serving
48 to preferentially localize and anchor one or both ends of cohesin loops. Genes with promoter-
49 proximal CTCF binding sites have been shown to have increased dependence on distal
50 enhancers⁹⁻¹¹, although the exact mechanisms involved are not well understood.

51
52 Although conventional 3C techniques give an impression of static loops, cohesin-mediated
53 chromatin loops are actually dynamic with an extrusion rate of ~1kb/s¹². Recent live cell-
54 imaging studies of two TADs found that the fully extruded state with a loop formed between two
55 convergent CTCF-bound anchors was present only 3-30% of the time^{13,14}. While these findings
56 suggest that CTCF loops spend the vast majority of their time partially-extruded, the partially-
57 extruded state has not yet been well characterized.

58
59 Several studies have found evidence of promoter-proximal CTCF binding sites (CBS) having
60 large impacts on EP contact frequencies and transcription⁹⁻¹¹. Putting this together with the high
61 prevalence of partially extruded CTCF-mediated loops, we hypothesize that promoter-proximal
62 CTCF sites enable gene regulation by halting cohesin on one side while cohesin continues to
63 extrude on the other side. Enhancers then slow down extrusion, thus enabling an increase in EP
64 contacts without requiring a fully extruded loop. The relationship between EP contacts and
65 transcription can be nonlinear such that small increases in EP contacts may cause large changes
66 in transcription^{3,4}. As a result, even minor decreases in extrusion rate through enhancer regions
67 may affect gene expression.

68
69 The ability of MNase to efficiently digest naked DNA while sparing protein-bound DNA has
70 been employed in various strategies to footprint the binding sites of proteins such as transcription
71 factors with near base-pair resolution¹⁵⁻¹⁸. A key advantage of using MNase over sonication-
72 based protocols is the shorter fragment size obtained, which directly leads to higher resolution
73 TF binding site identification. More recently, MNase DNA fragmentation has also been applied
74 to proximity ligation assays to map 3D genome architecture with nucleosome (~150 bp)
75 resolution, enabling precise characterization of 3D architecture including at TAD boundaries and
76 punctate enhancer-promoter interactions¹⁹⁻²². Since MNase HiChIP enables precise
77 characterization of both TF-binding and 3D contacts, it is uniquely poised to define how CTCF
78 enables 3D contacts.

79

80 To better characterize the partially extruded chromatin loop state, we first develop a
81 computational technique for high-resolution footprinting of CTCF using MNase HiChIP data.
82 We then employ this to study how, through its interaction with the looping factor cohesin, CTCF
83 can facilitate long-range DNA contacts. We further characterize how the length of loops
84 extruded by cohesin is affected by local chromatin state factors such as enhancer and RNA Pol II
85 density.

86

87 **Results**

88

89 MNase HiChIP generates short, TF-protected and longer, histone-protected DNA 90 fragments

91 We used Micrococcal nuclease (MNase) HiChIP²³ with a CTCF antibody to profile 3D
92 architecture in K562 cells, generating 150 bp reads with over 380 million unique pairwise
93 contacts across four replicates. Briefly, following cell fixation with DSG and formaldehyde,
94 chromatin is digested by MNase, immunoprecipitated to enrich for CTCF-bound DNA, and free
95 ends are then ligated. After reverse-crosslinking, the resulting ligation products are sequenced
96 from both ends and the mapping locations of the paired reads can be used to infer chromosomal
97 locations of the physically interacting loci. In cases where the pre-ligation fragments are shorter
98 than the read length it is also possible to infer the fragment length as the ligation junction
99 position will be observed within one or both of the reads. If multiple fragments within a read are
100 short enough to be aligned to distinct genomic locations, this is termed an ‘observed ligation’
101 (Fig. 1a, Supp Fig 1).

102

103 As expected, due to the preference of MNase to selectively cleave DNA not shielded by bound
104 proteins and the high abundance of histones in chromatin (Fig. 1b), the predominant fragment
105 length is approximately 150 bp, indicative of cuts between nucleosomes²⁴ (Fig 1c). We also
106 noted a distribution of shorter fragment lengths, with 20% representing lengths shorter than 120
107 bp (Fig. 1d). A metaplot centered on CTCF binding site motifs shows an enrichment of 30-60 bp
108 fragments suggesting that these shorter fragments represent CTCF-bound DNA (Fig. 1c)^{2,25,26}.
109 Consistent with this, we find that short (<80 bp) fragments have a 10-fold higher overlap
110 frequency with CTCF motifs than long (>120 bp) fragments (Fig. 1d). This is similar to data
111 from the MNase-based CUT&RUN assay that also results in short fragments protected by small
112 proteins such as transcription factors¹⁷.

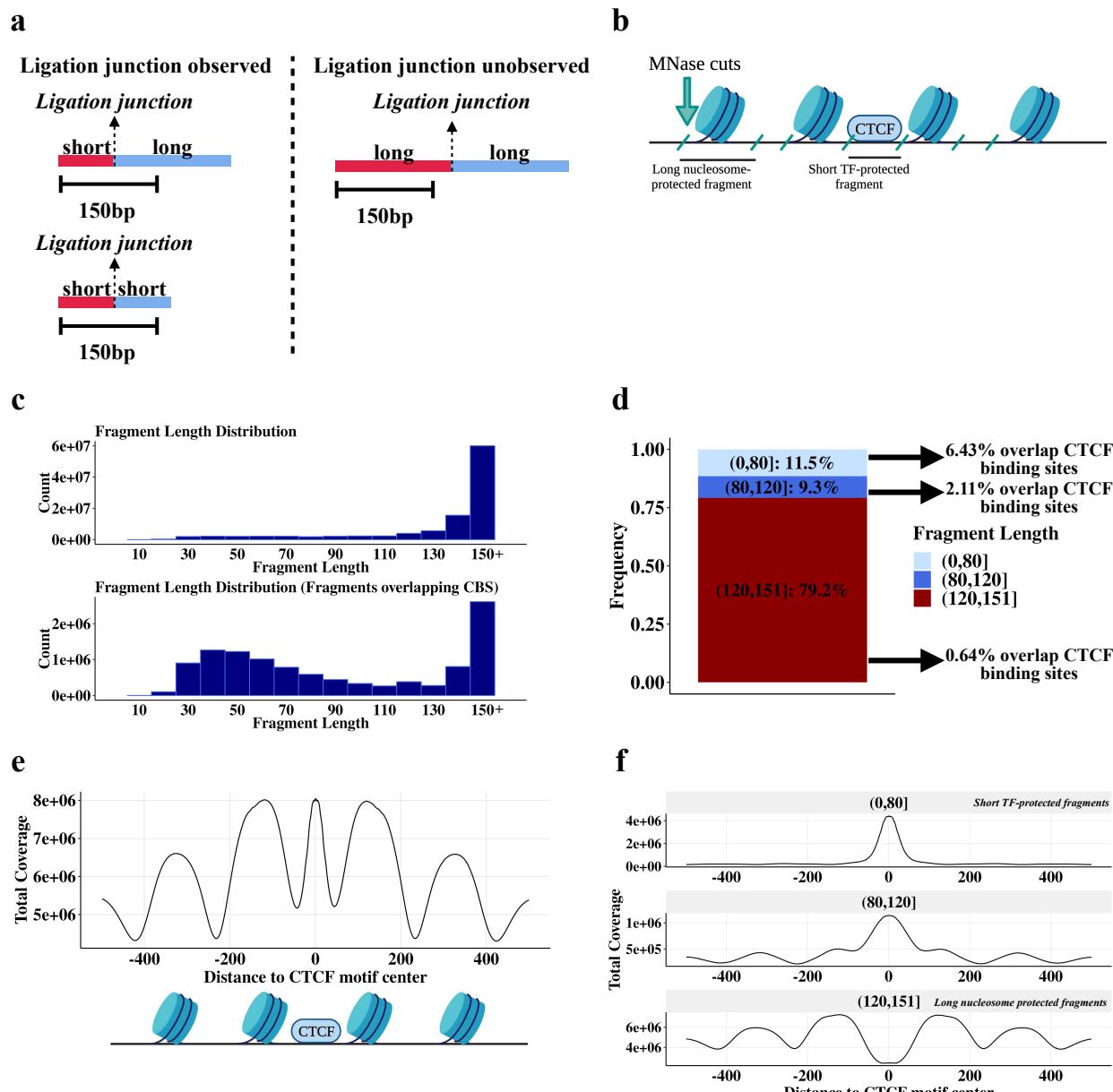
113

114 Fragment pileups at CTCF motif loci (Fig. 1e) show a strong enrichment of short fragments
115 centered on the CTCF motif sequence, and a concomitant depletion of long fragments at motifs
116 (Fig. 1f). Long fragments, in contrast, show peaks with a strong ~200 bp periodicity adjacent to
117 the central CTCF binding site (Fig. 1f). This is consistent with the ability of CTCF to precisely
118 position a series of nucleosomes adjacent to its binding site²⁵. Note that while long (>120 bp)
119 fragments are depleted at CTCF binding sites, they still represent a significant fraction of reads at

120 these sites (Fig. 1c). This likely reflects that CTCF motif loci without a bound CTCF are
 121 frequently instead occupied by histones²⁵, and even CTCF motifs with very strong CTCF ChIP-
 122 seq signal are not always occupied by a CTCF.

123

124 In summary, long fragments correspond to nucleosome-protected DNA whereas short fragments
 125 arise from TF-protected DNA. This is due to the different sizes of CTCF and histone octamers,
 126 with nucleosomes protecting about twice the amount of DNA that CTCF protects²⁵. Since
 127 MNase cuts around bound proteins, the different protein sizes directly translate to different
 128 fragment lengths. Accordingly, we next filter out long, nucleosome-protected fragments and
 129 focus on short, TF-protected fragments to identify CBS.



130

131 **Fig. 1** MNase ChIP data contains short (~<80 bp) CTCF-protected fragments and

132 longer (~>120 bp) nucleosome-protected fragments. **a** Schematic illustrating relationship
133 between short fragments and observed ligations. **b** Schematic illustrating how the fragment
134 length results from MNase cutting around bound proteins of different sizes. **c** Fragment length
135 distribution for all fragments (top plot) and fragments overlapping occupied CTCF motifs (lower
136 plot). Occupied CTCF motifs are defined here as CTCF motifs within 30 bp of a CTCF ChIP-seq
137 peak summit. **d** Boxplot quantifying the frequency of different fragment lengths genome-wide
138 and how often each fragment length group overlaps an occupied CTCF motif. Occupied CTCF
139 motifs are defined here as CTCF motifs within 30 bp of a CTCF ChIP-seq peak summit. **e**
140 Fragment coverage metaplot +/- 500 bp around CTCF binding sites. Schematic below the
141 coverage metaplot illustrates the proteins producing these peaks. **f** Plot (e) stratified by fragment
142 length.

143

144 *FactorFinder leverages the strand-specific bimodal distribution of short fragments*
145 *around CBS to obtain precise detection of CTCF binding*

146 In order to characterize CTCF-mediated chromatin loop interactions, we first set out to map
147 CTCF loop anchors with high resolution. We take advantage of the difference in fragment
148 lengths associated with CTCF-bound vs nucleosome-bound DNA to focus only on likely CTCF-
149 bound fragments. Fragment lengths can be determined for all fragments with length less than 150
150 bp; the 150 bp read length results in censoring of fragments longer than 150 bp. While exact
151 fragment lengths can be obtained for all fragments shorter than 150 bp, observed ligations
152 require a shorter fragment length. This is because observed ligations require distinct mapping of
153 fragments on either side of the ligation junction. Since at least ~25 bp are required to align a
154 sequence to the reference genome, this results in fragments characterized as observed ligations
155 having a maximum fragment length of ~125 bp, sufficient for the identification of most CTCF-
156 protected DNA fragments. Consequently, the fraction of informative, CTCF-protected fragments
157 decreases with shorter sequencing read length (Supp Fig 1). The effect of subsetting the CTCF
158 HiChIP dataset to only short fragments (<125 bp, identified by the proxy of an observed
159 ligation), is shown in Fig 2a,b. These shorter, presumably CTCF-protected fragments, are
160 overwhelmingly located immediately adjacent to CTCF motifs.

161

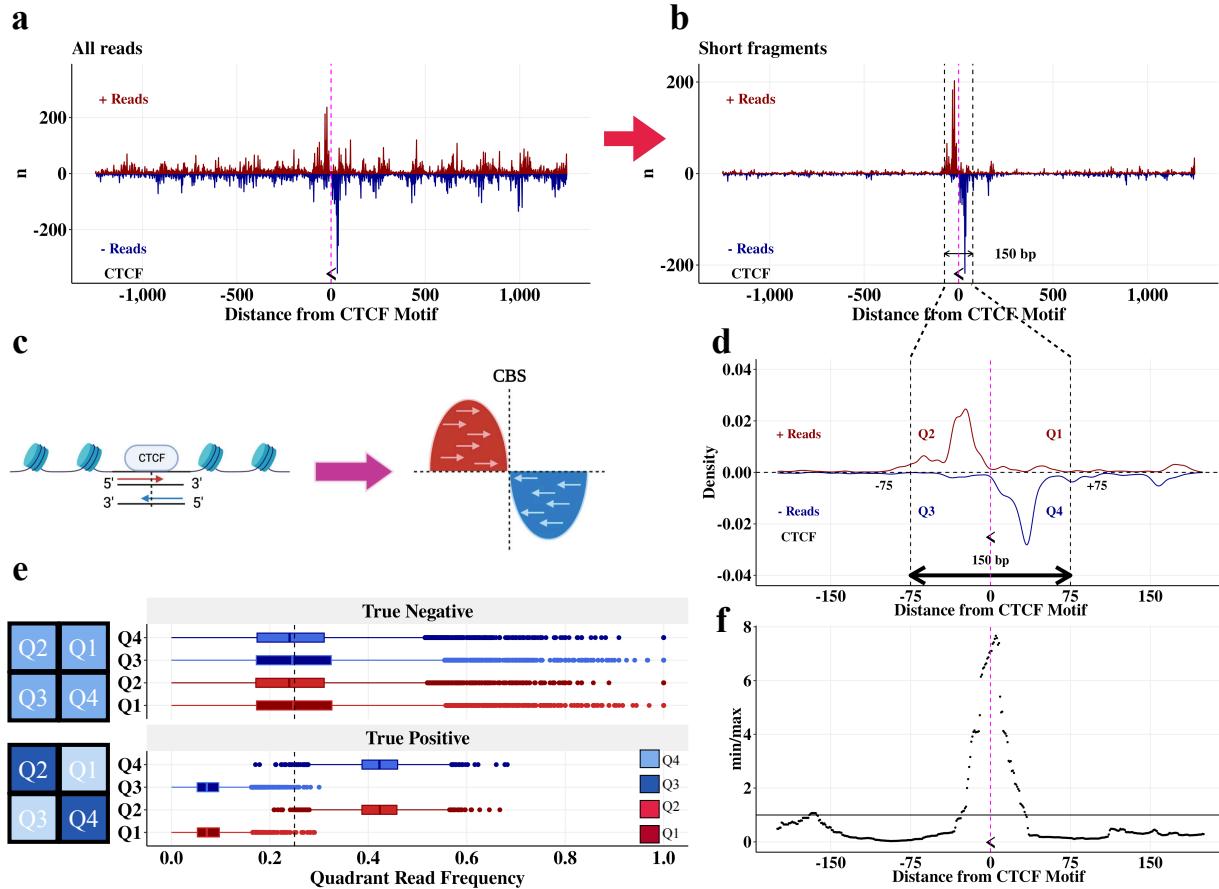
162 Sequencing of short, CTCF-protected fragments results in a bimodal read distribution centered
163 on the CBS, with read 5' location peaks observed upstream (positive strand) and downstream
164 (negative strand) of the CBS (Fig. 2c). We refer to these regions as quadrants 2 and 4 (Q2 and
165 Q4) respectively (Fig. 2d, e). In contrast, reads from the positive strand downstream of the CBS
166 (Q1) and negative strand upstream of the CBS (Q3) correspond to fragments with MNase cut
167 sites underneath CTCF-protected DNA, and therefore reflect a lack of CTCF occupancy. CTCF
168 binding therefore produces an enrichment of reads in Q2,Q4 and a depletion of reads in Q1,Q3
169 (Fig. 2e). At sites without protein binding, MNase can cut at any location resulting in no
170 enrichment of reads in Q2 and Q4 compared to Q1 and Q3 (Fig. 2e). As a result, we can

171 determine CTCF binding by testing if there are significantly more reads in Q2 and Q4 than Q1
172 and Q3 (Fig. 2f).

173
174 We can consider each read as an independent draw from a multinomial distribution with four
175 categories corresponding to the four quadrants. Under the null hypothesis, each read has equal
176 probability of belonging to any of the four quadrants $Q_i, i \in \{1,2,3,4\}$. Because true CTCF
177 binding induces a strong read pile-up in *both* quadrants 2 and 4 in addition to a depletion of reads
178 in quadrants 1 and 3 (Fig. 2d, e, f), we test for an enrichment of reads in Q2 and Q4 compared to
179 Q1 and Q3 by estimating the *FactorFinder* statistic $\hat{\alpha} = \frac{\min(n_2, n_4)}{\max(n_1, n_3)}$, where n_i is the number of
180 reads in Q_i . We then test if $\hat{\alpha}$ is significantly greater than 1. Note that min and max are used to
181 enforce that both quadrants 2 and 4 must have more reads than both quadrants 1 and 3; using the
182 average would enable read pile-ups that occur in quadrant 2 or 4 (but not both) to be spuriously
183 called as CTCF binding events.

184
185 To evaluate the significance of $\hat{\alpha}$ at a particular total read count $N = \sum_{i=1}^4 n_i$, we simulated 100
186 million samples under the null hypothesis that each fragment is equally likely to occur in any of
187 the four quadrants. This was done at each total read count ranging from 5 to 500. P-values at read
188 counts beyond 500 are very similar to those at 500, so 500+ read counts are treated as bins with
189 500 total read count (Supp Fig 2). The empirical CDF of the 100 million $\log_2(\hat{\alpha})$ at a given total
190 read count was then computed and used to evaluate the probability of observing a value more
191 extreme than $\log_2(\hat{\alpha})$ under the null hypothesis. The empirical CDF was evaluated at a sequence
192 of possible $\log_2(\hat{\alpha})$ between 0 and 5 at step sizes of 0.01 (this corresponds to $\hat{\alpha} \in [1, 32]$.) This
193 approach produces the same p-values as using $\hat{\alpha}$ instead of $\log_2(\hat{\alpha})$, but using the log enables
194 smaller step size at large values of $\hat{\alpha}$. After acquiring the grid of p-values for each $\hat{\alpha}$ at a given
195 read count N , we match the observed $\hat{\alpha}$ at a read count of N with the corresponding p-value from
196 the table. Because this approach only requires quadrant-specific read counts to match with the
197 given table of p-values, it is very computationally efficient. Furthermore, by using the
198 multinomial framework we place no assumptions on the reads within each quadrant being
199 distributed as poisson, negative binomial, or another distribution. The only assumption we make
200 is that in the event of no CTCF binding, the reads are equally distributed amongst the four
201 quadrants. We have shown this assumption holds in Figures 2c, d, e.

202
203 In brief, we have shown that short fragments exhibit a strand-specific, bimodal distribution
204 centered on the CBS. This distribution arises from MNase cutting around a bound CTCF and
205 subsequent sequencing 5' to 3' of the DNA. Significance is assessed through a multinomial
206 framework, which has the advantage of not placing any assumptions on the distribution of reads
207 within each quadrant. Now that we have explored the theory behind *FactorFinder*, we
208 demonstrate its ability to identify CBS with high resolution and accuracy.



209

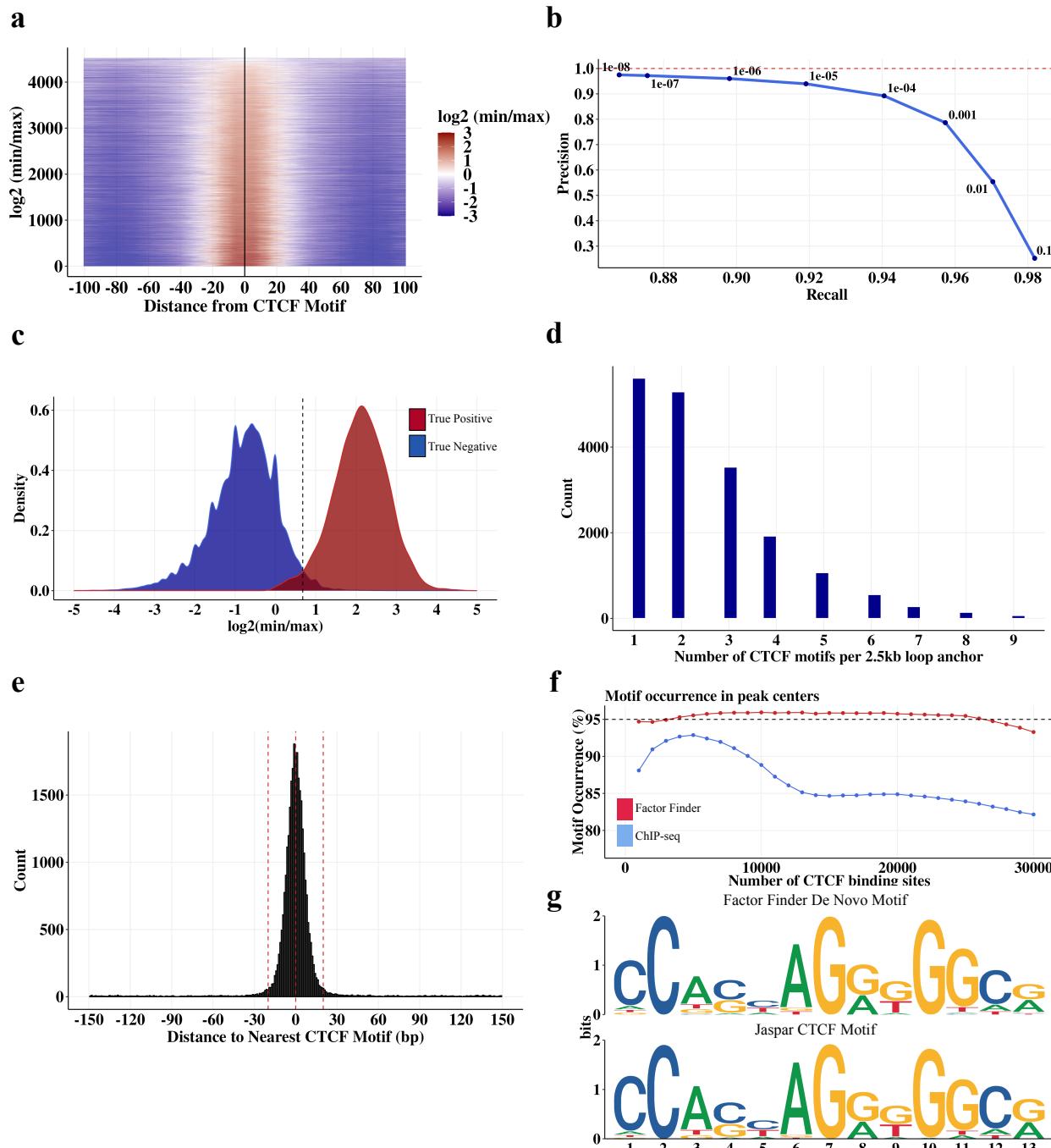
210 **Fig. 2** True CTCF binding sites have a bimodal strand-specific distribution centered on the
211 CTCF motif. **a** Unfiltered reads +/- 1250 bp around a CTCF binding site located on the negative
212 strand (chr1: 30,779,763 - 30,779,781). The midpoint of the CTCF motif is marked with the
213 symbol “<”, representing that it is on the negative strand, and a pink line. **b** Plot (a) filtered to
214 observed ligations (equivalently, short fragments.) **c** Schematic demonstrating the bimodal read
215 pile-up around a CTCF binding site. **d** Plot (b) as a density plot and zoomed in on the CTCF
216 motif, with quadrant annotations. **e** Distributions of reads in quadrants for true negative and true
217 positive CTCF binding sites in DNA loop anchors. True positives are defined as CTCF motifs
218 that are the only CTCF motif in a loop anchor and within 30 bp of a CTCF ChIP-seq peak. True
219 negatives are areas of the loop anchors with one CTCF motif that are at least 200 bp from the
220 CTCF motif. Schematics of the quadrant read pile-up patterns are shown next to the
221 corresponding true positive and true negative boxplots. **f** *FactorFinder* statistic ($\hat{a} = \frac{\min(n_2, n_4)}{\max(n_1, n_3)}$)
222 for plot (d) peaks at the CTCF motif.

223

224 Model evaluation

225 *FactorFinder* uses a biologically-informed model that takes advantage of the distribution of short
226 fragments around a CTCF binding site to pinpoint CTCF binding. Additionally, our use of a
227 multinomial framework for significance evaluation avoids placing any distributional assumptions

228 on the reads within a quadrant. We then sought to benchmark our CTCF binding site
229 identification performance using CTCF motif locations²⁷, CTCF ChIP-seq peaks²⁸, and loop
230 anchors identified by FitHiChIP at 2.5kb resolution²⁹.
231
232 We define a high stringency true positive set of CTCF binding sites as CTCF motifs in loop
233 anchors that are located within 30 bp of a CTCF ChIP-seq peak summit. To avoid ambiguity due
234 to multiple closely spaced motifs, we further selected only those motifs that are unique within a
235 2.5kb loop anchor. Using this true positive set, we observe that the *FactorFinder* statistic,
236 $\log_2(\hat{\alpha}) = \log_2\left(\frac{\min(n_2, n_4)}{\max(n_1, n_3)}\right)$ has signal greater than 0 (equivalently, $\hat{\alpha} > 1$) almost exclusively
237 within 20 bp of the CTCF motif center and centered on 0 bp from the CTCF motif center (Fig.
238 3a). Using this same set of true positive sites (false negatives are the regions of the loop anchors
239 >200 bp from a CTCF motif), we achieve > 90% precision and > 90% recall at a p-value
240 threshold of 1e-05, and maintain high recall and precision at all p-value thresholds < 1e-05 (Fig.
241 3b). This high level of recall and precision is achieved because of the very different
242 *FactorFinder* statistic distributions for true positives and true negatives (Fig. 3c).
243
244 Because 70% of loop anchors defined with 2500 bp resolution contain multiple CTCF motifs
245 (Fig. 3d), higher levels of precision are often needed to determine the specific CTCF motif(s)
246 mediating a CTCF loop. Examining the effectiveness of *FactorFinder* genome-wide, we observe
247 that almost all *FactorFinder* peak summits (93%) are within 20 bp of a CTCF motif center, with
248 a median separation of 5 bp (Fig. 3e). Quantifying accuracy using motif occurrence within 20 bp
249 of a peak summit, we find that *FactorFinder* maintains ~95% motif occurrence while ChIP-seq
250 declines to less than 85% motif occurrence (Fig. 3f). Applying the motif discovery tool
251 STREME³⁰ to 30 bp sequences centered on the *FactorFinder* peak summit produces a motif
252 sequence that exactly matches the core JASPAR CTCF motif (Fig. 3g), further supporting
253 *FactorFinder*'s ability to identify true CTCF binding sites.



254
255
256
257
258
259
260
261
262

Fig. 3 CTCF binding sites identified by *FactorFinder* with single basepair resolution in MNase K562 CTCF HiChIP data. **a** Heatmap of $\log_2(\text{min/max})$ as a function of distance between *FactorFinder* peak center and CTCF motif center within loop anchors. Only CTCF motifs that are unique within a loop anchor and within 30 bp of a CTCF ChIP-seq peak are used. **b** Precision recall curve for true negative and true positive CTCF binding sites in DNA loop anchors. True positives are defined as in (a). True negatives are areas of the loop anchors in (a) that are at least 200 bp from the one CTCF motif. Precision is calculated as $\text{TP} / (\text{TP} + \text{FP})$, recall is calculated as $\text{TP} / (\text{TP} + \text{FN})$. **c** *FactorFinder* statistic density plots using the same set of true positives and

263 true negatives as **(b)**. **d** Distribution of the number of CTCF motifs in a 2.5kb loop anchor. **e**
264 Histogram with 1 bp bin size depicting *FactorFinder* resolution for all peaks genome-wide (not
265 just in loop anchors). **f** Motif occurrence in ChIP-seq and *FactorFinder* peak centers genome-
266 wide. Motif occurrence is calculated as % peak centers within 20 bp of CTCF motif. Only peak
267 centers within 150 bp of a CTCF motif are used for this figure. **g** 30 bp sequences centered on
268 genome-wide *FactorFinder* peak centers produce a de novo motif (top) that matches the core
269 JASPAR CTCF motif (bottom).

270

271 *CTCF and Cohesin occupancy footprints*

272 We next examined the length characteristics of MNase HiChIP fragments overlapping individual
273 CTCF motifs, to infer the presence and identity of the protein occupying the locus. For motifs
274 with non-zero coverage, we observed long, 150+ bp fragments, as shown for three representative
275 motifs in Figure 4a. These fragments likely represent cells with a nucleosome located at the
276 motif locus, and are observed at CTCF motifs genome-wide (Fig. 1c). In addition, for a large
277 subset of CTCF motifs, we also observed short, sub-nucleosome sized (<115 bp) fragments (Fig.
278 4a, Fig. 1c), likely instead representing DNA protected by CTCF.

279

280 A closer examination of the TF-scale fragments at *FactorFinder*-identified bound motifs reveals
281 that they tend to exhibit a skew towards the downstream side of the CTCF motif (Fig. 4a, b, c),
282 suggesting a preferred location for the protein(s) protecting the region from MNase cleavage. We
283 considered cohesin as a potential candidate, given a recent finding that cohesin is stabilized on
284 DNA through a specific interaction with the N terminus of the CTCF protein², which localizes to
285 the downstream side of the CTCF binding site.

286

287 Given CTCF's role in mediating DNA looping we investigated whether the CTCF-adjacent
288 protected footprint might relate to 3D architecture within the cell. We used HiChIP pairwise
289 interaction data where each ligation event reflects a single-cell point-to-point contact, to classify
290 each CTCF motif-overlapping fragment as either 'upstream' or 'downstream', depending on its
291 relationship to its interaction partner. Upstream fragments have long range contacts downstream
292 of the motif, and therefore have looping contacts in the same direction as a chromatin loop
293 mediated by cohesin bound to the N terminus of the CTCF protein. Examining the difference in
294 coverage downstream and upstream of CBS genome-wide, we observe that upstream fragments
295 overlapping CBS with an adjacent strong RAD21 ChIP-seq peak have substantially more
296 adjacent coverage in the ~60 bp region downstream compared to upstream of the motif, while
297 downstream fragments and CBS with weak adjacent RAD21 ChIP-seq peaks exhibit no
298 difference (Fig. 4d). This finding further suggests that the CTCF-adjacent factor is associated
299 with loop formation.

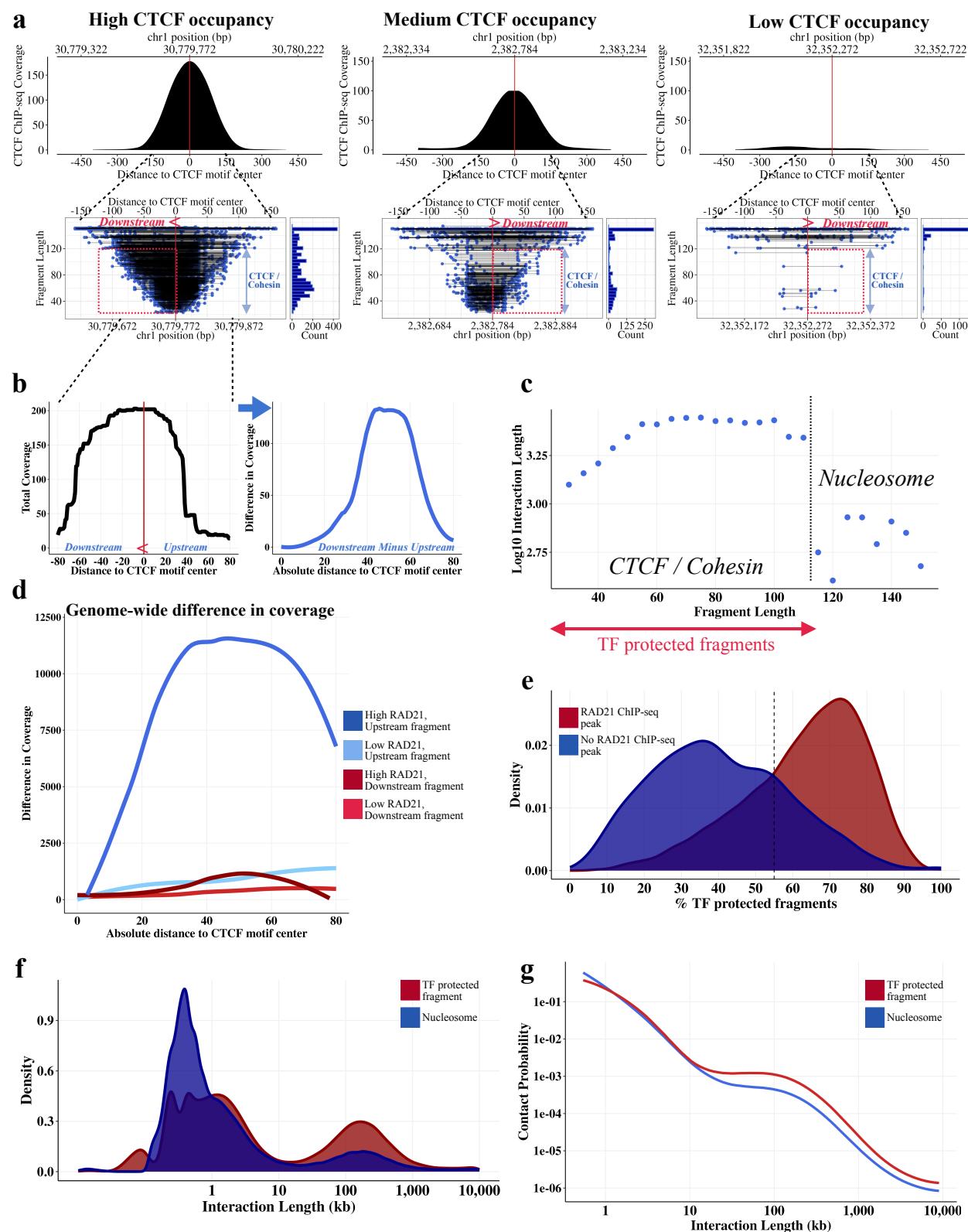
300

301 To further investigate whether the TF footprints identified at CTCF motifs might relate to an
302 architectural role, we used HiChIP data to characterize their interaction patterns. We found that

303 TF-protected fragments (<115 bp) had contacts at substantially longer genomic distances than
304 nucleosome-protected fragments (Fig. 4c), suggesting that the TF presence may facilitate long
305 range interactions. Furthermore, we computed the frequency of TF-protected fragments at all
306 *FactorFinder*-identified CTCF bound sites, and found that it is strongly associated with the
307 presence of a RAD21 ChIP-Seq peak at the motif²⁸ (Fig 4e).

308

309 Examination of the interaction length distribution shows that, as expected, the majority of
310 interactions occur within a linear separation of less than 10kb. The fraction of long-range
311 (>10kb) interactions, however, is significantly enriched (3.5-fold, $p < 10^{-10}$) for short TF-
312 protected fragments as would be expected if these footprints represent CTCF/cohesin (Fig. 4f).
313 Similarly, an examination of the P(s) curve, showing contact probability as a function of linear
314 distance, reveals a decreased attenuation in contact probability at longer interaction lengths (Fig.
315 4g). Taken together, these findings suggest that we can classify CTCF HiChIP interaction data
316 based on footprint/fragment size as involving either unoccupied CTCF sites that tend to have
317 short-range chromatin interactions, or CTCF/cohesin occupied sites that, presumably through
318 loop extrusion, are able to make long-range contacts.



319

320 **Fig. 4** Cohesin and CTCF-protected fragments identified in CTCF MNase HiChIP. **a** High,
321 medium, and low CTCF occupied motifs. Cohesin footprint is observed downstream of the CBS

322 for high and medium CTCF occupancy motifs. For each occupancy level, CTCF ChIP-seq (top)
323 and all fragments overlapping the CTCF motif (bottom left) are depicted, along with the
324 corresponding fragment length histogram (bottom right). **b** Locus-specific high CTCF occupancy
325 figure from **(a)** as a coverage plot (left figure), difference in coverage between downstream and
326 upstream coverage (right figure). **c** Plotting median log₁₀ interaction length as a function of
327 fragment length suggests presence of nucleosome vs TF-protected fragments. Only left
328 fragments overlapping CTCF (+) motifs with start and end at least 15 bp from the CTCF motif
329 were included in this graph to remove confounding by MNase cut site. Using this figure, we are
330 approximating CTCF +/- cohesin-protected fragments as those with fragment length < 115, start
331 and end at least 15 bp from the motif center. **d** Difference in coverage (downstream - upstream)
332 across all CBS shows an increase in coverage downstream of the CTCF motif for upstream
333 fragments underlying CBS with a strong adjacent RAD21 ChIP-seq peak. **e** CTCF motifs that
334 have a nearby RAD21 ChIP-seq peak (within 50 bp) have a larger proportion of TF-protected
335 fragments. **f** TF-protected fragments have a noticeably larger bump in density of long range
336 interactions compared to nucleosome-protected fragments. Fragments were first filtered to those
337 with start and end at least 15 bp from the motif. TF-protected fragments were then defined as
338 fragments with length < 115 bp while nucleosome-protected fragments are fragments with length
339 at least 115 bp. **g** P(S) curve for fragments depicted in **(f)**.

340

341 *Active enhancers and gene transcription hinder cohesin-mediated loop extrusion*
342 Using the techniques described above, MNase HiChIP enables us to simultaneously locate CBS
343 at high resolution, identify footprints of bound proteins, and interrogate specific chromatin
344 contacts at the single molecule level. We next sought to leverage these data to characterize
345 cohesin extrusion dynamics in a range of genomic contexts.

346

347 We first estimated the frequency of fully extruded CTCF-CTCF chromatin loops genome-wide.
348 By obtaining fragments overlapping CTCF binding sites and estimating the fraction of
349 interaction partners overlapping a downstream convergent CTCF motif, we obtain 5% as the
350 genome-wide frequency of the fully extruded CTCF-CTCF state.. We also find a wide CBS to
351 CBS variability with an estimated range of ~1-10% (Fig. 5a). This suggests that most CTCF-
352 anchored chromatin contacts at the single-cell level are in the ‘extruding’ state, rather than
353 joining two CTCF sites. These ranges are consistent with two recent locus-specific live cell
354 imaging studies, which found that the fully extruded loop state is rare at the *Fbn2* TAD¹³ and an
355 engineered TAD on chr15¹⁴, occurring ~3-6%¹³ and ~20-30% of the time¹⁴ respectively. Note
356 that the 20-30% estimate corresponds to a loop existing between any combination of three CBS
357 (+) and three CBS (-).

358

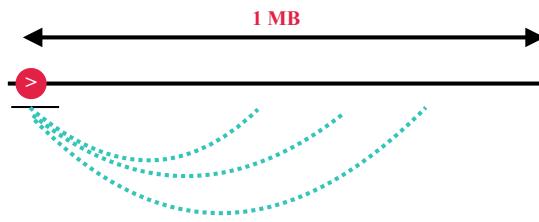
359 We next sought to use our data to examine how cohesin extrusion is impacted by chromatin
360 context. Since HiChIP libraries are a snapshot of millions of cells, we can estimate dynamic
361 extrusion parameters (primarily the average loop size extruded by cohesin³¹) from the interaction

362 length distribution. To determine the impact of chromatin state on cohesin extrusion, we first
363 annotated the 1 MB regions downstream of *FactorFinder* identified CBS with ChromHMM
364 states³² (Fig. 5b) to characterize the DNA through which a cohesin anchored at the CBS would
365 extrude through. Due to the highly correlated nature of ChromHMM annotations (Fig. 5c, d), we
366 then divided the genome into three main chromatin state categories to uniquely classify each 1
367 MB region as either active, polycomb/bivalent or quiescent (Fig. 5d). CTCF/cohesin-protected
368 fragments overlapping CBS were accordingly annotated with the corresponding motif-level
369 chromatin state group, and extruded loop size estimates were obtained for each chromatin state
370 based on the fragment-level interaction lengths.

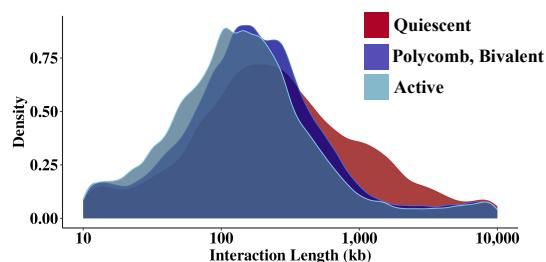
371
372 Interestingly, we find that cohesin extrudes 1.75 times further through quiescent regions (252kb)
373 than through active regions (144kb), corresponding to a difference in average extruded loop size
374 of ~110kb, $p < 10^{-10}$ (Fig. 5e, Supp Fig. 3, Supp Fig. 4 right). The P(s) curve, a plot of interaction
375 decay with distance, confirms a depletion of the longest-range interactions in active regions (Fig
376 5f). This estimate for quiescent regions is consistent with a live cell imaging study of the *Fbn2*
377 locus in the absence of transcription that estimated a processivity of 300kb¹³. As quiescent
378 regions are characterized by low TF binding, low transcription, and minimal histone
379 modifications³³, we hypothesized that the substantial difference in extruded loop size relates to
380 gene activity and enhancer density obstructing loop extrusion. Consistent with this, we found
381 that higher levels of H3K27ac and RNA Pol II binding in the 1MB region downstream of the
382 CBS strongly correlate with lower average extruded loop size (Fig. 5g).

383
384 We sought to establish that the observed differences in loop extrusion length as a function of
385 chromatin state are not confounded by locus-specific effects on cohesin extrusion. Each CBS has
386 locus-specific genetic architecture and a different number of overlapping fragments, so we fit a
387 linear mixed effects model to account for this group-level heterogeneity. Specifically, we
388 compute the ‘cohesin effect’ on loop length, defined as the increase in average interaction length
389 for CTCF/cohesin bound fragments compared to nucleosome bound fragments for each CBS.
390 Controlling for the background interaction frequency of a region in this way confirms that
391 cohesin-associated loops are significantly shorter in active chromatin (Supp Fig. 4 left). Taken
392 together, these findings imply that gene and enhancer activity impede cohesin translocation (Fig.
393 5h).

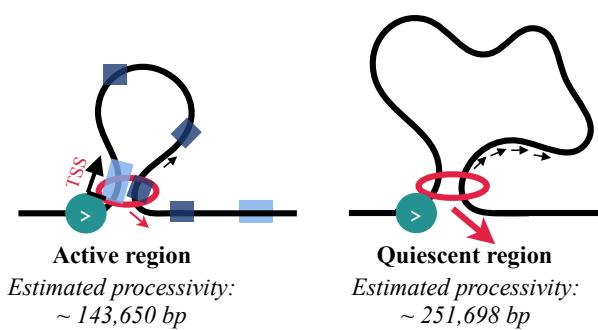
394



f



g



395

396 **Fig. 5** Cohesin extrudes further through quiescent regions than active regions. **a** Most CTCF-
397 mediated looping contacts do not reflect the fully extruded state. Estimate is obtained using left

398 TF-protected (start and end at least 15 bp from motif center, length < 115) fragments that overlap
399 *FactorFinder* identified CBS (+) and have an interaction length greater than 10kb. For each CBS
400 with at least 50 long-range TF-protected fragments overlapping the motif, % convergent is
401 calculated as the number of interaction partners overlapping CTCF (-) motifs / total number of
402 fragments at motif. Because this estimate is conditional on CTCF binding at the anchor, we
403 divide estimates by two to account for the ~50% occupancy of CTCF³⁴. **b** Depiction of how
404 regions were annotated using ChromHMM. Correlation (**c**) and fragment (**d**) heatmaps for
405 ChromHMM annotated unique 1 MB regions downstream of left fragments overlapping CTCF
406 (+) binding sites. All other plots in this figure are filtered to TF-protected (fragment length < 115
407 bp, start and end at least 15 bp from motif center) fragments. Density (**e**) and P(S) curves (**f**) for
408 chromatin state clusters shown in (**c,d**), filtered to the top 20%. Chromatin annotations making
409 up each cluster are added together and quantiles are obtained to determine fragments in the top
410 20% of active chromatin, quiescent chromatin, and bivalent / polycomb chromatin. **g** Ridge plots
411 for the bottom 10% quantile (“Low”) and top 10% quantile (“High”) of H3K27ac bp and number
412 of RNAPII binding sites. ChIP-seq from ENCODE was used to annotate 1 MB downstream of
413 left fragments overlapping CBS (+) for this figure. **h** Diagram illustrating differences in
414 extrusion rates between active and quiescent chromatin states, with numbers obtained from Supp
415 Fig. 3.

416

417 Discussion

418

419 We have developed *FactorFinder*, a transcription factor footprinting method for MNase HiChIP
420 data and used it to identify CTCF binding sites with near base-pair resolution. We show that the
421 DNA protection footprints of nucleosomes and transcription factors can be readily distinguished
422 based on pre-ligation fragment size and strand origin and use these features to identify CTCF
423 binding sites. Significance is then assessed through a multinomial approach, which avoids
424 placing distributional assumptions on read counts. Using this method, the median distance
425 between *FactorFinder* peak summits and motif center is 5 bp, with 93% of peak summits
426 identified within 20 bp of a CTCF motif center.

427

428 We then leverage this methodological advance to investigate how chromatin state affects cohesin
429 extrusion dynamics. A close examination of CTCF-protected fragments revealed an additional
430 CTCF-adjacent footprint downstream of the CBS, which we propose represents cohesin given its
431 positioning relative to looping orientation as well as its strong association with both long range
432 interactions and cohesin occupancy. We estimated the frequency with which a CTCF bound
433 locus forms a loop with a downstream CTCF site and found that it varies considerably from CBS
434 to CBS, with a genome-wide range from ~1-10%. This is consistent with recent live-cell imaging
435 work that found that CTCF-mediated loops predominantly exist in the partially extruded state at
436 two studied loci^{13,14}.

437

438 We next sought to characterize how cohesin impacts genome contacts in different chromatin
439 contexts. To this end, we employed our high-resolution *FactorFinder* identified CBS and
440 HiChIP 3D contact information to look at differences in extruded loop size in regions with
441 different chromatin states. We observe an approximately 2-fold increase in extruded loop size
442 comparing quiescent chromatin to active chromatin, and this effect is similarly observed when
443 examining the impact of H3K27ac and RNAPII binding. Our finding that RNAPII binding
444 obstructs cohesin-mediated loop extrusion is consistent with two recent studies that investigated
445 RNAPII's impact on cohesin through RNAPII and enhancer perturbations³⁵ as well as polymer
446 simulations, CTCF depletion, and Wapl knockout experiments³⁶. These substantial differences in
447 average extruded loop size observed for different levels of RNAPII binding and H3K27ac
448 suggest that gene and enhancer activity obstruct cohesin-mediated loop extrusion.

449

450 The obstruction of cohesin by gene and enhancer activity implies a model of CTCF-mediated
451 gene regulation where a fully extruded, stable, and convergent CTCF-CTCF loop is not required
452 for CTCF to mediate enhancer-promoter contacts. Instead, a promoter-proximal CTCF can halt
453 cohesin next to the TSS of a gene while cohesin continues to extrude on the other side,
454 effectively behaving as an enhancer recruiter. Cohesin slowing down through enhancer regions
455 would then enable an enrichment of enhancer-promoter contacts without requiring a stable
456 CTCF-CTCF loop (Fig. 6). This attenuation in cohesin extrusion may also provide a mechanism
457 relating gene regulation to the presence of RNAPII at enhancers³⁷.



458

459 **Fig. 6** Schematic of proposed model whereby single promoter-proximal CTCF sites enable an
460 enrichment of enhancer-promoter contacts.

461

462 The dynamic CTCF-mediated enhancer-promoter contact model proposed here is consistent with
463 recent findings that promoter proximal CTCFs have important roles in gene regulation^{9–11}, that
464 enhancer-promoter contacts are unstable^{38,39}, and that CTCF and cohesin-mediated chromatin
465 loops are dynamic^{13,14}. The dynamic nature of EP contacts has contributed to the development of
466 the “kiss and kick” model⁴⁰ as a potential explanation for how enhancers and promoters come

467 into contact but move away from each other at the time of transcription. Our findings are
468 compatible with the “kiss and kick” model, but additionally suggest a potential mechanism by
469 which distal enhancers can locate gene promoters without being stuck in a stable conformation.
470 This model would use promoter- or enhancer-proximal CTCF sites to enable distal enhancers to
471 both come into contact with gene promoters and subsequently disengage during transcription. In
472 this way, CTCF’s role in long-range enhancer promoter contact would be as a dynamic
473 functional element recruiter instead of mediating continual stable contact between distal
474 enhancers and gene promoters.

475

476 **Materials and methods**

477

478 CTCF MNase HiChIP

479 Four MNase K562 CTCF HiChIP (150 bp paired-end) libraries were generated using the Cantata
480 Bio / Dovetail Genomics MNase HiChIP kit. CTCF MNase HiChIP was performed as described
481 in the Dovetail HiChIP MNase Kit protocol v.2.0. Briefly, 5 million K562 cells per sample were
482 crosslinked with 3mM DSG and 1% formaldehyde and digested with 1ul MNase (“YET”
483 samples) or 2ul MNase (“GW” samples) in 100ul of 1X nuclease digestion buffer. Cells were
484 lysed with 1X RIPA containing 0.1% SDS, and CTCF ChIP was performed using 1500ng of
485 chromatin (40-70% mononucleosomes) and 500 ng of CTCF antibody (Cell Signaling, cat #:
486 3418). Protein A/G beads pull-down, proximity ligation, and library preparation were done
487 according to the protocol. Libraries were sequenced to a read depth of ~172 million paired end
488 reads per sample on the Illumina Nextseq 2000 platform.

489

490 Software implementation

491 Preprocessing, analysis and figure code used in this paper are available at
492 https://github.com/aryeelab/cohesin_extrusion_reproducibility. Data figures in this paper were
493 made in R v.4.1.2 using ggplot.

494

495 Data availability

496 Raw and Processed HiChIP data produced in this study will be uploaded to NCBI GEO (GSE
497 Record ID pending).
498 K562 ChIP-seq RAD21 BED file (Accession ID: ENCFF330SHG), CTCF BED file (Accession
499 ID: ENCFF736NYC), CTCF bigWig signal value (Accession ID: ENCFF168IFW), RNAPII
500 BED file (Accession ID: ENCFF355MNE), and H3K27ac BED file (Accession ID:
501 ENCFF544LXB) were obtained from ENCODE, and CTCF motifs were obtained from the R
502 package *CTCF*²⁷ (annotation record: AH104729, documentation:
503 <https://bioconductor.org/packages/release/data/annotation/vignettes/CTCF/inst/doc/CTCF.html>).

504

505 **Methods**

506 Data Processing

507 4 replicates of K562 MNase CTCF HiChIP data were aligned to the reference genome using the
508 BWA-MEM algorithm⁴¹. Ligation events were then recorded using pairtools parse v. 0.3.0⁴²,
509 PCR duplicates were removed, and the final pairs and bam files were generated. HiChIP loop
510 calls were then made using FitHiChIP Peak to Peak²⁹ with 2.5kb loop anchor bin size. The
511 MNase HiChIP processing protocol is based on guidelines from
https://hichip.readthedocs.io/en/latest/before_you_begin.html. Reproducible code is available at
513 https://github.com/aryeelab/cohesin_extrusion_reproducibility.

514

515 Identification of significant motifs

516 We use CTCF motifs identified as significant ($p < 1e-05$) by *FactorFinder* as the set of CTCF
517 binding sites. This p-value threshold was chosen based on the precision recall curve (Fig. 3b),
518 and corresponds to a maximum FDR q-value of 3e-04.

519

520 Multiple Testing

521 For genome-wide footprinting analysis adjustment for multiple testing, CTCF motifs are
522 assigned the p-value of the closest *FactorFinder* sliding window. The Benjamini-Hochberg
523 method⁴³ was used to obtain q-values.

524

525 Estimating cohesin footprints

526 The cohesin footprint is observed by obtaining motif-level coverage estimates +/- 80 bp around
527 CBS, summing up the coverage across all motifs (within strata), and subtracting the upstream
528 coverage from the downstream (downstream coverage - upstream coverage) at each base pair.
529 Note that downstream and upstream are defined relative to the motif strand, so downstream is to
530 the “left” of CBS (-) and to the “right” of CBS (+) in terms of reference genome base pairs. The
531 aforementioned strata are defined by RAD21 ChIP-seq signal level (high vs low) and whether
532 the fragment is the upstream or downstream interaction partner in its pair. RAD21 ChIP-seq high
533 and low correspond to the top 25% and bottom 25% of ChIP-seq signal value of the adjacent
534 (within 50 bp of CBS) RAD21 ChIP-seq peak. Note that only mid-size (fragment length between
535 80 and 120), long range fragments (interaction length > 10kb) are used for this analysis.

536

537 Estimating the fully extruded state

538 We estimated a genome-wide range for the fully extruded state by obtaining CTCF/cohesin-
539 protected upstream fragments overlapping CBS (+) and estimating the fraction of interaction
540 partners overlapping a downstream convergent negative strand CTCF motif. CBS (+) were
541 required to have at least 50 CTCF/cohesin-protected upstream fragments overlapping the motif
542 to enable sufficient sample size for the motif-specific percent convergent calculation. We then
543 accounted for CTCF occupancy (estimated as ~50%)³⁴ by dividing this estimate by two. The
544 point estimate (5%) is the number of interaction partners overlapping a downstream convergent

545 negative strand CTCF motif genome-wide / the total number of fragments genome-wide, and the
546 range (1-10%) are the 1st and 99th percentile of the CBS-level CTCF-CTCF chromatin loop
547 estimate.

548

549 Determining extruded loop size as a function of chromatin state

550 We used upstream fragments overlapping CTCF binding sites (+) for this analysis. 1 MB regions
551 downstream of the CBS (+) were annotated using ChromHMM³² to quantify the percentage of bp
552 assigned to each of the 15 chromatin states. To simplify annotation, we grouped the 15
553 chromatin states into three categories (quiescent, polycomb/bivalent, and active) based on their
554 correlation (Fig 5c). Regions were clustered using Ward's hierarchical clustering method⁴⁴ (Fig
555 5d.). For extrusion dynamics analyses (Fig 5e,f,h), each of the three chromatin categories was
556 represented by the 20% of regions with the highest fraction of DNA in this state. Extruded loop
557 size was then estimated as the average log10 interaction length for each annotation. Only long
558 range TF-protected fragments (start and end at least 15 bp from the motif center, length < 115,
559 interaction length > 10kb) were included in this estimate.

560

561 Similarly, high/low H3K27ac corresponds to the top 10% and bottom 10% of the number of
562 basepairs covered by H3K27ac ChIP-seq peaks in the 1 MB regions downstream of CBS (+).
563 High/low RNAPII corresponds to the top 10% and bottom 10% of the number of RNAPII ChIP-
564 seq peaks located in the 1 MB regions downstream of CBS (+). Extruded loop size estimates
565 were obtained in the same way for these annotated regions; long range TF-protected fragments
566 were used to estimate the average log10 interaction length.

567

568 Directionality of CBS-adjacent nucleosome position signal

569 Interestingly, the strength of the nucleosome positioning signal is related to the orientation of the
570 DNA contact. Stratifying nucleosome-bound fragments based on whether they are the upstream
571 or downstream long-range (>10kb) fragment in a pair (effectively single-cell left or right loop
572 anchor) produces a differential nucleosome signal inside and outside the loop (Supp Fig. 5). For
573 both upstream and downstream nucleosome-bound fragments, the nucleosome closest to the
574 CTCF binding site and inside the loop exhibits a substantially stronger signal than the closest
575 nucleosome outside the loop. HiChIP ligations are unlikely to fully account for this signal as a
576 previous study using MNase-seq also showed a directional nucleosome preference around CBS
577 (see Fig. 1a), although this result was not noted in the text²⁵.

578

579 **Disclosures**

580

581 Dovetail Genomics/Cantata Bio provided reagents and sample processing for HiChIP
582 experiments. M.B. and M.S.B were employees at Dovetail Genomics during the course of this
583 research. M.J.A has financial and consulting interests unrelated to this work in SeQure Dx and
584 Chroma Medicine. M.J.A's interests are reviewed and managed by Dana Farber Cancer Institute.

585 J.K.J. is a co-founder of and has a financial interest in SeQure, Dx, Inc., a company developing
586 technologies for gene editing target profiling. JKJ also has, or had during the course of this
587 research, financial interests in several companies developing gene editing technology: Beam
588 Therapeutics, Blink Therapeutics, Chroma Medicine, Editas Medicine, EpiLogic Therapeutics,
589 Excelsior Genomics, Hera Biolabs, Monitor Biotechnologies, Nvelop Therapeutics (f/k/a ETx,
590 Inc.), Pairwise Plants, Poseida Therapeutics, and Verve Therapeutics. J.K.J.'s interests were
591 reviewed and are managed by Massachusetts General Hospital and Mass General Brigham in
592 accordance with their conflict of interest policies.

593

594 **Funding**

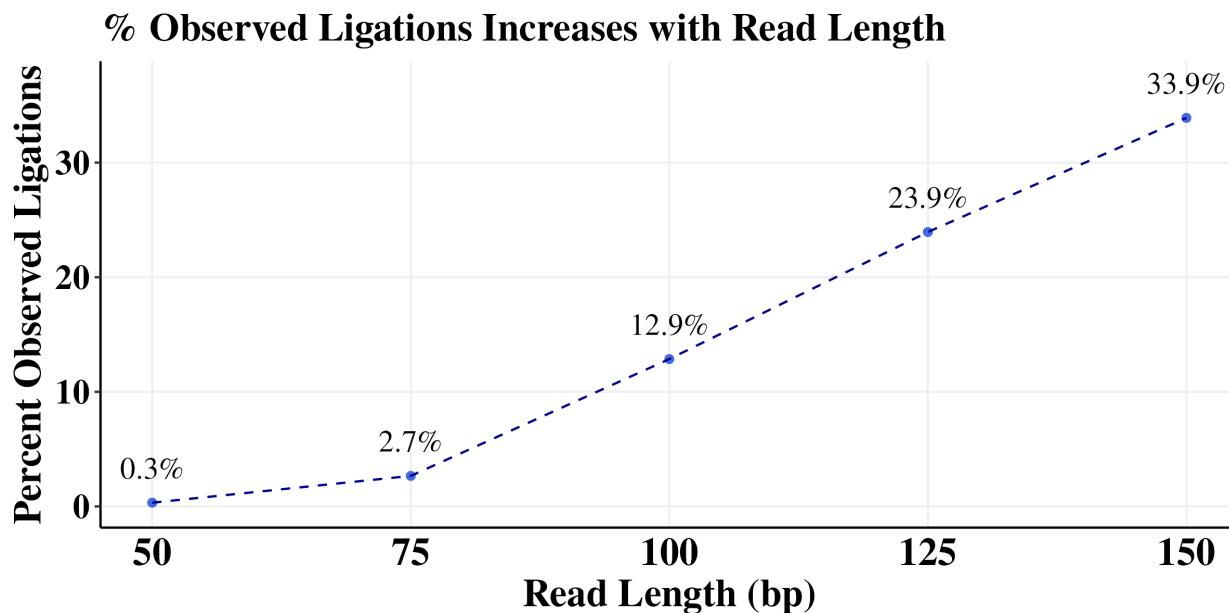
595

596 This work was supported by the National Institutes of Health grants RM1HG009490 (MJA, JKJ,
597 CS), R35GM118158 (JKJ), T32GM135117 (CS), and a Career Development Award from the
598 American Society of Gene & Cell Therapy (YET). The content is solely the responsibility of the
599 authors and does not necessarily represent the official views of the American Society of Gene &
600 Cell Therapy. Dovetail Genomics / Cantata Bio supported data generation costs.

601

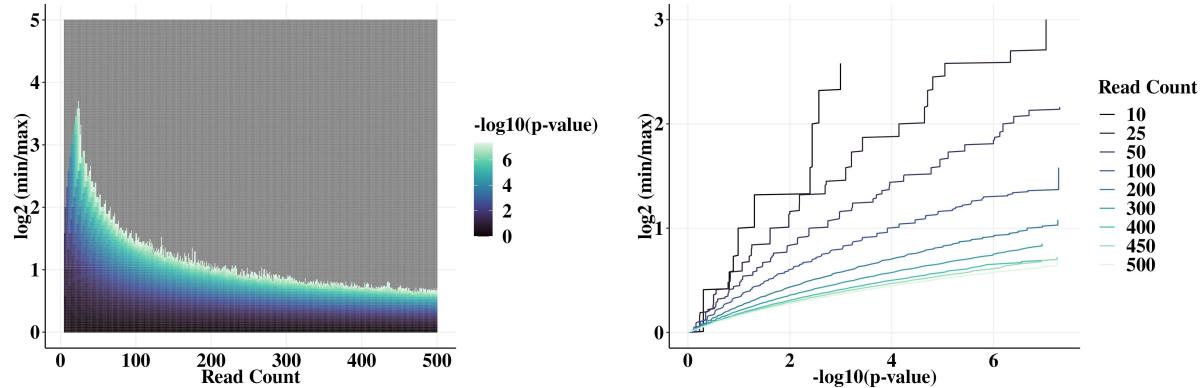
602 **Supplementary Figures**

603



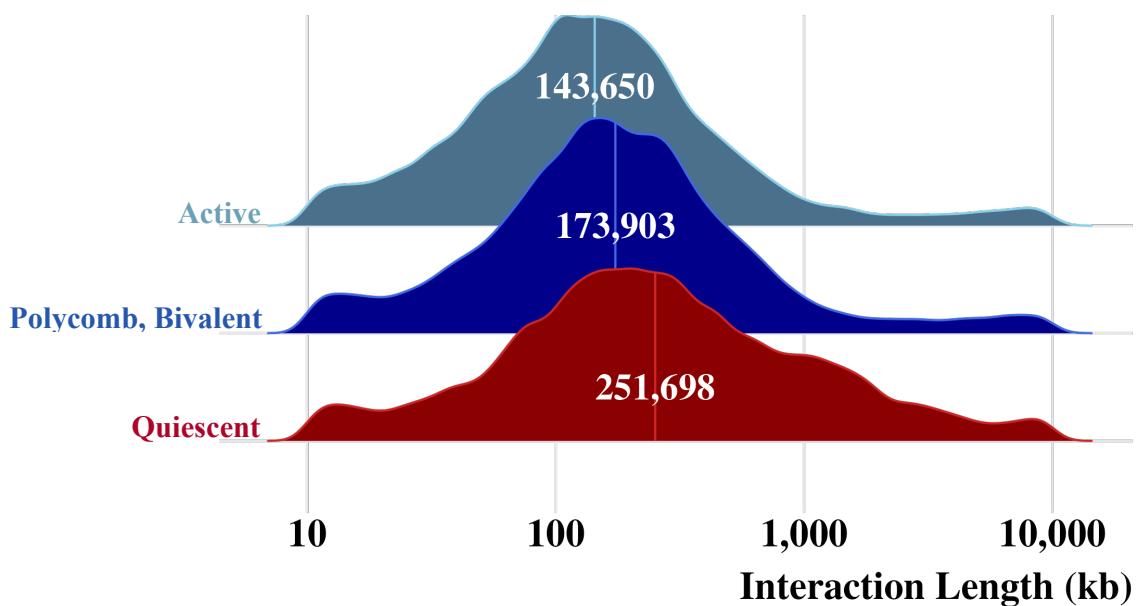
604

605 **Supplementary Figure 1.** Percent observed ligations increases with read length.



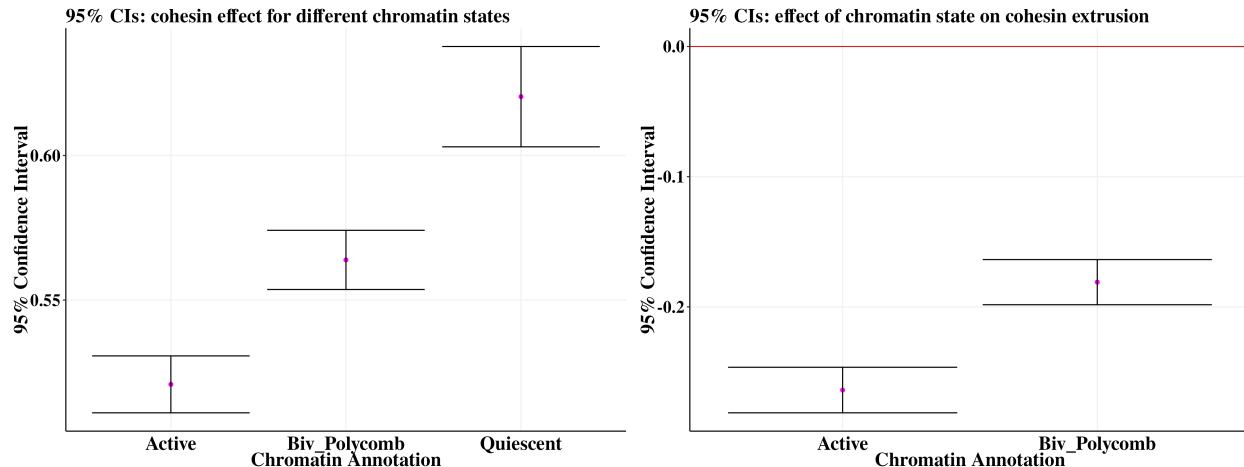
606
607
608
609

Supplementary Figure 2. The probability of observing a high *FactorFinder* statistic under the null hypothesis is higher at low read counts.



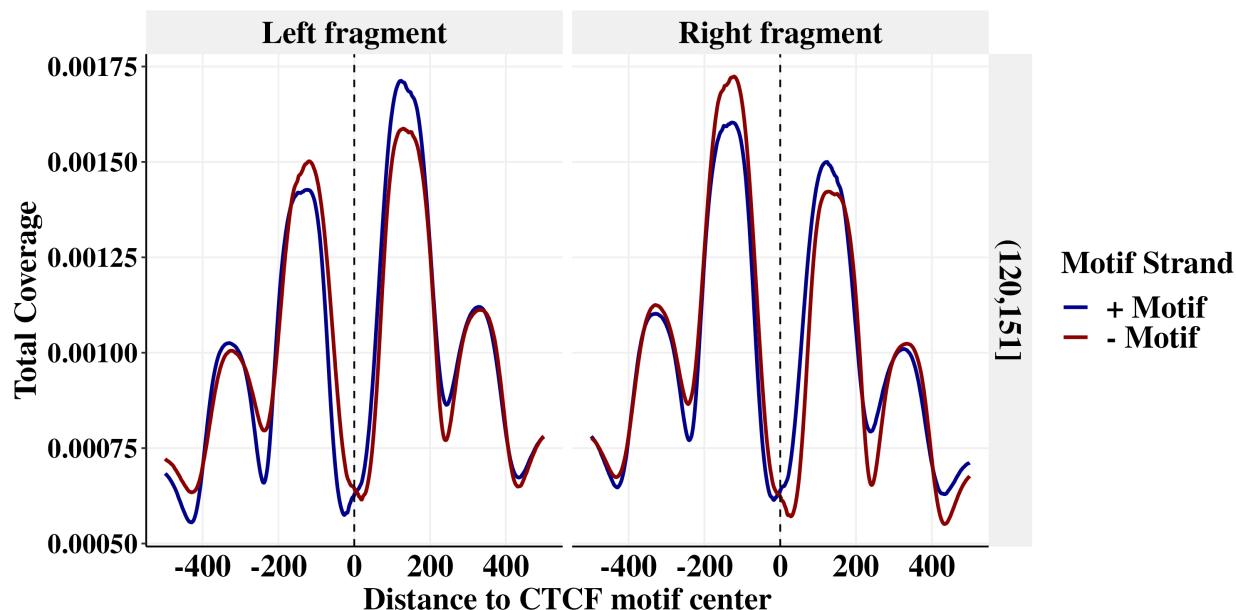
610
611
612
613

Supplementary Figure 3. Cohesin extrudes significantly further through quiescent regions than active regions.



614
615
616
617
618
619

Supplementary Figure 4. Controlling for locus-specific variation with linear mixed models does not attenuate the relationship between chromatin state and extruded loop size. Note that for the figure on the right, the group that active and bivalent polycomb are being compared to is quiescent.



620
621

Supplementary Figure 5. Nucleosomes are preferentially positioned inside the loop.

622

623 References

624
625
626
627
628

1. Fudenberg, G. *et al.* Formation of Chromosomal Domains by Loop Extrusion. *Cell Rep.* **15**, 2038–2049 (2016).
2. Li, Y. *et al.* The structural basis for cohesin–CTCF-anchored loops. *Nature* **578**, 472–476 (2020).

629 3. Xiao, J. Y., Hafner, A. & Boettiger, A. N. How subtle changes in 3D structure can create
630 large changes in transcription. *eLife* **10**, e64320.

631 4. Zuin, J. *et al.* Nonlinear control of transcription through enhancer–promoter interactions.
632 *Nature* **604**, 571–577 (2022).

633 5. Lupiáñez, D. G. *et al.* Disruptions of Topological Chromatin Domains Cause Pathogenic
634 Rewiring of Gene-Enhancer Interactions. *Cell* **161**, 1012–1025 (2015).

635 6. Flavahan, W. A. *et al.* Insulator dysfunction and oncogene activation in IDH mutant gliomas.
636 *Nature* **529**, 110–114 (2016).

637 7. Hirayama, T., Tarusawa, E., Yoshimura, Y., Galjart, N. & Yagi, T. CTCF Is Required for
638 Neural Development and Stochastic Expression of Clustered Pcdh Genes in Neurons. *Cell*
639 *Rep.* **2**, 345–357 (2012).

640 8. Katainen, R. *et al.* CTCF/cohesin-binding sites are frequently mutated in cancer. *Nat. Genet.*
641 **47**, 818–821 (2015).

642 9. Kubo, N. *et al.* Promoter-proximal CTCF-binding promotes long-range-enhancer dependent
643 gene activation. *Nat. Struct. Mol. Biol.* **28**, 152–161 (2021).

644 10. Schuijers, J. *et al.* Transcriptional Dysregulation of MYC Reveals Common Enhancer-
645 Docking Mechanism. *Cell Rep.* **23**, 349–360 (2018).

646 11. Cerdá-Smith, C. G. *et al.* Integrative PTEN Enhancer Discovery Reveals a New Model of
647 Enhancer Organization. 2023.09.20.558459 Preprint at
648 <https://doi.org/10.1101/2023.09.20.558459> (2023).

649 12. Davidson, I. F. *et al.* DNA loop extrusion by human cohesin. *Science* **366**, 1338–1345
650 (2019).

651 13. Gabriele, M. *et al.* Dynamics of CTCF- and cohesin-mediated chromatin looping revealed by
652 live-cell imaging. *Science* **376**, 496–501 (2022).

653 14. Mach, P. *et al.* Cohesin and CTCF control the dynamics of chromosome folding. *Nat. Genet.*
654 **54**, 1907–1918 (2022).

655 15. Kasinathan, S., Orsi, G. A., Zentner, G. E., Ahmad, K. & Henikoff, S. High-resolution
656 mapping of transcription factor binding sites on native chromatin. *Nat. Methods* **11**, 203–209
657 (2014).

658 16. Henikoff, J. G., Belsky, J. A., Krassovsky, K., MacAlpine, D. M. & Henikoff, S. Epigenome
659 characterization at single base-pair resolution. *Proc. Natl. Acad. Sci.* **108**, 18318–18323
660 (2011).

661 17. Skene, P. J. & Henikoff, S. An efficient targeted nuclease strategy for high-resolution
662 mapping of DNA binding sites. *eLife* **6**, e21856 (2017).

663 18. Gutin, J. *et al.* Fine-Resolution Mapping of TF Binding and Chromatin Interactions. *Cell*
664 *Rep.* **22**, 2797–2807 (2018).

665 19. Hua, P. *et al.* Defining genome architecture at base-pair resolution. *Nature* **595**, 125–129
666 (2021).

667 20. Krietenstein, N. *et al.* Ultrastructural Details of Mammalian Chromosome Architecture. *Mol.*
668 *Cell* **78**, 554–565.e7 (2020).

669 21. Hsieh, T.-H. S. *et al.* Resolving the 3D Landscape of Transcription-Linked Mammalian
670 Chromatin Folding. *Mol. Cell* **78**, 539-553.e8 (2020).

671 22. Goel, V. Y., Huseyin, M. K. & Hansen, A. S. Region Capture Micro-C reveals coalescence
672 of enhancers and promoters into nested microcompartments. *Nat. Genet.* 1–9 (2023)
673 doi:10.1038/s41588-023-01391-1.

674 23. Mumbach, M. R. *et al.* HiChIP: efficient and sensitive analysis of protein-directed genome
675 architecture. *Nat. Methods* **13**, 919–922 (2016).

676 24. Chereji, R. V., Bryson, T. D. & Henikoff, S. Quantitative MNase-seq accurately maps
677 nucleosome occupancy levels. *Genome Biol.* **20**, 198 (2019).

678 25. Fu, Y., Sinha, M., Peterson, C. L. & Weng, Z. The Insulator Binding Protein CTCF Positions
679 20 Nucleosomes around Its Binding Sites across the Human Genome. *PLOS Genet.* **4**,
680 e1000138 (2008).

681 26. Hashimoto, H. *et al.* Structural basis for the versatile and methylation-dependent binding of
682 CTCF to DNA. *Mol. Cell* **66**, 711-720.e3 (2017).

683 27. Dozmorov, M. G. *et al.* CTCF: an R/bioconductor data package of human and mouse CTCF
684 binding sites. *Bioinforma. Adv.* **2**, vbac097 (2022).

685 28. Luo, Y. *et al.* New developments on the Encyclopedia of DNA Elements (ENCODE) data
686 portal. *Nucleic Acids Res.* **48**, D882–D889 (2020).

687 29. Bhattacharyya, S., Chandra, V., Vijayanand, P. & Ay, F. Identification of significant
688 chromatin contacts from HiChIP data by FitHiChIP. *Nat. Commun.* **10**, 4221 (2019).

689 30. Bailey, T. L. STREME: accurate and versatile sequence motif discovery. *Bioinformatics* **37**,
690 2834–2840 (2021).

691 31. Gassler, J. *et al.* A mechanism of cohesin-dependent loop extrusion organizes zygotic
692 genome architecture. *EMBO J.* **36**, 3600–3618 (2017).

693 32. Ernst, J. & Kellis, M. Chromatin-state discovery and genome annotation with ChromHMM.
694 *Nat. Protoc.* **12**, 2478–2492 (2017).

695 33. Hoffman, M. M. *et al.* Integrative annotation of chromatin elements from ENCODE data.
696 *Nucleic Acids Res.* **41**, 827–841 (2013).

697 34. Cattoglio, C. *et al.* Determining cellular CTCF and cohesin abundances to constrain 3D
698 genome models. *eLife* **8**, e40164 (2019).

699 35. Barshad, G. *et al.* RNA polymerase II dynamics shape enhancer–promoter interactions. *Nat.*
700 *Genet.* 1–11 (2023) doi:10.1038/s41588-023-01442-7.

701 36. Banigan, E. J. *et al.* Transcription shapes 3D chromatin organization by interacting with loop
702 extrusion. *Proc. Natl. Acad. Sci.* **120**, e2210480120 (2023).

703 37. Arnold, P. R., Wells, A. D. & Li, X. C. Diversity and Emerging Roles of Enhancer RNA in
704 Regulation of Gene Expression and Cell Fate. *Front. Cell Dev. Biol.* **7**, (2020).

705 38. Transcription decouples estrogen-dependent changes in enhancer-promoter contact
706 frequencies and spatial proximity | bioRxiv.
707 <https://www.biorxiv.org/content/10.1101/2023.03.29.534720v2.abstract>.

708 39. Benabdallah, N. S. *et al.* Decreased Enhancer-Promoter Proximity Accompanying Enhancer

709 Activation. *Mol. Cell* **76**, 473–484.e7 (2019).

710 40. Pownall, M. E. *et al.* Chromatin expansion microscopy reveals nanoscale organization of
711 transcription and chromatin. *Science* **381**, 92–100 (2023).

712 41. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler
713 transform. *Bioinformatics* **25**, 1754–1760 (2009).

714 42. Abdennur, N. *et al.* Pairtools: from sequencing data to chromosome contacts. *bioRxiv*
715 2023.02.13.528389 (2023) doi:10.1101/2023.02.13.528389.

716 43. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and
717 Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B Methodol.* **57**, 289–300
718 (1995).

719 44. Murtagh, F. & Legendre, P. Ward’s Hierarchical Agglomerative Clustering Method: Which
720 Algorithms Implement Ward’s Criterion? *J. Classif.* **31**, 274–295 (2014).