

## Pervasive and programmed nucleosome distortion patterns on single mammalian chromatin fibers

Marty G Yang<sup>1</sup>, Hannah J Richter<sup>1,9</sup>, Simai Wang<sup>1,9</sup>, Colin P McNally<sup>1</sup>, Nicole Harris<sup>1</sup>, Simaron Dhillon<sup>3,6</sup>, Michela Maresca<sup>2</sup>, Elzo de Wit<sup>2</sup>, Holger Willenbring<sup>3,4,5</sup>, Jacquelyn Maher<sup>3,6</sup>, Hani Goodarzi<sup>7,8#</sup>, Vijay Ramani<sup>1,7,10#</sup>

### AFFILIATIONS

<sup>1</sup>: Gladstone Institute of Data Science and Biotechnology, Gladstone Institutes, San Francisco, CA 94158

<sup>2</sup>: Division of Gene Regulation, The Netherlands Cancer Institute, Amsterdam, Netherlands

<sup>3</sup>: Liver Center, University of California, San Francisco, San Francisco, CA 94143

<sup>4</sup>: Division of Transplant Surgery, Department of Surgery, University of California, San Francisco, San Francisco, CA, 94143

<sup>5</sup>: Eli and Edythe Broad Center of Regeneration Medicine and Stem Cell Research, University of California San Francisco, San Francisco, CA, 94143

<sup>6</sup>: Division of Gastroenterology, Department of Medicine, University of California San Francisco, San Francisco, CA, 94143

<sup>7</sup>: Department of Biochemistry and Biophysics, University of California San Francisco, San Francisco, CA 94158

<sup>8</sup>: Arc Institute, Palo Alto, CA 94304

<sup>9</sup>: these authors contributed equally

<sup>10</sup>: lead contact

#: correspondence to [hani.goodarzi@ucsf.edu](mailto:hani.goodarzi@ucsf.edu) & [vijay.ramani@gladstone.ucsf.edu](mailto:vijay.ramani@gladstone.ucsf.edu)

### ABSTRACT

We present a genome-scale method to map the single-molecule co-occupancy of structurally distinct nucleosomes, subnucleosomes, and other protein-DNA interactions via long-read high-resolution adenine methyltransferase footprinting. Iteratively Defined Lengths of Inaccessibility (IDLI) classifies nucleosomes on the basis of shared patterns of intranucleosomal accessibility, into: i.) minimally-accessible chromatosomes; ii.) octasomes with stereotyped DNA accessibility from superhelical locations (SHLs)  $\pm 1$  through  $\pm 7$ ; iii.) highly-accessible unwrapped nucleosomes; and iv.) subnucleosomal species, such as hexasomes, tetrasomes, and other short DNA protections. Applying IDLI to mouse embryonic stem cell (mESC) chromatin, we discover widespread nucleosomal distortion on individual mammalian chromatin fibers, with  $>85\%$  of nucleosomes surveyed displaying degrees of intranucleosomally accessible DNA. We observe epigenomic-domain-specific patterns of distorted nucleosome co-occupancy and positioning, including at enhancers, promoters, and mouse satellite repeat sequences. Nucleosome distortion is programmed by the presence of bound transcription factors (TFs) at cognate motifs; occupied TF binding sites are differentially decorated by distorted nucleosomes compared to unbound sites, and degradation experiments establish direct roles for TFs in structuring binding-site proximal nucleosomes. Finally, we apply IDLI in the context of primary mouse hepatocytes, observing evidence for pervasive nucleosomal distortion *in vivo*. Further genetic experiments reveal a role for the hepatocyte master regulator FOXA2 in directly impacting nucleosome distortion at hepatocyte-specific regulatory elements *in vivo*. Our work suggests extreme—but regulated—plasticity in nucleosomal DNA accessibility at the single-molecule level. Further, our study offers an essential new framework to model transcription factor binding, nucleosome remodeling, and cell-type specific gene regulation across biological contexts.

## 1 INTRODUCTION

2 Regulated DNA accessibility governs cell-type specific transcription. DNA accessibility is dictated in part by the  
3 octameric nucleosome, which compacts ~147 basepairs (bp) of DNA around two dimers of histone proteins H2A  
4 and H2B flanking a tetramer of histone proteins H3 and H4. Our canonical understanding of how chromatin  
5 accessibility regulates specificity of protein-DNA interactions is simple: nucleosomes and positive transcriptional  
6 regulatory factors compete with one another for DNA binding, particularly at active *cis*-regulatory elements<sup>1</sup>.  
7 Thus, nucleosome occupancy negatively regulates transcription, and multiple essential coactivators act by  
8 modulating nucleosome occupancy directly via chromatin remodeling.

9  
10 Work over the last three decades has challenged the simplicity of this model. The synthesis of genome-scale  
11 mapping techniques and high-resolution structural biology has revealed ‘altered’ nucleosomal structures across  
12 eukaryotic epigenomes, including hexasomes (*i.e.* nucleosomes missing a single H2A/H2B dimer)<sup>2-4</sup>, overlapping  
13 dinucleosomes (*i.e.* hexasome-octasome complexes)<sup>5-7</sup>, and other noncanonical histone-containing complexes<sup>8</sup>.  
14 Work from the groups of Widom and others has illustrated that primary DNA sequence can guide occupancy and  
15 modulate the stability of nucleosome-wrapped DNA<sup>9-11</sup>. ATP-dependent chromatin remodeling can rely on  
16 structural distortion of nucleosomes (for the imitation switch [ISWI] remodeler SNF2h<sup>12</sup>), or the presence of  
17 hexasomes (for INO80<sup>13,14</sup>). Thus, dynamic nucleosome composition and structure are physiologically-relevant:  
18 nuclear regulatory processes—including ATP-dependent chromatin remodeling, transcription, replication, and  
19 histone chaperone activity—generate these species and render them important marks of genomic activity. Perhaps  
20 most importantly, all of these species must ultimately compete with essential sequence-specific transcriptional  
21 factors (TFs) to determine cell-type-specific chromatin accessibility and gene transcription<sup>15</sup>. Holistically  
22 understanding how nucleosomes are structured and positioned with respect to TFs necessitates high-resolution *in*  
23 *vivo* measurement of nucleosomal structures at TF binding sites across diverse contexts.

24  
25 Technical limitations have confounded accurate assessment of canonical and noncanonical nucleosome structures  
26 *in vivo*. Highly resolved *in situ* imaging techniques (*e.g.* [cryo]electron tomography<sup>16,17</sup>; chromatin expansion  
27 microscopy<sup>18</sup>) are blind to DNA sequence, while standard genomic mapping techniques rely on digestion of the  
28 chromatin fiber<sup>3,4,19,20</sup>. Digestion is especially problematic in the context of mapping subnucleosomal particles: a  
29 hexasome-sized nucleolytic fragment, for instance, could either derive from a true hexasome, from a remodeled  
30 nucleosome with increased DNA accessibility between a single H2A/H2B dimer and the H3/H4 tetramer, or from  
31 transient unwrapping of DNA at nucleosomal entry / exit sites (*i.e.* nucleosomal breathing)<sup>21,22</sup>. Short read single-  
32 molecule methyltransferase (MTase) footprinting methods (*e.g.* NoME-seq<sup>23</sup>; MAPit<sup>24</sup>; dSMF<sup>25</sup>) solve some of  
33 these issues but are themselves limited by i.) the maximum read-length (600 bp) possible on an Illumina  
34 sequencer, and ii.) use of MTases with specificity for GpC or CpG dinucleotides. Single-molecule long-read  
35 footprinting methods that use promiscuous adenine MTases (*e.g.* the single-molecule adenine methylated  
36 oligonucleosome sequencing assay [SAMOSA]<sup>26,27</sup>; Fiber-seq<sup>28</sup>) theoretically solve many of these limitations,  
37 but to date have not been applied towards systematically interrogating noncanonical nucleosome structure.

38  
39 Here, we build on computational methods previously developed by our group<sup>29</sup> to nondestructively resolve—for  
40 the first time—the subunit architecture of chromatin fibers down to individual H2A/H2B dimers and H3/H4  
41 tetramers. Harnessing this ability to map structurally distinct classes of nucleosomes on chromatin fibers genome-  
42 wide *in vivo*, we make a startling discovery: pervasive, but regulated patterns of ‘distorted’ nucleosomes with  
43 varying degrees of accessible nucleosomal DNA, across both active and inactive mammalian epigenomic regions.  
44 Acute depletion of transcription factors (CTCF; SOX2) in mouse embryonic stem cells (mESCs) reveals distinct  
45 modes by which essential trans-acting factors distort and position nucleosomes on chromatin fibers. Extension of  
46 our approach to data generated in primary mouse hepatocytes reveals that widespread nucleosome distortion  
47 occurs *in vivo*. By genetically perturbing the master regulatory factor FOXA2 in hepatocytes *in vivo*, we  
48 demonstrate a direct role for this ‘pioneer’ factor in positioning and distorting nucleosomes at FOXA2-bound  
49 regulatory regions. Taken as a whole, our study provides a sequence-resolved foundation for quantifying and  
50 studying causes and consequences of nucleosome structural heterogeneity genome-wide.

## 1 RESULTS

### 2 **Footprinting the subunit architecture of the nucleosome core particle on individual chromatin fibers**

3 We previously introduced the single-molecule adenine methylated oligonucleosome sequencing assay  
4 (SAMOSA), which combines genome-wide chromatin footprinting by nonspecific adenine MTase EcoGII with  
5 native detection of methylated adenine via long-read PacBio sequencing<sup>26</sup>. Inferring protein-DNA contacts from  
6 EcoGII methylation is nontrivial, and we later introduced a neural network – hidden markov model (NN-HMM)  
7 that explicitly accounts for: i.) sequence biases of the EcoGII enzyme, and ii.) sequence biases influencing the  
8 kinetics of PacBio sequencing, to infer nucleosome ‘footprint’ positions along sequenced DNA molecules<sup>29</sup>. Our  
9 original approach – which has successfully mapped chromatin *in vivo*<sup>27</sup>, *in vitro*<sup>29,30</sup>, and in the context of DNA  
10 replication<sup>31</sup> – was calibrated to detect mononucleosomes, the predominant feature of chromatin.

11  
12 We hypothesized that our model – which uses an NN to define emission probabilities and requires only a single,  
13 user-defined parameter (the transition probability; hereafter, ‘*t*’) within a two-state HMM – might be able to detect  
14 the subunit structure of mononucleosomes (**Figure 1A**). To test this hypothesis, we collated a mixture of published  
15 and unpublished SAMOSA datasets for the E14 mESC line (44 sequencing libraries to yield 88.87 Gb of  
16 sequenced mESC DNA; library metadata summarized in **Supplementary Table 1**), and called footprints varying  
17 the *t* parameter (which represents the likelihood of switching between hidden ‘inaccessible’ and ‘accessible’  
18 model states). Plotting the distributions of called footprint sizes across a wide range of *t* values confirmed that  
19 increasing *t* resulted in successively smaller footprint calls (hereafter referred to as ‘subfootprints’; **Figure 1B**).  
20 We reasoned that if subfootprints were biologically ‘patterned,’ then due to the two-fold symmetry of the  
21 nucleosome about the dyad, accessibility surrounding subfootprints should symmetrically shift when plotted  
22 about the center of the originating footprint (due to nucleosomal ‘breathing’ at the entry / exit site<sup>32</sup>); further, we  
23 reasoned that these footprints, when plotted as footprint size versus midpoint position ‘horizon plots’ (analogous  
24 to V-plots for micrococcal nuclease sequencing [MNase-seq] data<sup>19</sup>), would reflect the composition of the  
25 nucleosome, including H2A/H2B dimers, H3/H4 tetramer, hexasomes, and “breathing” nucleosomes (**Figure 1C-**  
26 **D**).

27  
28 Visualization of data in both ways confirmed that our analysis of subfootprints captured the expected  
29 intranucleosomal accessibility and patterns of protein-DNA contacts from nucleosomes *in vivo*. At four different  
30 *t* values, accessibility demonstrated a symmetric pattern about called footprint midpoints (**Figure 1E**). Most  
31 strikingly, horizon plot visualization of subfootprints revealed clear signal enrichment at precise positions (**Figure**  
32 **1F**), indicative of subfootprinting of breathing octasomes (vertical line; feature 1), hexasomes (feature 2),  
33 tetramers (feature 3), and dimers (feature 4). Importantly, this result was quantitatively robust across multiple  
34 biological replicates (**Extended Data Figure 1A-B**; median correlation  $\pm$  median absolute deviation [MAD] =  
35  $0.906 \pm 0.010$ ), and robust to per-sample methylation amount (measured through the mononucleosome-  
36 dinucleosome ratio [MDR]<sup>33</sup>; see **Methods**; **Extended Data Figure 1C**). Finally, accessibility and footprint /  
37 subfootprint positions on individual fibers were easily visualized along with bulk ATAC-seq data<sup>34</sup> at the *Sox2*  
38 locus (**Figure 1G**; **Extended Data Figure 1D**). Taken together, these results demonstrate our ability to footprint  
39 the subunit architecture of the nucleosome on individual chromatin fibers in a mammalian epigenome. Hereafter,  
40 we refer to this computational strategy—dependent on iteratively scoring footprinted molecules—as Iteratively  
41 Defined Lengths of Inaccessibility, or IDLI.

### 42 **The mammalian epigenome is composed of a constellation of structurally distinct nucleosomes**

43 We next sought to evaluate structural heterogeneity as defined by IDLI, at the resolution of individual nucleosome  
44 footprints. We devised a classification approach (**Extended Data Figure 2A**) that leverages the Leiden  
45 community detection<sup>35</sup> and UMAP visualization<sup>36</sup> algorithms to classify and visualize structurally distinct  
46 nucleosome footprints defined by IDLI. We applied this approach to 795,765 sampled nucleosomes from 145,065  
47 sampled E14 mESC chromatin fibers falling within 8 distinct epigenomic regions, including both mappable  
48 domains of specific post-translationally modified histones and typically ‘unmappable’ domains comprised of low-  
49 complexity repeat sequence (mouse major satellite, minor satellite, and telomeric sequence). We further binned  
50

1 clusters into ‘groups’ through hierarchical clustering (**Extended Data Figure 2B**) and examined the underlying  
2 properties of each cluster of nucleosomes. Specifically, we examined three properties (**Extended Data Figure**  
3 **2C**): i.) per-cluster footprint sizes; ii.) per-cluster horizon plots; and iii.) average accessibility of footprints at the  
4 most permissive calculated  $t$  value used for clustering (*i.e.*,  $t = 0.071$ ). Our joint analysis and visualization pipeline  
5 (UMAP colored by cluster shown in **Figure 2A**) resulted in 14 distinct structural classes (*i.e.* ‘nucleosome types;’  
6 **Figure 2B**), which fell into seven larger groups.

7  
8 Nucleosome types displayed patterns consistent with known structural features of chromatin fibers (summarized  
9 in key for **Figure 2A**). Group I types were consistent with chromatosomes (*i.e.* nucleosome core particles  
10 associated with linker histone H1), owing to their long footprint sizes and positioning of a short subfootprint in  
11 the horizon plot, mirroring the expected additional DNA protection by an H1 molecule; group V, VI, and VII  
12 clusters, conversely, were the most accessible footprinted particles, demonstrating patterns consistent with  
13 breathable nucleosomes (*i.e.* nucleosomes with a full complement of histone proteins and accessibility at  
14 superhelical location [SHL]  $\pm 6-7$ ), hexasomes, tetrasomes, and smaller subnucleosomal particles (*e.g.* H2A/H2B  
15 dimers; transcription factors), respectively. Finally, we also observed ‘intermediate’ clusters, which demonstrated  
16 unique, focal patterns of accessibility within the nucleosome core particle, including at SHL-2 (**Figure 2C**) and  
17 at SHL-3.5 (**Figure 2D**). We speculate that these states may reflect specific nucleosome-cofactor interactions (*e.g.*  
18 interaction between remodelers<sup>37</sup> or high-mobility group [HMG] proteins<sup>38</sup> and SHL-2), or reflect the dynamic  
19 assembly/disassembly process between H2A/H2B dimers and the H3/H4 tetramer. The relative abundance of  
20 nucleosome types was concordant over multiple independent SAMOSA footprinting experiments (**Extended**  
21 **Data Figure 2D**; median  $\pm$  MAD Pearson’s  $r = 0.978 \pm 0.024$ ), and nucleosome type classification was highly  
22 robust to the number of nucleosomes sampled (**Extended Data Figure 2E**). Together these experiments suggest  
23 that the majority (86.6% of footprints falling in groups II – VII) of nucleosomes in E14 mESCs display some  
24 degree of ‘intranucleosomal’ accessibility, implying pervasive nucleosome distortion epigenome-wide.

25  
26 We hypothesized that nucleosome types reflect a previously unseen layer of chromatin-based regulation. To test  
27 this hypothesis, we quantified differences in the distribution of nucleosome types across the E14 mESC  
28 epigenome via a series of Fisher’s Exact Enrichment tests, and visualized resultant data as a heatmap of False  
29 Discovery Rate (FDR) corrected effect sizes (**Figure 2E**). Strikingly, we found that distinct nucleosome types  
30 were significantly depleted or enriched across a wide range of both mappable epigenomic domains (defined by  
31 overlap with published histone modification ChIP-seq data) and unmappable epigenomic regions. In predicted  
32 promoter regions (defined by H3K4me3), for instance, we observed significant enrichment for clusters 13 and 14  
33 (indicative of two types of short subnucleosomal protection; cluster 13 odds ratio [C13 O.R.] = 1.670,  $q = 1.16$   
34  $\times 10^{-294}$ ; C14 O.R. = 1.856,  $q < 1.79 \times 10^{-308}$ ). In mouse minor satellite sequence (representing the centromeric  
35 region), we observed a significant enrichment for clusters 2 and 7 (indicative of chromatosomes and nucleosomes  
36 accessible at SHL3.5; C2 O.R. = 2.468,  $q < 1.79 \times 10^{-308}$ ; C7 O.R. = 2.559,  $q < 1.79 \times 10^{-308}$ ). Importantly, these  
37 enrichment patterns were quantitatively robust over multiple independent replicate SAMOSA experiments  
38 (**Extended Data Figure 2F**), suggesting that these are *bona fide* biological differences in the representation of  
39 nucleosome types across the epigenome. These analyses suggest that epigenome-specific processes regulate  
40 nucleosome types.

#### 41 42 **Single-molecule co-occupancy of structurally distorted nucleosomes**

43 A critical advantage of single-molecule footprinting approaches is their unique ability to measure the molecular  
44 co-occurrence of protein-DNA interactions<sup>39,40</sup>. Owing to the multi-kilobase length-scale of SAMOSA  
45 footprinting data, IDLI extends such analyses to the co-occupancy of several nucleosomes per fiber. To take  
46 advantage of this, we next assessed whether nucleosome types display statistically enriched co-occurrence at  
47 single-molecule resolution. We implemented an analysis pipeline (**Extended Data Figure 3A**), wherein we  
48 computed the likelihood that trinucleosome stretches contain a greater number of co-occurring nucleosome groups  
49 than expected by chance. We visualized test results in two ways: first, as a heatmap of effect sizes (**Extended**  
50 **Data Figure 3B**), and second, as waterfall plots of these effect sizes, ranked and colored by FDR-corrected

1 statistical significance (**Extended Data Figure 3C**). Our analysis revealed striking and significant patterns of  
2 occupancy at the ‘group’ level, including a marked enrichment of ‘homotypic’ arrangements specifically within  
3 active versus repressed epigenomic domains (*e.g.* significant enrichment of VI-VI-VI trinucleosomes in  
4 H3K4me1 chromatin; O.R. = 3.189,  $q = 4.48 \times 10^{-3}$ ). Our analysis also uncovered triplet arrangements that were  
5 found enriched on both active and facultatively repressed chromatin, including a symmetric pattern of IV-V-IV  
6 (cartoon shown in **Extended Data Figure 3D**). Single-molecule visualization of fibers falling within each of four  
7 different chromatin types (**Extended Data Figure 3E**) highlight the flexibility of these analyses, while illustrating  
8 trinucleosomal co-occurrence motifs at single-molecule resolution and at any desired sequence genome-wide.  
9 Together, these analyses highlight the technical utility of nucleosome classification in uncovering context-specific  
10 structural motifs on chromatin fibers. Further, our proof-of-concept reveals specific co-occurrence patterns that  
11 are both unique and shared across diverse epigenomic regions.

### 13 **The interplay between nucleosome type, transcription factors, and DNA sequence**

14 We next sought to study nucleosome types in the context of TFs and their cognate DNA sequence motifs. We  
15 reasoned that the position and identity of nucleosome types in reference to TFs could be analyzed in two  
16 complementary ways (**Figure 3A**): first, we could assess the relative abundance of nucleosome types surrounding  
17 bound TF motifs compared to unbound controls (*i.e.* a ‘footprint-centric’ view of nucleosome structure analogous  
18 to analyses above); second, we could assess the relative abundance and positioning of IDLI footprints /  
19 subfootprints centered at motifs themselves (*i.e.* a ‘motif-centric’ view of nucleosome distortion). These analyses  
20 relate different structural aspects: footprint-centric views are agnostic to motif position but provide a holistic view  
21 of nucleosome structure encompassing motifs (**Figure 3B**); motif-centric views quantify how DNA sequence  
22 positions footprints, but only capture the substructure of regions of a footprint overlapping with the motif (**Figure**  
23 **3C**). Interpreting IDLI footprints in both ways is essential, as how TFs engage their motifs in the context of  
24 nucleosomes remains contentious<sup>15,41,42</sup>. Finally, to query whether TFs directly control nucleosome types, we  
25 incorporated small-molecule sensitive degron perturbation<sup>43</sup> within this experimental framework (**Figure 3D**).

26  
27 We applied our approach first to the essential TF CCCTC-binding factor (CTCF), which plays diverse roles within  
28 the nucleus and has a number of features that make it readily amenable for this analysis: its long residence time<sup>44</sup>,  
29 ease-of-footprinting<sup>45</sup>, and well-defined set of E14 mESC binding sites<sup>46</sup>. Importantly, our analysis of CTCF  
30 serves as a valuable control to validate our ability to detect different nucleosome configurations (and potentially  
31 directly footprint TF binding itself) at sequence-specific recognition motifs. After integrating single-molecule  
32 footprinting data from CTCF degron experiments<sup>33,47</sup> (reproducibility of degradation experiments summarized in  
33 **Extended Data Figure 4A-E**), we clustered IDLI footprints using both footprint-centric (nucleosome types) and  
34 motif-centric pipelines, resulting in 13 (**Extended Data Figure 4F**; F1-F13) and 17 clusters (**Extended Data**  
35 **Figure 4G**; M1-M17), respectively, which we further grouped through hierarchical clustering (**Figure 3E-F**).  
36 Cluster abundances for both footprint-centric and motif-centric CTCF analyses were highly reproducible  
37 (**Extended Data Figure 4H-I**).

38  
39 Footprint-centric clusters qualitatively matched nucleosome types defined in epigenome-wide analyses,  
40 demonstrating patterns consistent with chromatosomes, intranucleosomally accessible particles, breathable  
41 nucleosomes, subnucleosomes, and short protections (**Extended Data Figure 4F**). We computed the extent to  
42 which these nucleosome types differed at ChIP-backed CTCF motifs versus random motif matches unlikely to  
43 bind CTCF (**Figure 3G**; reproducibility of enrichment shown in **Extended Data Figure 4J**). We observed  
44 significant enrichment of subnucleosome and short footprints at bound CTCF motifs specifically (*e.g.* F13 O.R.  
45 = 1.309;  $q < 1.79 \times 10^{-308}$ ; Fisher’s Exact Test), accompanied by significant depletion for less accessible  
46 nucleosome types. Depletion of CTCF protein (**Figure 3H**) resulted in significant loss of short footprints (degron  
47 F13 O.R. = 0.681;  $q = 6.93 \times 10^{-103}$ ; reproducibility of effect size shown in **Extended Data Figure 4K**). These  
48 analyses suggest that CTCF protein dictates the composition of structurally distinct nucleosomes (specifically, to  
49 favor partially unwrapped states), and that CTCF can generate its own short footprint. Further, acute depletion  
50 demonstrates that CTCF directly regulates the abundance of short footprints only, as its loss results in a  
51 concomitant gain of all other nucleosome types.

1  
2 We repeated a similar set of analyses for motif-centric clusters (**Extended Data Figure 4G**). These clusters fell  
3 into groups that could be broadly categorized as: i.) nucleosomes of varying breathability with a dyad centered  
4 over the CTCF motif (group I); ii.) translationally-positioned nucleosomes with the CTCF motif precisely placed  
5 at the nucleosome entry / exit (group II); iii.) nucleosomes with the CTCF motif occurring within the nucleosome  
6 core particle (groups III, IV); and iv.) short footprints of varying sizes, precisely positioned over the CTCF motif  
7 (group V; possibly reflective of multi-modal binding by the constituent zinc fingers of CTCF<sup>45,48</sup>). Enrichment  
8 analysis of these clusters at bound versus control motifs (**Figure 3I**; reproducibility of enrichment shown in  
9 **Extended Data Figure 4L**), revealed that group V footprints were significantly enriched at true sites versus  
10 control (*e.g.* M15 O.R. = 30.939;  $q < 1.79 \times 10^{-308}$ ), and degradation experiments (**Figure 3J**; reproducibility of  
11 effect size shown in **Extended Data Figure 4M**) confirmed that these short footprints were generated by CTCF  
12 itself (degron M15 O.R. = 0.186;  $q < 1.79 \times 10^{-308}$ ). Together, these analyses establish a technical framework for  
13 studying nucleosome heterogeneity at TF binding sites, suggest that there are no specific preferred nucleosome  
14 orientations surrounding occupied CTCF motifs, and demonstrate that CTCF binding is the predominant driver  
15 of motif-dependent footprint positioning in E14 mESCs.

### 17 **SOX2 protein directly regulates nucleosome distortion at cognate binding sites**

18 While strong CTCF motifs are sufficient to outcompete well-positioned nucleosomes in synthetic contexts<sup>49</sup>, how  
19 motifs for the pluripotency factor SOX2 contend with nucleosomes remains unclear. Genomic, structural, and  
20 other biochemical assays have offered conflicting results on how SOX2 engages chromatin, particularly regarding  
21 SOX2 interactions within the nucleosome<sup>50,51</sup>, versus at the nucleosome entry / exit site<sup>49,52</sup>. To provide molecular  
22 insight into these possible binding modes for SOX2, we repeated the analyses above at SOX2-OCT4 composite  
23 motifs<sup>53</sup>, including data from a SOX2-dTAG mESC line<sup>54</sup> (degradation characterization and reproducibility in  
24 **Extended Data Figure 5A-F**). Footprint-centric analyses revealed 14 nucleosome types (F1-F14), which fell into  
25 similar groups as above (**Figure 4A**; **Extended Data Figure 5G**; reproducibility of cluster abundances shown in  
26 **Extended Data Figure 5I**). Differential analysis of nucleosome types at bound versus control motifs (**Figure 4C**;  
27 reproducibility of enrichment shown in **Extended Data Figure 5K**) revealed a significant enrichment for  
28 breathing nucleosomes, subnucleosomes, and short footprints at bound motifs (*e.g.* F13 O.R. = 1.420;  $q = 1.42 \times$   
29  $10^{-247}$ ), with depletion of inaccessible nucleosomal species. We speculate that these patterns may reflect the  
30 activities of ATP-dependent chromatin remodeling complexes and / or SOX2 altering nucleosome structure at  
31 these *cis*-regulatory elements. Like CTCF depletion, acute degradation of SOX2 protein (**Figure 4D**;  
32 reproducibility of effect size shown in **Extended Data Figure 5L**) resulted in significant depletion of short  
33 footprints (degron F14 O.R. = 0.871,  $q = 1.79 \times 10^{-3}$ ). Paradoxically, acute depletion of SOX2 is also accompanied  
34 by a significant decrease in the proportion of inaccessible nucleosome species (*e.g.* degron F9 O.R. = 0.713,  $q =$   
35  $1.95 \times 10^{-5}$ ) in the vicinity of cognate binding motifs. This result reveals an unexpected role for SOX2 protein in  
36 binding free DNA to generate short footprints and stabilizing nucleosomal species in proximal regions, possibly  
37 by reducing the extent of remodeling at loci with SOX2-OCT4 motifs.

38  
39 To further investigate this potential multimodality, we next performed a motif-centric analysis at these sites.  
40 Motif-centric clustering yielded 15 different clusters (M1-M15) falling in seven broad groups (**Figure 4B**) with  
41 patterns similar to those observed at CTCF sites (**Extended Data Figure 5H**; reproducibility of cluster  
42 abundances shown in **Extended Data Figure 5J**). Enrichment analyses at bound motifs compared to unbound  
43 controls revealed a striking pattern of translationally positioned nucleosomes, subnucleosomes, and short  
44 footprints with respect to bound motifs (**Figure 4E**; reproducibility of enrichment shown in **Extended Data**  
45 **Figure 5M**), and further analysis in the context of protein degradation (**Figure 4F**) revealed no significant effect  
46 of the protein itself on these distributions. Importantly, these findings are corroborated by orthogonal studies:  
47 enrichment of what appears to be a breathable nucleosome precisely centered over the motif (C2 O.R. = 1.238,  $q$   
48  $= 3.32 \times 10^{-13}$ ) aligns with recently published genomic mapping studies centered on OCT4, demonstrating that  
49 OCT4 can bind an altered nucleosome core particle directly<sup>55</sup>. Moreover, the most abundant enriched states (C3-  
50 C5) harbor a motif precisely placed at the nucleosome entry / exit site, consistent with work demonstrating this  
51 preference<sup>52</sup>. Finally, the lack of significant effect of SOX2 degradation on these trends implies that additional

1 regulatory factors must be required to reposition nucleosomes and enable SOX2 binding, consistent with recent  
2 genomic studies perturbing SWI/SNF remodelers<sup>56,57</sup>. Taken together, these analyses suggest multiple modes of  
3 interaction between SOX2 and nucleosomes at bound motifs; motif-encoded interactions with breathable (or  
4 noncanonical) nucleosome core particles and nucleosome edges occur concurrently with motif-independent  
5 interactions with stable nucleosome core particles.

### 6 7 **Widespread and regulated nucleosome distortion *in vivo***

8 We next sought to determine if the landscape of regulated nucleosome types exists beyond cultured mESCs, in a  
9 tissue *in vivo*. We generated reproducible single-molecule accessibility data in purified mouse hepatocytes for  $n$   
10 = 7 mice, including a published genetic model<sup>58</sup> heterozygous for a *Foxa2* allele lacking a critical nucleosome-  
11 interacting domain (*i.e.* ‘FOXA2-ΔHx’; characterization and reproducibility of heterozygous deletion in  
12 **Extended Data Figure 6A-E**). We then performed footprint-centric IDLI analyses (motif-centric tests all  
13 insignificant and not shown) on footprints drawn from regions surrounding ChIP-backed FOXA2 binding sites<sup>59</sup>  
14 and matched negative controls, to assess whether nucleosome types exist in primary cells.

15  
16 Footprint-centric analyses yielded 15 nucleosome types (F1-F15) / seven groups, corresponding qualitatively to  
17 structurally distinct (sub)nucleosomal protections seen in all mESC analyses (**Figure 5A**; **Extended Data Figure**  
18 **6F**). This suggests that widespread nucleosome distortion is a general property of mammalian chromatin, even  
19 within a terminally differentiated primary cell type. Computing the relative enrichment of these clusters at ChIP-  
20 backed FOXA2 binding sites compared to random motif matches (**Figure 5B**; reproducibility of enrichment  
21 shown in **Extended Data Figure 6G**), we observed a similar TF-dependent pattern to that seen for SOX2 in  
22 mESCs, with breathable (sub)nucleosomes and short footprints significantly enriched (*e.g.* F12 O.R. = 2.284,  $q$   
23 =  $2.16 \times 10^{-68}$ ), and inaccessible nucleosomes significantly depleted at *bona fide* binding sites. Differential  
24 nucleosome type analysis in FOXA2-ΔHx heterozygous hepatocytes versus wildtype littermate controls revealed  
25 significant depletion of two nucleosome types: F7 (predicted accessibility at SHL-3.5; mutant O.R. = 0.624,  $q$   
26 =  $6.45 \times 10^{-11}$ ) and F11 (one of 5 chromosome-like types; mutant O.R. = 0.820,  $q$  =  $8.02 \times 10^{-4}$ ), providing further  
27 evidence that the FOXA2 Hx motif controls nucleosome structure (reproducibility of effect size shown in  
28 **Extended Data Figure 6H**). Taken together, these analyses demonstrate the *in vivo* relevance of nucleosome  
29 distortion at cell-type-specific regulatory elements, and illustrate multimodal regulation of nucleosome structure  
30 by a histone-binding domain of FOXA2.

### 31 32 **DISCUSSION**

33 We report the discovery of pervasive, patterned, and programmed nucleosome distortion across mammalian  
34 epigenomes (**Figure 6**). Using a newly developed analytical pipeline we term “IDLI” we reveal that the majority  
35 of nucleosomes across diverse epigenomic states exist in a partially accessible state (**Figure 6A**), ranging from  
36 focal accessibility at specific sites along the nucleosomal wrap, to symmetrically unwound / breathable  
37 nucleosomes and subnucleosomes. Our pipeline also classifies short DNA protections – given the raw abundance  
38 of these protections, these are most likely owed to the histone H2A/H2B dimer, though our analyses at TF binding  
39 sites show a quantifiable (and expected) contribution of TFs to the overall number of short protections.

40  
41 IDLI analyses capture regulated patterns of nucleosome distortion genome-wide (**Figure 6B**). Our proof-of-  
42 concept study demonstrates that nucleosome types—which we interpret as structurally distinct nucleosomes—are  
43 specifically enriched and depleted across epigenomic domains; further, our first-in-class single-fiber co-  
44 occupancy analyses demonstrate the existence of statistically-significant co-occurrence of structurally distinct  
45 nucleosomes on individual chromatin fibers. These analyses are uniquely enabled by the IDLI pipeline, which  
46 operates on long-read high-resolution footprinting data. Future work must assess the significance of these co-  
47 occurrence patterns at even higher resolution (*e.g.* at actively transcribed promoters, enhancers, or gene bodies,  
48 or at individual TF binding sites, which we have preliminarily computed for CTCF and SOX2 in mESCs to  
49 demonstrate feasibility; **Extended Data Figure 7**). Given the short residence times of most TFs<sup>60</sup>, however, we  
50 stress that a comprehensive TF-specific co-accessibility analysis will require SAMOSA data with increased  
51 genomic coverage specifically at target TF binding sites. In the future, this may be achieved via additional

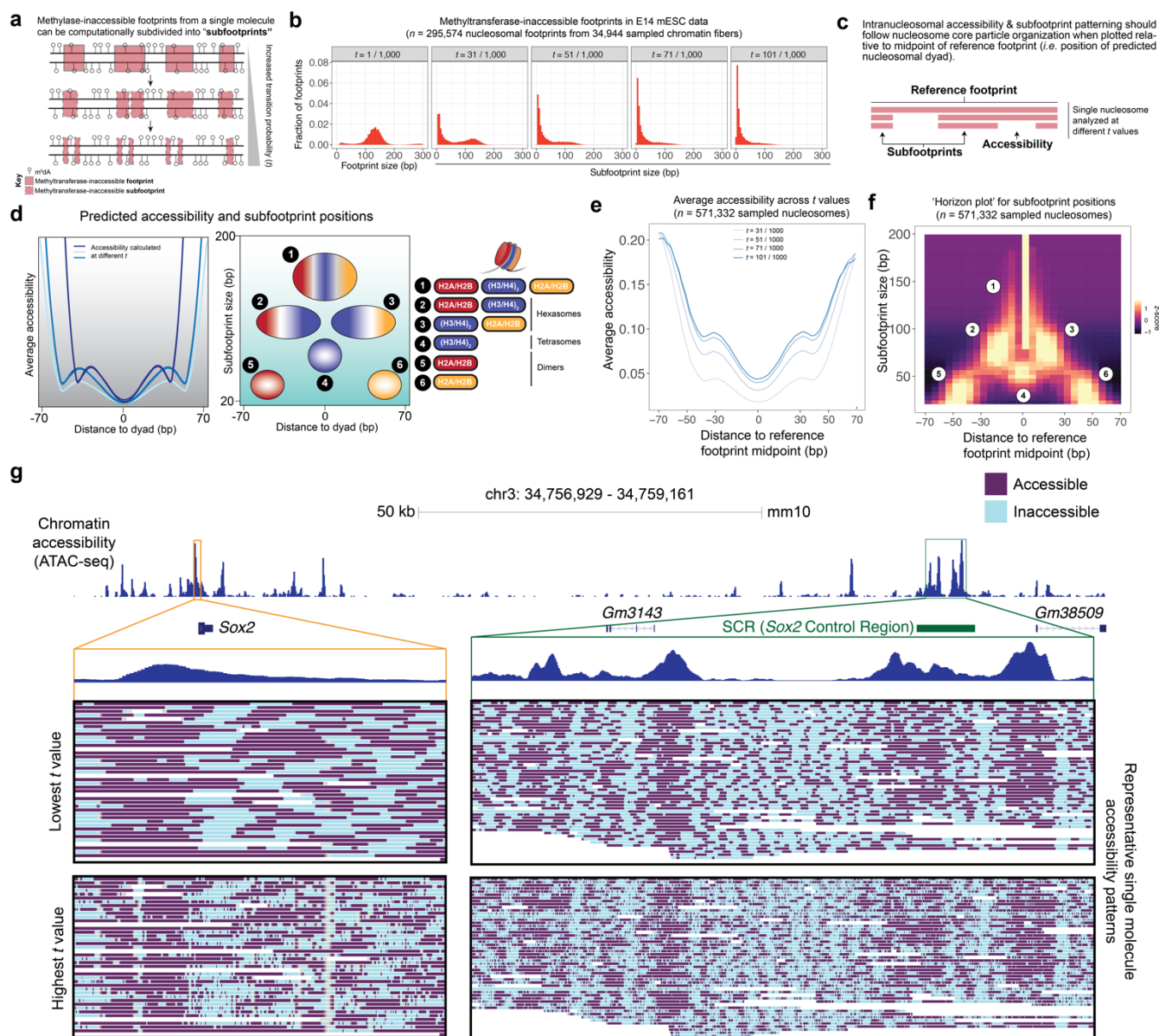
1 technology improvements (*e.g.* higher-throughput multiplex target capture of TF binding sites<sup>61</sup>, analogous to  
2 multiplex bisulfite sequencing<sup>40,62</sup>).

3  
4 Finally, we show that the TFs CTCF, SOX2, and FOXA2 directly program nucleosome distortion at cognate  
5 binding sites (**Figure 6C**). Leveraging small molecule sensitive degrons (CTCF; SOX2) and a mutant *Foxa2*  
6 allele defective for histone binding, we provide compelling genetic evidence that TF dosage directly influences  
7 the structure and positioning of nucleosomes at cognate motifs in cell culture and *in vivo*. Our results for SOX2  
8 and FOXA2 imply multiple modes by which TFs interact with DNA motifs and / or nucleosomes. For both TFs,  
9 depletion (or mutation) paradoxically results in a loss of less-accessible nucleosome types, signaling a direct role  
10 for these factors in stabilizing nucleosomes. In the case of SOX2 specifically, this phenotype appears to be motif-  
11 independent; motif-centric analyses, conversely, demonstrate a preferred positioning state of nucleosomes with  
12 the motif oriented at the nucleosome entry / exit site. Our results thus harmonize seemingly conflicting results in  
13 the field regarding SOX2 binding. First, binding and nucleosome positioning by SOX2 is clearly motif-encoded,  
14 and SOX2 is not sufficient to dictate nucleosome positioning at binding sites; these results are consistent with  
15 studies demonstrating the importance of ATP-dependent chromatin remodeling to SOX2 motif accessibility<sup>56,57,63</sup>,  
16 as well as studies demonstrating the inability of SOX2 to overcome nucleosomal occlusion<sup>49,52</sup>. Second, SOX2  
17 itself can stabilize nucleosomes in the vicinity of SOX2-OCT4 motifs; we speculate that this ‘on-nucleosome’  
18 regulatory mode may relate to observed interactions between SOX2 and the nucleosome dyad<sup>50</sup>, and further note  
19 structural evidence that SOX2 and other HMG-box proteins can interact with the nucleosome directly (*e.g.* at  
20 SHL2<sup>38,51</sup>, though we acknowledge no specificity for SHL-2 in our data). Finally, these properties appear to be  
21 shared at bound FOXA2 motifs in primary mouse hepatocytes. Compared to unbound control motifs, FOXA2-  
22 regulated sites are specifically enriched for accessible (sub)nucleosomes and short protections, supporting  
23 previous models for a motif-encoded grammar for these interactions<sup>64</sup>. Altering the dosage of intact FOXA2  
24 protein by replacing one allele with FOXA2-ΔHx has a significant impact, however, on the observed distribution  
25 of nucleosome types at FOXA2 binding sites, resulting in loss of nucleosomes with accessibility at SHL-3.5, and  
26 chromatosomes; these findings are consistent with the canonical definitions of TF pioneering<sup>65</sup>, in which the  
27 nucleosome interaction itself instructs cell fate<sup>58</sup>.

28  
29 Looking ahead, we anticipate the broad applicability of the IDLI pipeline to a wide variety of problems in  
30 chromatin biology. ATP-dependent chromatin remodeling<sup>29,30</sup>, nucleosome assembly, transcription<sup>66</sup>,  
31 replication<sup>31</sup>, and all other nuclear processes regularly generate structurally distorted nucleosomes, but prevailing  
32 approaches to study this distortion rely on *in vitro* biochemical and biophysical assays. Moving forward, we  
33 envision a routine experimental framework that complements highly resolved biochemistry with near-nucleotide-  
34 precise snapshots of structurally distorted states *in vivo*. Combining bottom-up biochemistry, top-down structural  
35 classification presented here, and complementary advances in *in situ* chromatin imaging<sup>67</sup> will ultimately enable  
36 deeper quantitative insight into the *in vivo* biochemistry of essential nuclear regulatory processes.

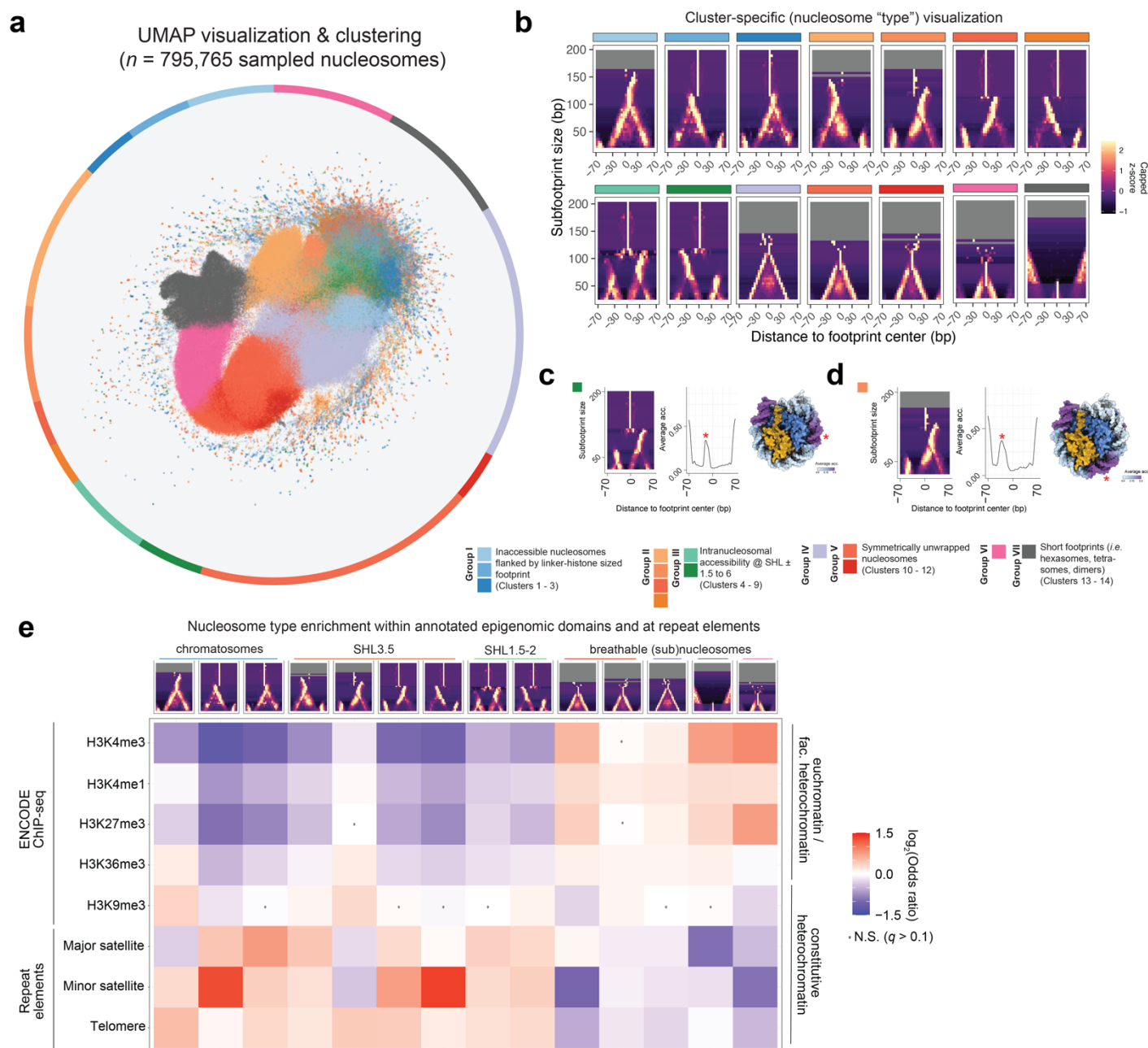


## MAIN TEXT FIGURES



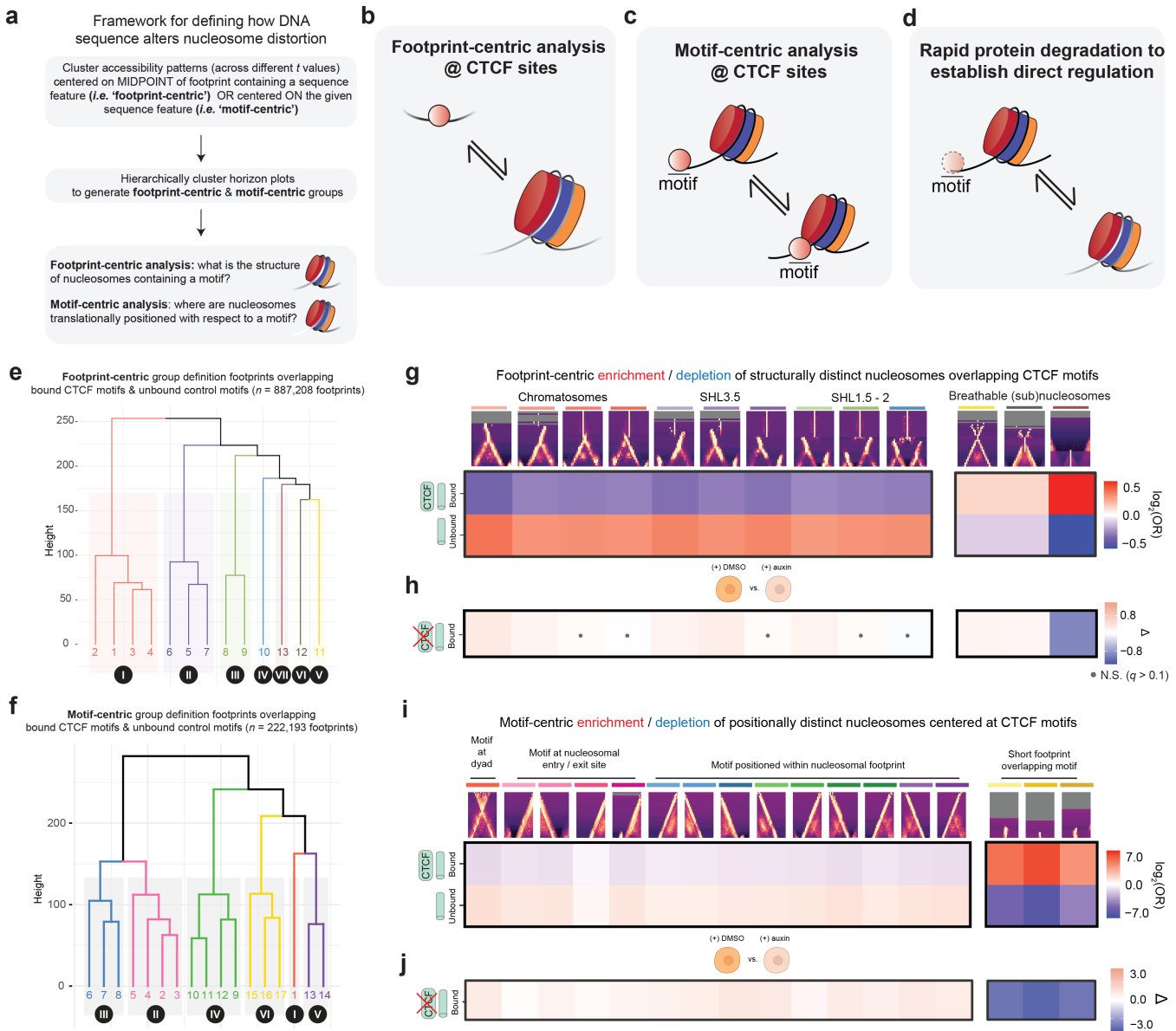
**Figure 1: Computational framework for defining subunit architecture and organization of nucleosome core particles from *in vivo* long-read methyltransferase footprinting data.** **A.)** Sequencing reads from single-molecule adenine methylated oligonucleosome sequencing assay [SAMOSA] data are analyzed with a two-state hidden Markov model (HMM) to define contiguous stretches of methylase-inaccessible nucleotides ('footprints') from EcoGII-treated chromatin. As shown in this schematic, individual footprints are subdivided into shorter 'subfootprints' at higher transition probabilities ( $t$ ; the likelihood of switching between the 'inaccessible' and 'accessible' model states). **B.)** Distributions of footprint and subfootprint sizes from randomly sampled footprints from E14 mESCs across a range of  $t$  values. Shown on the left is the footprint size distribution from the transition probability ( $t = 1 / 1,000$ ) used in our prior work, calibrated to principally detect mononucleosome-sized footprints. Increasing  $t$  values results in a greater proportion of progressively shorter footprints from an identical set of chromatin fibers. **C.)** We named this computational framework wherein intra-nucleosomal accessibility within single footprints are systematically analyzed across a range of  $t$  values: Iteratively Defined Lengths of Inaccessibility (or IDLI, for short). We hypothesized that by tuning the transition probability, we should be able to recapitulate intra-nucleosome accessibility and subfootprint patterns reflective of the organization of the nucleosome core particle. **D.)** Predicted outcomes for nucleosomal footprints in E14 mESC data analyzed with

1 IDLI. **Left:** Mean accessibility should be patterned symmetrically with respect to the footprint midpoints (*i.e.*  
2 predicted nucleosomal dyad positions) at different  $t$  values. **Right:** Predicted sizes and positions of subfootprints  
3 (plotted relative to footprint midpoints) should reflect the subunit architecture of the octameric nucleosome,  
4 including: (1) histone octamers, (2-3) hexasomes (lacking an H2A-H2B dimer), (4) H3-H4 tetramers, and (5-6)  
5 H2A-H2B dimers. **E.)** Measured mean accessibility with respect to reference (*i.e.* from  $t = 1 / 1,000$ ) footprint  
6 midpoints reveals expected symmetric pattern about predicted dyad positions. **F.)** ‘Horizon plot’ visualization  
7 (with respect to footprint midpoints) captures enrichment of previously-noted subfootprint sizes and positions,  
8 including octameric nucleosomes that exhibit spontaneous ‘unspooling’ of DNA from entry / exit sites (Feature  
9 1; vertical line centered at footprint midpoint). **G.)** Single-molecule accessibility patterns (EcoGII-accessible and  
10 -inaccessible DNA in purple and teal, respectively) at *Sox2* promoter (left; depicted in orange) and *Sox2* Control  
11 Region [SCR] (right; depicted in green). Data from representative identical single EcoGII-footprinted fibers are  
12 shown for the lowest ( $t = 1 / 1,000$ ; top) and highest ( $t = 101 / 1,000$ ; bottom) transition probabilities used in IDLI.  
13 Bulk-average chromatin accessibility data (from ATAC-seq) in mESCs are also shown across the entire locus  
14 (top) and at specific regulatory loci (bottom).  
15



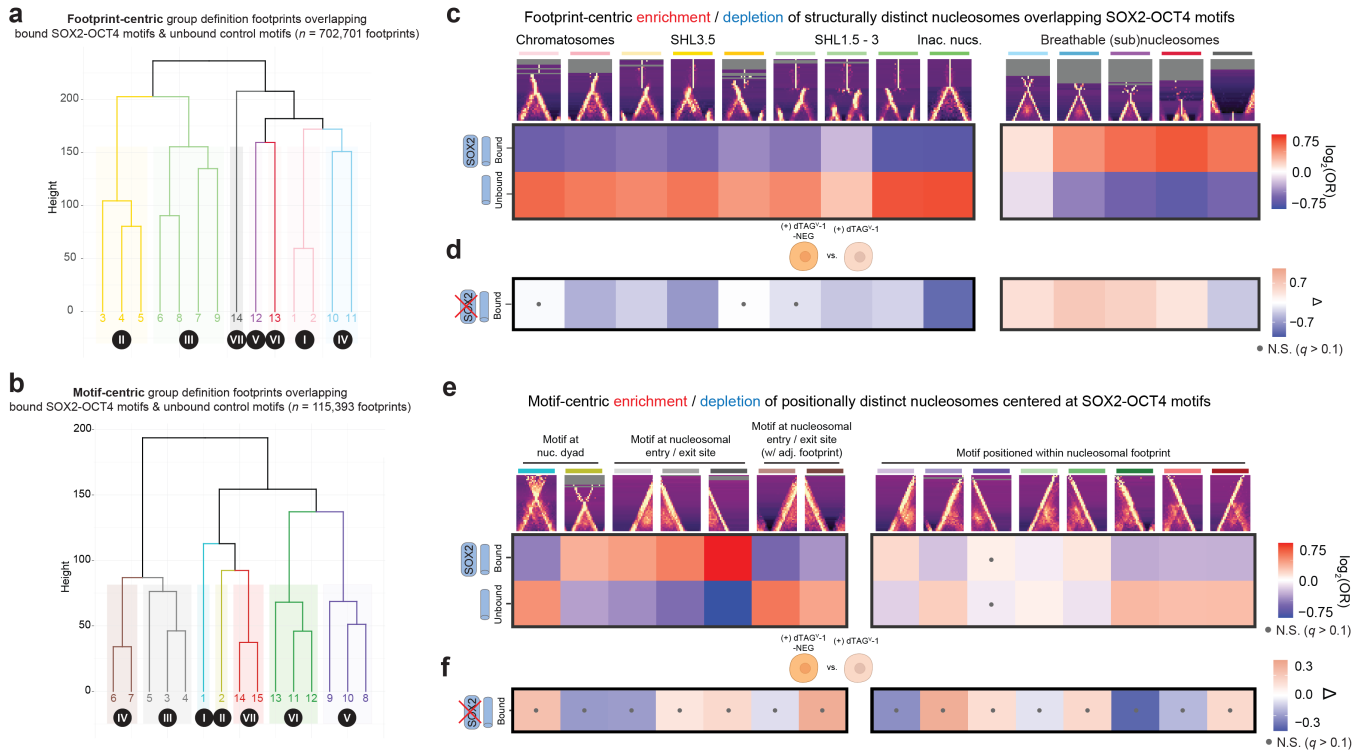
**Figure 2: Structurally distinct nucleosomes are enriched in an epigenomic domain- and repeat-sequence-specific manner.** **A.)** UMAP visualization of  $n = 795,765$  footprints within  $\pm 500$  bp of midpoints of histone post-translational modification-defined epigenomic domains (in mESCs) or 'unmappable' repeat elements. Individual footprints within UMAP are colored based on Leiden-clustering assignments of their intra-nucleosomal accessibility (*i.e.* distortion patterns). Proportions of each cluster are depicted as a pie chart surrounding the UMAP plot. Key for individual nucleosome types and respective 'groups' (as defined by hierarchical clustering) are shown on bottom. **B.)** 'Horizon plots' for 14 distinct structural classes of footprints, including: chromosomes (Group I), octameric nucleosomes with focal intra-nucleosomal accessibility (Group II-III), symmetrically breathing nucleosomes (Group IV-V), and shorter footprints consistent with hexasomes, tetramers, and dimers (Group VI-VII). **C.)** 'Horizon plot' (left), mean accessibility (middle), and structural visualization (right) of C9 footprints, which are nucleosome core particles accessible at superhelical location [SHL]  $\pm 2$  (denoted by red asterisk). **D.)** As in (C), but for C5 footprints, which are nucleosome core particles accessible at SHL  $\pm 3.5$ . **E.)** Heatmap of  $\log_2$ -transformed odds ratios (ORs) for different nucleosome types across epigenomic domains (H3K4me3-marked promoters, H3K4me1-marked enhancers, H3K27me3-marked Polycomb-repressed

1 chromatin, H3K36me3-marked actively transcribed gene bodies, and H3K9me3-marked constitutive  
2 heterochromatin) and repeat elements (mouse major satellite, minor satellite, and telomeric sequences). ORs  
3 designated with a grey dot indicate those that are not statistically significant from Fisher's exact tests (*i.e.* Storey's  
4  $q$ -value  $> 0.1$ ).  
5

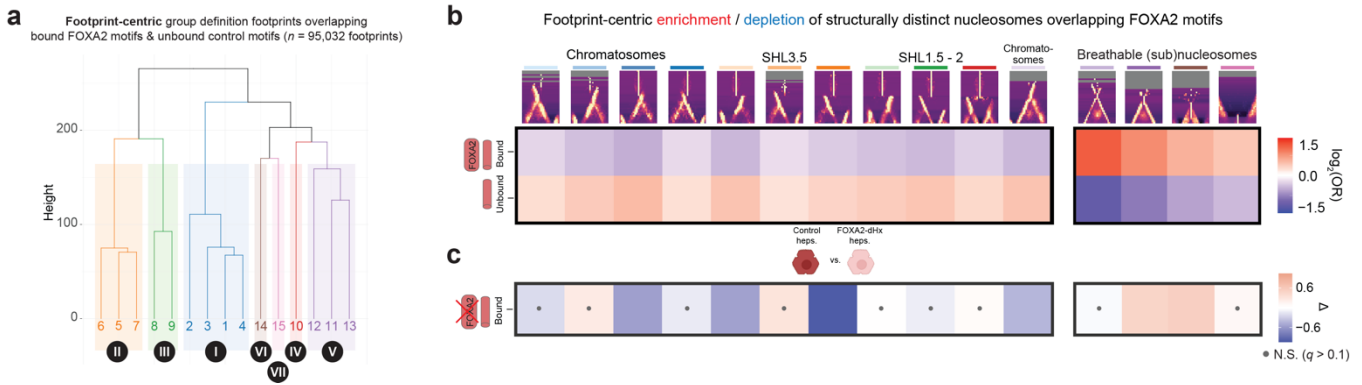


**Figure 3: Footprint- and motif-centric analyses of nucleosomes at CTCF motifs in wild-type and CTCF-depleted mESCs.** **A.)** Framework for defining nucleosome types and translational positions for footprints at TF-binding motifs of interest. Footprint-centric analyses (*i.e.* clustering accessibility centered on footprint midpoints) quantify distortion patterns for nucleosomes within a 1-kb window centered on a given motif of interest. Motif-centric analyses (*i.e.* clustering accessibility centered on motif sequence) address how nucleosomes are translationally positioned with respect to a given motif. **B-C.)** Footprint-centric and motif-centric analyses were performed in E14 mESC data by comparison of bound vs. randomly sampled motifs for CCTCC-binding factor (CTCF). **D.)** Similar quantification was performed at bound CTCF motifs in mESCs expressing an AID-tagged form of CTCF by comparison of distortion patterns and translational positions before vs. after auxin-mediated CTCF degradation (for 6 hr). **E-F.)** Hierarchical clustering of 'horizon plot' data from footprint-centric and motif-centric analyses at bound and unbound CTCF motifs reveals 7 distinct 'groups' of nucleosome types and 6 distinct 'groups' of translational positions. **G.)** Heatmap of  $\log_2$ -transformed odds ratios (ORs) for assessing enrichment / depletion of different nucleosome types at bound vs. randomly sampled CTCF motifs in E14 mESCs. ORs designated with a grey dot indicate those from Fisher's exact test that are not statistically significant (*i.e.* Storey  $q$ -value  $> 0.1$ ). **H.)** Heatmap of effect sizes ( $\Delta$ ) for nucleosome types at bound CTCF motifs in auxin vs. DMSO-treated CTCF-AID mESCs. Effect sizes designated with a grey dot indicate those from Fisher's exact test that are

1 not statistically significant (*i.e.* Storey  $q$ -value  $> 0.1$ ). **I-J.)** As in **(G-H)**, but for CTCF motif-associated  
2 translational positions.  
3

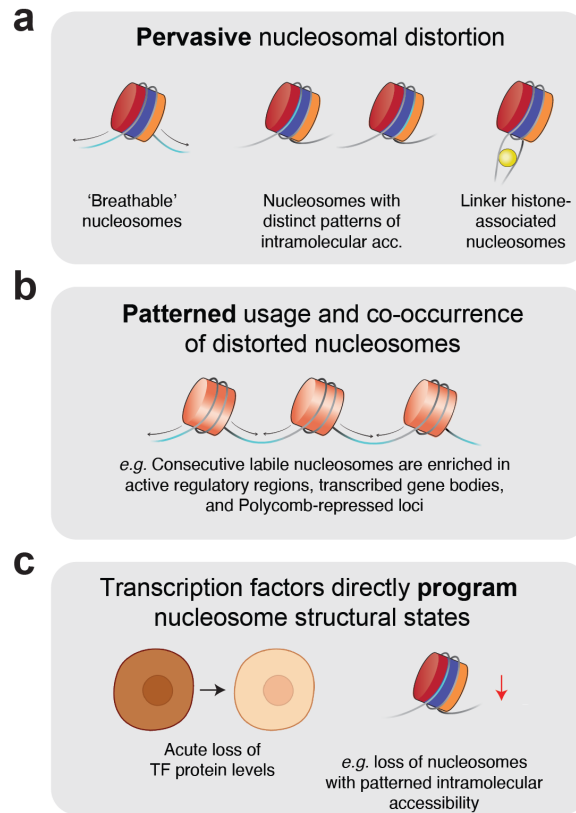


**Figure 4: Footprint- and motif-centric analyses of nucleosomes at SOX2-OCT4 motifs in wild-type and SOX2-depleted mESCs. A-B.)** Hierarchical clustering of ‘horizon plot’ data from footprint-centric and motif-centric analyses at bound and unbound SOX2-OCT4 composite motifs reveals 7 distinct ‘groups’ of nucleosome types and 7 distinct ‘groups’ of translational positions. **C.)** Heatmap of  $\log_2$ -transformed odds ratios (ORs) for assessing enrichment / depletion of different nucleosome types at bound vs. randomly sampled SOX2-OCT4 composite motifs in E14 mESCs. ORs designated with a grey dot indicate those from Fisher’s exact test that are not statistically significant (*i.e.* Storey  $q$ -value  $> 0.1$ ). **D.)** Heatmap of effect sizes ( $\Delta$ ) for nucleosome types at bound SOX2-OCT4 composite motifs in dTAG<sup>V-1</sup> vs. dTAG<sup>V-1</sup>-NEG-treated (control) SOX2-dTAG mESCs. Effect sizes designated with a grey dot indicate those from Fisher’s exact test that are not statistically significant (*i.e.* Storey  $q$ -value  $> 0.1$ ). **E-F.)** As in (C-D), but for SOX2 motif-associated translational positions.



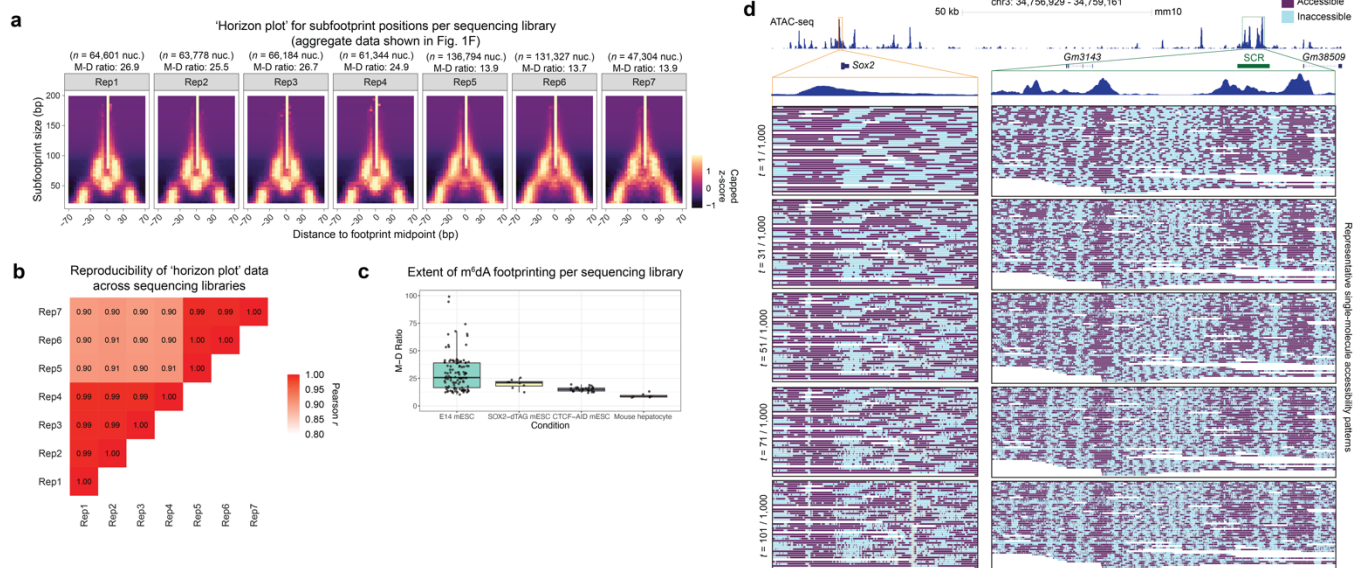
1  
2  
3 **Figure 5: Footprint-centric analyses of nucleosomes at FOXA2 motifs in wild-type primary hepatocytes**  
4 **and hepatocytes derived from heterozygous FOXA2-ΔHx mice. A.)** Hierarchical clustering of ‘horizon plot’  
5 data from footprint-centric analyses at bound and unbound FOXA2 motifs reveals 7 distinct ‘groups’ of  
6 nucleosome types. **B.)** Heatmap of  $\log_2$ -transformed odds ratios (ORs) for assessing enrichment / depletion of  
7 different nucleosome types at bound vs. randomly sampled FOXA2 motifs in C57BL/6J hepatocytes. ORs  
8 designated with a grey dot indicate those from Fisher’s exact test that are not statistically significant (*i.e.* Storey  
9  $q$ -value  $> 0.1$ ). **C.)** Heatmap of effect sizes ( $\Delta$ ) for nucleosome types at bound FOXA2 motifs in heterozygous  
10 FOXA2-ΔHx hepatocytes vs. wild-type hepatocytes (derived from age-matched littermate controls). Effect sizes  
11 designated with a grey dot indicate those from Fisher’s exact test that are not statistically significant (*i.e.* Storey  
12  $q$ -value  $> 0.1$ ).  
13



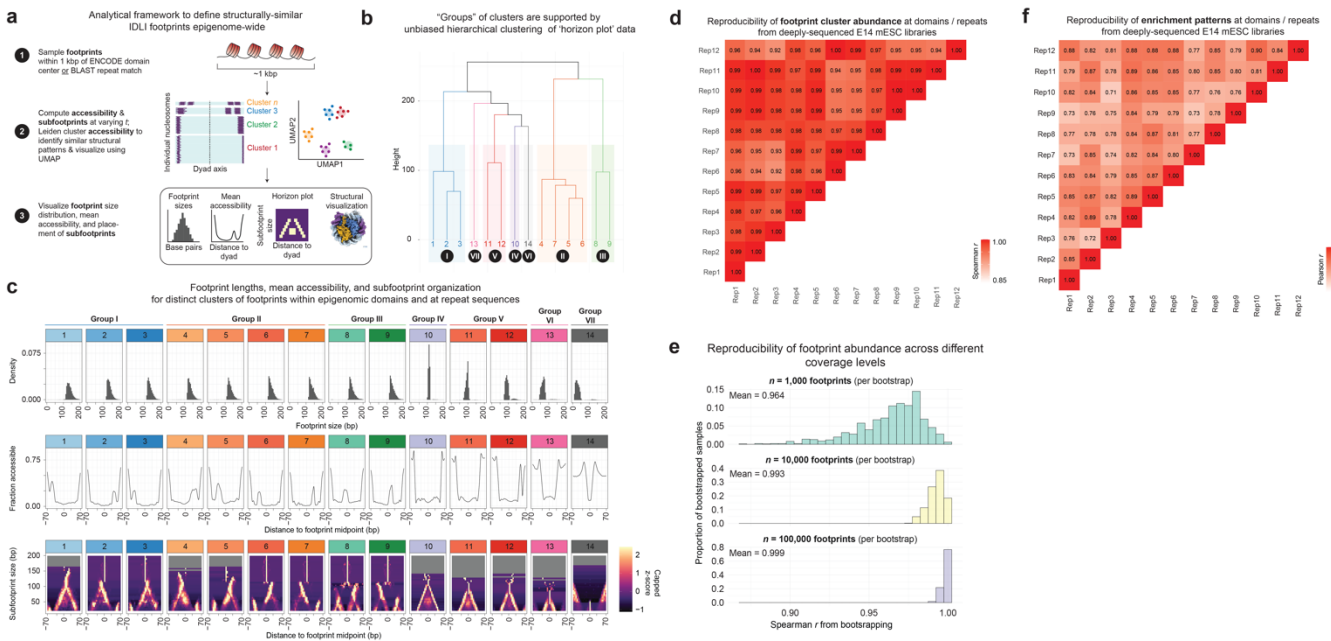


1  
2  
3 **Figure 6: IDLI reveals pervasive, patterned, and programmed nucleosome distortion within mammalian**  
4 **chromatin.** A.) Analyses of intranucleosomal accessibility from non-destructive footprinting data indicate that  
5 most nucleosomes surveyed are not entirely refractory to EcoGII methylation; instead, nucleosomes are often  
6 symmetrically 'unwound' or highly 'breathable', they can exhibit focal accessibility at specific superhelical  
7 locations, or they can exist as chromatosomes. B.) We find that patterns of nucleosomal distortion are enriched /  
8 depleted to different extents across epigenomic domains and repeat elements (not shown here). Moreover, we  
9 demonstrated that consecutive homotypic 'triplets' of nucleosomes are often statistically enriched on single fibers  
10 (*e.g.* 'breathing' nucleosomes tend to co-occur on individual chromatin fibers in active and facultatively repressed  
11 chromatin). C.) Across multiple cell types and genetic perturbations, we uncover that TFs are directly responsible  
12 for changes in the abundance of different nucleosome types at loci that harbor their cognate binding motifs. In the  
13 context of acute protein depletion (*i.e.* CTCF and SOX2-degron lines), these observations strongly suggest a direct  
14 role in TF-mediated programming of nucleosomal states.  
15

## EXTENDED DATA FIGURES

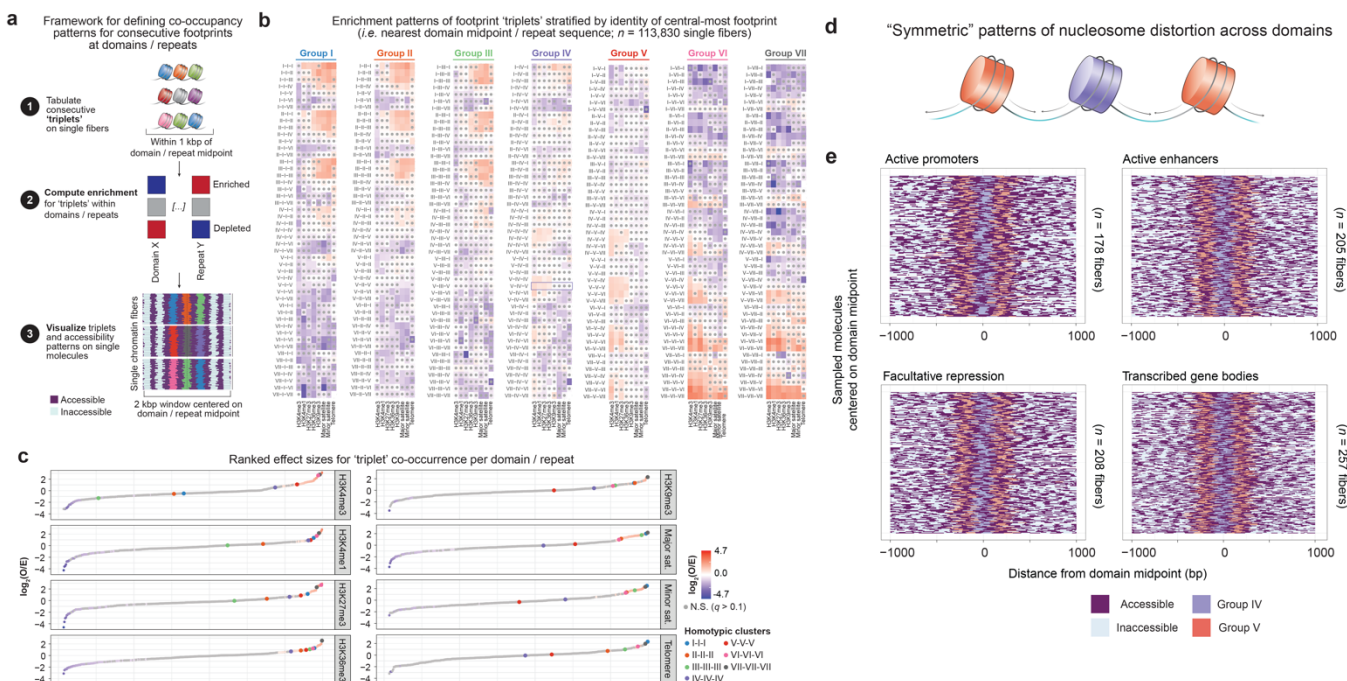


**Extended Data Figure 1: Reproducibility of subfootprint detection with IDLI framework, related to Figure 1. A.)** 'Horizon plots' for enrichment patterns of subfootprints of different sizes at varying positions with respect to footprint midpoints, plotted separately for nucleosomal footprints derived from individual biological replicates (for aggregate data shown in **Figure 1F**). Subfootprints are reliably detected across sequencing libraries with different extents of methyltransferase footprinting, as indicated by mononucleosome-dinucleosome ratio (MDR; value shown above each plot). **B.)** Pearson  $r$  values from comparison of 'horizon plot' data across sequencing libraries (from data shown in **Extended Data Figure 1A**). Samples with more similar MDRs exhibited quantitatively higher levels of correlation in reproducibility of 'horizon plot' data (e.g. Rep1-4 vs. Rep5-7). **C.)** Box-and-whisker plots of MDR values for SAMOSA data included in this study, wherein individual points represent data from single biological replicates. Data are stratified based on cell type and genotype. All libraries from mESCs included in this study are well-footprinted based on previously-defined criteria (i.e. MDR > 10). Primary hepatocyte samples (wild-type and FOXA2- $\Delta$ Hx) have slightly lower MDR values (MDR = 7.53 – 13.06), which we ascribe to slightly reduced efficiency of EcoGII footprinting in primary cells (compared to cell culture conditions) consequent to global differences in chromatin accessibility in pluripotent cells vs. terminally-differentiated cell types. **D.)** Single-molecule accessibility patterns (EcoGII-accessible and -inaccessible DNA in purple and teal, respectively) at *Sox2* promoter (left; depicted in orange) and *Sox2* Control Region [SCR] (right; depicted in green). Data from representative identical single EcoGII-footprinted fibers are shown for the full range of transition probabilities used in IDLI ( $t = 1, 31, 51, 71, \text{ and } 101 / 1000$ ). Bulk-average chromatin accessibility data (from ATAC-seq) in mESCs are also shown across the entire locus (top) and at specific regulatory loci (bottom).

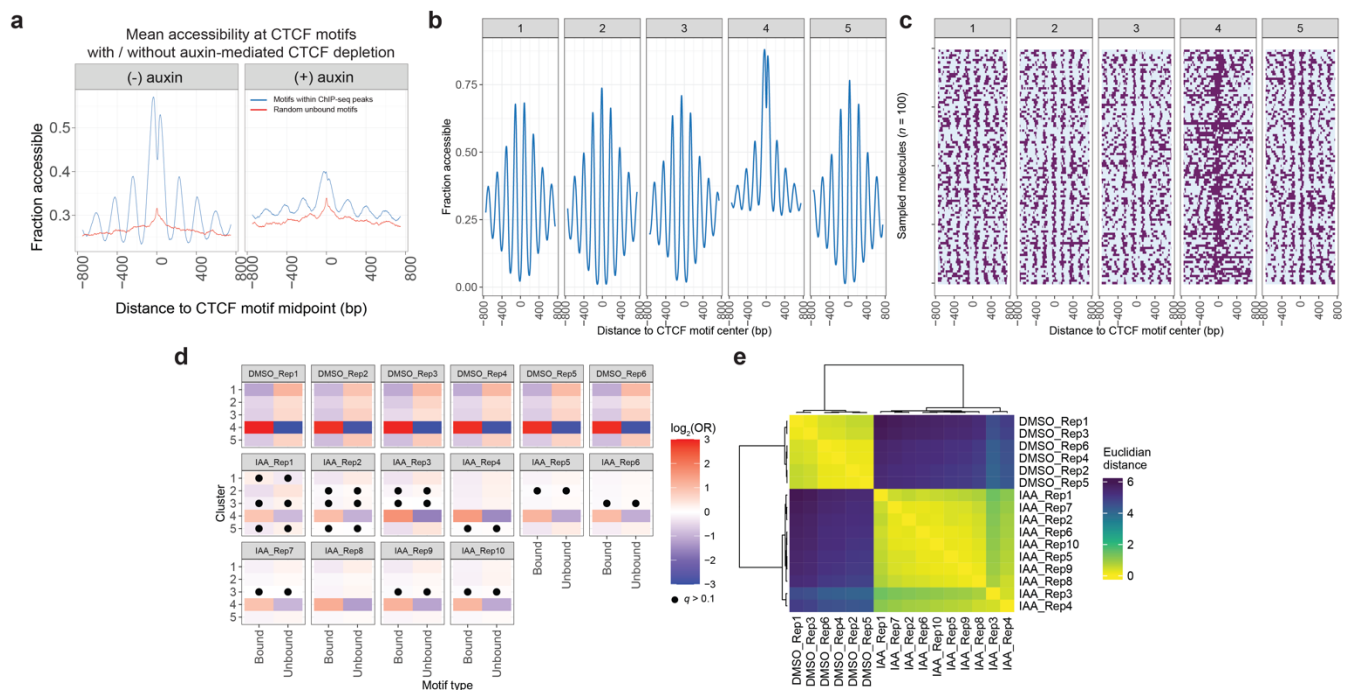


**Extended Data Figure 2: Clustering analysis for defining epigenomic domain- and repeat element-associated nucleosomal distortion patterns, related to Figure 2. A.)** Framework for clustering and visualization of footprints with shared structural features (*i.e.* patterns of intranucleosomal accessibility) within epigenomic domains and at repeat sequences. Footprints are sampled from within  $\pm 500$  bp window centered on either the midpoint of ENCODE-defined domains or from the top-scoring repeat for sequencing reads that contain one or more BLAST matches. Single-footprint accessibility patterns across  $t = 31, 51, \text{ and } 71 / 1,000$  are subject to Leiden clustering, and resultant data are shown as a UMAP visualization. To ascribe biological function to footprints with different Leiden assignments, we visualized: (1) footprint size distributions, (2) mean accessibility profiles with respect to footprint midpoints, (3) 'horizon plots' for detecting subfootprint placement, and (4) structural visualization of mean accessibility overlaid on the PDB structure for an interphase nucleosome (7KBD). **B.)** Hierarchical clustering of 'horizon plot' data from 14 Leiden-defined clusters of domain and repeat-associated footprints into 7 distinct groups with shared intra-nucleosomal accessibility patterns. **C.)** Analysis of individual Leiden-defined clusters of nucleosomal distortion patterns at epigenomic domains and repeat sequences. **Top:** Histograms of footprint sizes on a per-cluster basis. **Middle:** Mean accessibility traces centered on footprint midpoints on a per-cluster basis. **Bottom:** 'Horizon plots' for subfootprint sizes versus distance from footprint midpoints to assess subfootprint enrichment at particular locations on a per-cluster basis. z-scores are capped such that the maximum shown in the heatmap are the 95<sup>th</sup> percentile values within  $\pm 70$  bp of the respective footprint midpoints. **D.)** Spearman  $r$  values from comparison of proportions of nucleosome types at domains and repeats across biological replicates (from  $n = 12$  selected, deeply-sequenced mESC samples surveyed for both domain- and repeat-associated footprints). **E.)** Bootstrapping analysis to determine number of footprints required for reproducible cluster assignments. From  $n = 795,765$  domain- and repeat-associated footprints, we bootstrapped  $n = 1,000, 10,000, \text{ or } 100,000$  footprints per trial. Shown are the distribution of resultant Spearman  $r$  values (from a total of 10,000 trials per condition). These data suggest that our Leiden assignments for clusters of nucleosomal distortion patterns are reproducible and robust at the number of nucleosomes we have chosen for analysis. **F.)** Pearson  $r$  values from comparison of nucleosome type enrichment patterns across domains and repeats (aggregate data shown in **Figure 2E**) across biological replicates (from  $n = 12$  selected, deeply-sequenced mESC samples surveyed for both domain- and repeat-associated footprints).

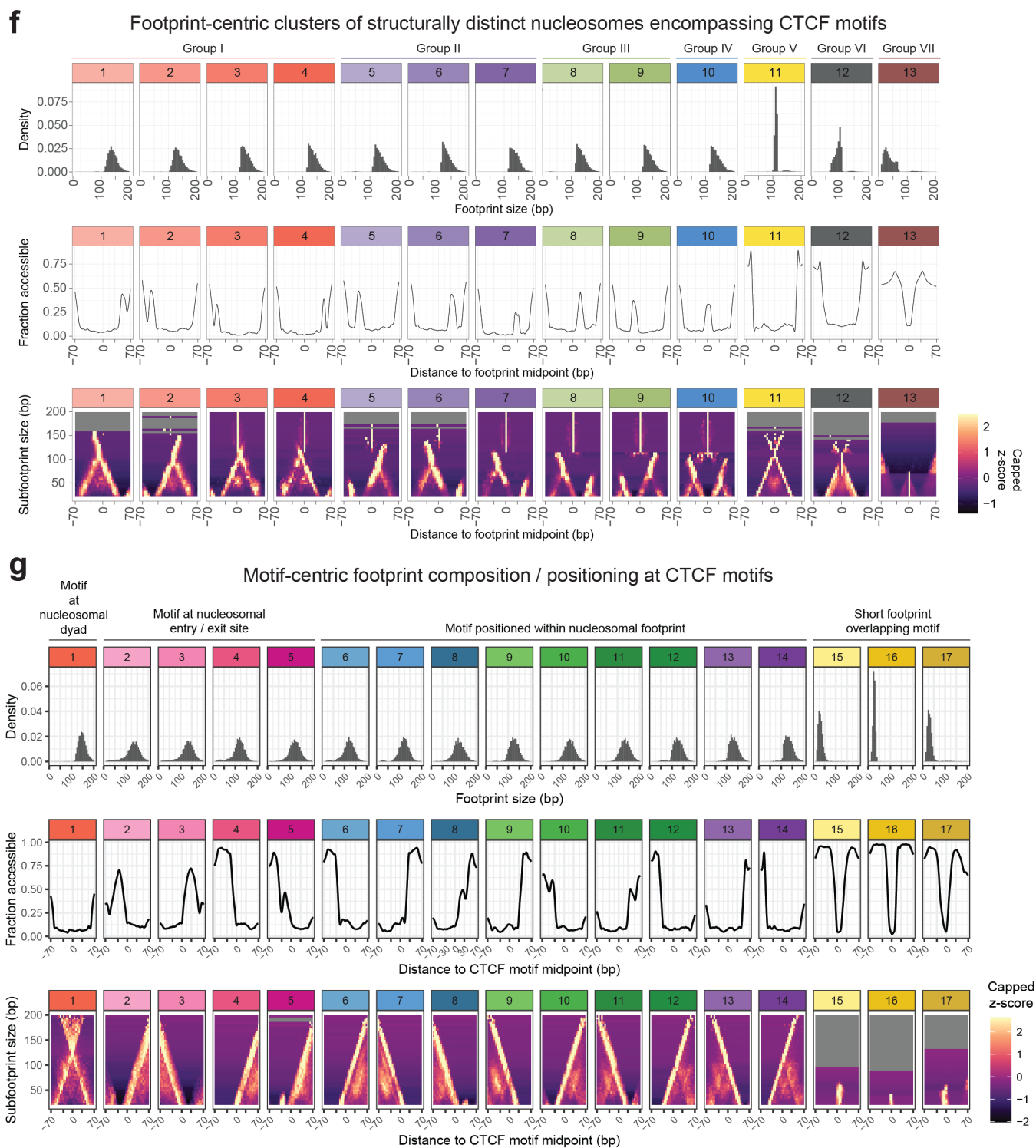
1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31



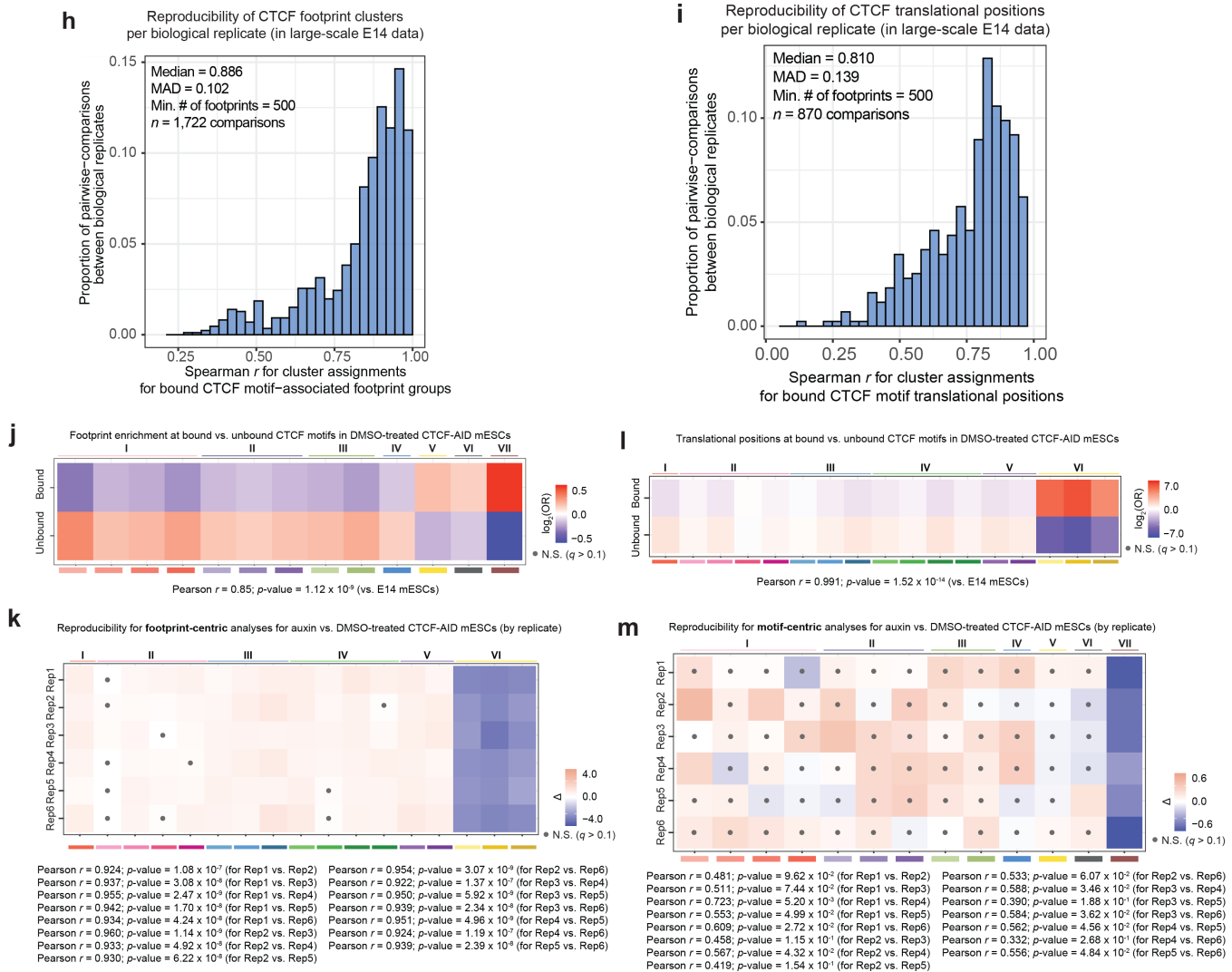
1  
2  
3 **Extended Data Figure 3: Co-occupancy analysis reveals enrichment of homotypic clusters of footprints**  
4 **with shared nucleosomal distortion patterns on individual fibers.** A.) Domain- and repeat-associated  
5 footprints were analyzed for co-occupancy patterns on single chromatin fibers. Instances of three consecutive  
6 footprints on individual sequencing reads nearest domain or repeat midpoints were tabulated based on their  
7 associated group assignments (as defined by hierarchical clustering in **Extended Data Figure 2B**). We required  
8 that at least three Leiden-classified footprints occurred within  $\pm 500$  bp of the domain or repeat midpoint of  
9 interest, excluding fibers with footprints  $> 200$  nt in length and/or with footprints from the least abundant (bottom  
10 10%) of Leiden-defined clusters. B.) Heatmap of  $\log_2$ -transformed O / E ratios computed for each possible 'triplet'  
11 on a per-domain or per-repeat basis from a total of  $n = 113,830$  single fibers. Data are stratified based on the group  
12 identity of the central-most footprint (i.e. footprint nearest the domain / repeat midpoint of interest). Effect sizes  
13 designated with a grey dot indicate those from Fisher's exact test that are not statistically significant (i.e. Storey  
14  $q$ -value  $> 0.1$ ). C.) Waterfall plots of ranked effect sizes stratified by locus indicates enrichment of homotypic  
15 clusters of footprints with shared nucleosomal distortion patterns. As in (B), grey dots used to indicate effect sizes  
16 for 'triplets' with  $q > 0.1$ . Red and blue dots are used to indicate 'triplets' that are enriched and depleted within a  
17 given locus, respectively. 'Triplets' that comprise homotypic clusters are highlighted by larger colored dots (see  
18 legend on bottom right). D.) Schematic for 'triplets' of symmetric pattern of 'breathable' nucleosomes (IV-V-IV),  
19 which are statistically enriched in active and facultatively repressed chromatin. E.) Single-molecule visualization  
20 of chromatin fibers with IV-V-IV 'triplets', wherein each row represents a unique sequenced fiber (total number  
21 of molecules indicated on right side of each plot). Accessibility patterns across a 2-kb window centered on domain  
22 midpoints of interest are also depicted (purple for EcoGII-accessible and teal for EcoGII-inaccessible DNA,  
23 respectively). Footprints from different groups are colored as described in associated legend; relative widths  
24 depict the number of nucleotides protected per footprint. Only data from epigenomic domains that are  
25 significantly enriched for the given triplet (IV-V-IV) are shown.



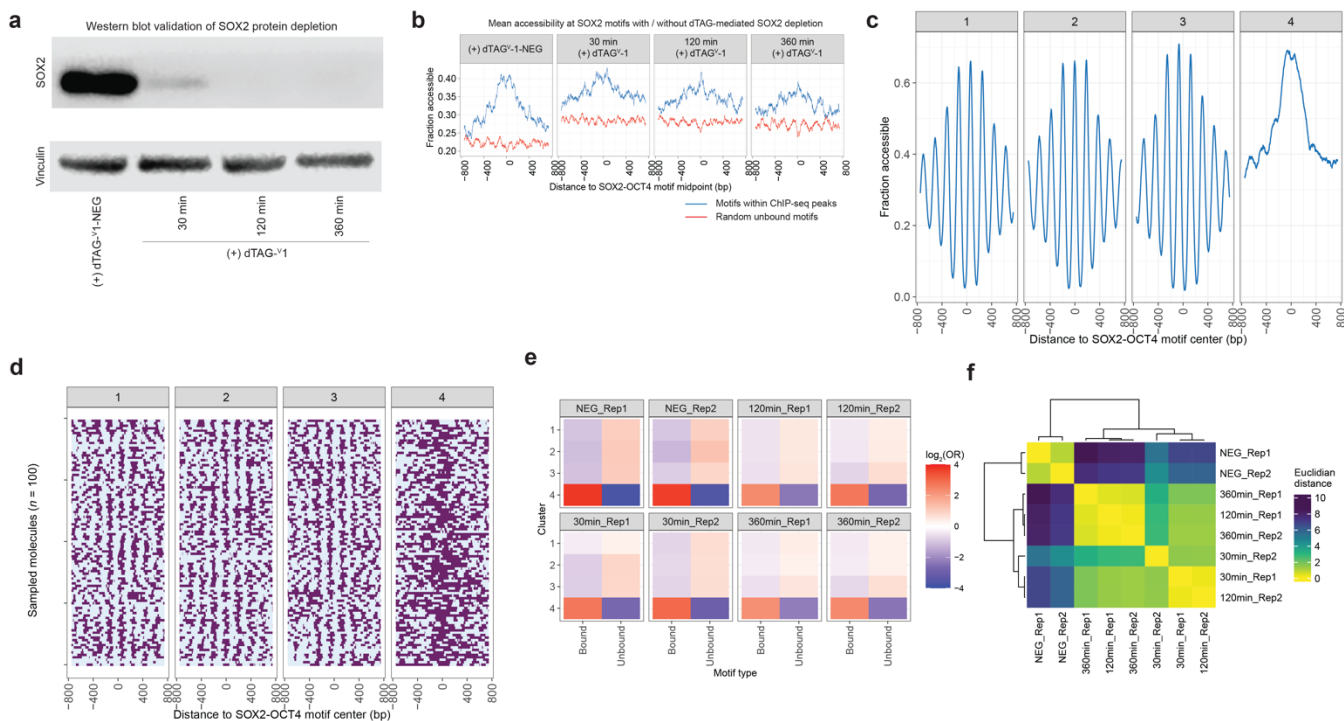
1  
2 **Extended Data Figure 4: Validation and reproducibility of CTCF-associated distortion patterns and**  
3 **translational positions, related to Figure 3. A.)** Fraction accessible across  $\pm 750$  bp window centered on bound  
4 (blue) and random (red) CTCF motifs. Comparison of data from DMSO- and auxin-treated CTCF-AID mESCs  
5 indicates that acute depletion of CTCF protein is associated with a loss in protection directly at bound CTCF  
6 motifs, a decrease in mean accessibility at bound CTCF loci, and an associated loss in phased nucleosomes  
7 flanking the nucleosome-free region. **B.)** Line plot representation of mean fraction accessible for Leiden-defined  
8 clusters at CTCF motifs in CTCF-AID mESCs. **C.)** Sampled single molecules underlying the averages shown as  
9 lineplots for each cluster. Each line represents 1,500 nucleotides extracted from an individual molecule, centered  
10 at the CTCF motif. **D.)** Heatmap representation of enrichment (red) or depletion (blue) of clusters comparing  
11 fibers at bound vs. unbound CTCF motifs. Data are shown for DMSO- and auxin-treated (designated as 'IAA')  
12 CTCF-AID mESC samples, stratified by biological replicate. Black dots mark Fisher's exact tests where  $q > 0.1$   
13 (not significant). **E.)** Heatmap of Euclidian distances computed from enrichment data in (D); replicates cluster by  
14 drug treatment, suggesting that loss of CTCF protein is associated with a reproducible change in accessibility  
15 states at CTCF motif loci.  
16



1  
2  
3 **Extended Data Figure 4: (continued).** **F.)** Analysis of individual Leiden-defined clusters of nucleosome types  
4 at CTCF motifs. **Top:** Histograms of footprint sizes on a per-cluster basis. **Middle:** Mean accessibility traces  
5 centered on footprint midpoints on a per-cluster basis. **Bottom:** ‘Horizon plots’ for subfootprint sizes versus  
6 distance from footprint midpoints to assess subfootprint enrichment at particular locations on a per-cluster basis.  
7 z-scores are capped such that the maximum shown in the heatmap are the 95<sup>th</sup> percentile values within  $\pm 70$  bp of  
8 the respective footprint midpoints. **G.)** As in **(F)**, but for CTCF-associated translational positions. Accordingly,  
9 mean accessibility and ‘horizon plots’ are instead plotted relative to motif positions.



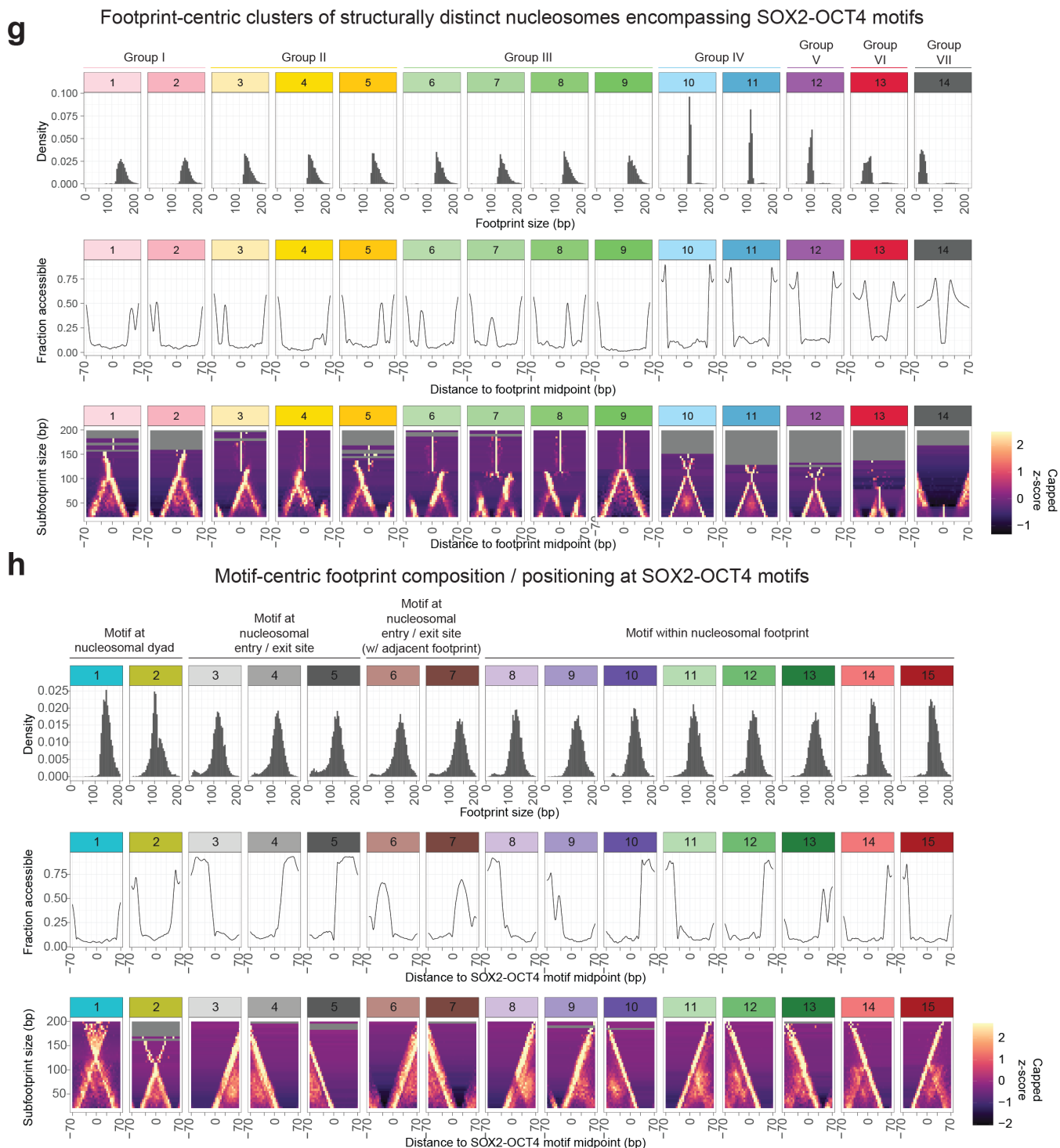
1  
2  
3 **Extended Data Figure 4: (continued). H.)** Distribution of Spearman's  $r$  values for pairwise comparisons of  
4 cluster abundances for CTCF motif-associated distortion patterns in large-scale E14 mESC dataset. A minimum  
5 of  $n = 500$  footprints was required in order to exclude lowly-sequenced libraries from this correlation analysis. **I.)**  
6 As in **(H)**, but for CTCF-associated translational positions. **J.)** Heatmap of  $\log_2$ -transformed odds ratios (ORs)  
7 for assessing enrichment / depletion of different nucleosome types at bound vs. randomly sampled CTCF motifs  
8 in DMSO-treated CTCF-AID mESCs. ORs designated with a grey dot indicate those from Fisher's exact test that  
9 are not statistically significant (*i.e.* Storey  $q$ -value  $> 0.1$ ). Correlation is computed relative to E14 mESC data  
10 (data from **Figure 3G**). **K.)** Heatmap of effect sizes ( $\Delta$ ) for nucleosome types at bound CTCF motifs in auxin vs.  
11 DMSO-treated CTCF-AID mESCs, stratified by biological replicate. Effect sizes designated with a grey dot  
12 indicate those from Fisher's exact test that are not statistically significant (*i.e.* Storey  $q$ -value  $> 0.1$ ). **L-M.)** As in  
13 **(J-K)**, but for CTCF-associated translational positions.



**Extended Data Figure 5: Validation and reproducibility of SOX2-associated distortion patterns and translational positions, related to Figure 4.** **A.)** Western blot validation of dTAG-mediated degradation of SOX2 protein indicates near-complete degradation within 30 min of dTAG<sup>V1</sup> treatment and undetectable levels of SOX2 by 2 hr and 6 hr. **B.)** Fraction accessible across  $\pm 750$  bp window centered on bound (blue) and random (red) SOX2-OCT4 composite motifs. Comparison of data from dTAG<sup>V1</sup>-NEG- and dTAG<sup>V1</sup>-treated SOX2-dTAG mESCs indicates that acute depletion of SOX2 protein is associated with a slight decrease in chromatin accessibility at SOX2-OCT4 motif loci that fall within ChIP-nexus peaks. **C.)** Line plot representation of mean fraction accessible for Leiden-defined clusters at SOX2-OCT4 composite motifs in SOX2-dTAG mESCs. **D.)** Sampled single molecules underlying the averages shown as lineplots for each cluster. Each line represents 1,500 nucleotides extracted from an individual molecule, centered at the SOX2-OCT4 motif. **E.)** Heatmap representation of enrichment (red) or depletion (blue) of clusters comparing fibers at bound vs. unbound SOX2-OCT4 composite motifs. Data are shown for dTAG<sup>V1</sup>-NEG- (designated as 'NEG') and dTAG<sup>V1</sup>-treated (time of treatment indicated) SOX2-dTAG mESC samples, stratified by biological replicate. Black dots mark Fisher's exact tests where  $q > 0.1$  (not significant). **F.)** Heatmap of Euclidian distances computed from enrichment data in (E); replicates cluster by drug treatment, suggesting that loss of SOX2 protein is associated with a reproducible change in accessibility states at SOX2-OCT4 motif loci.

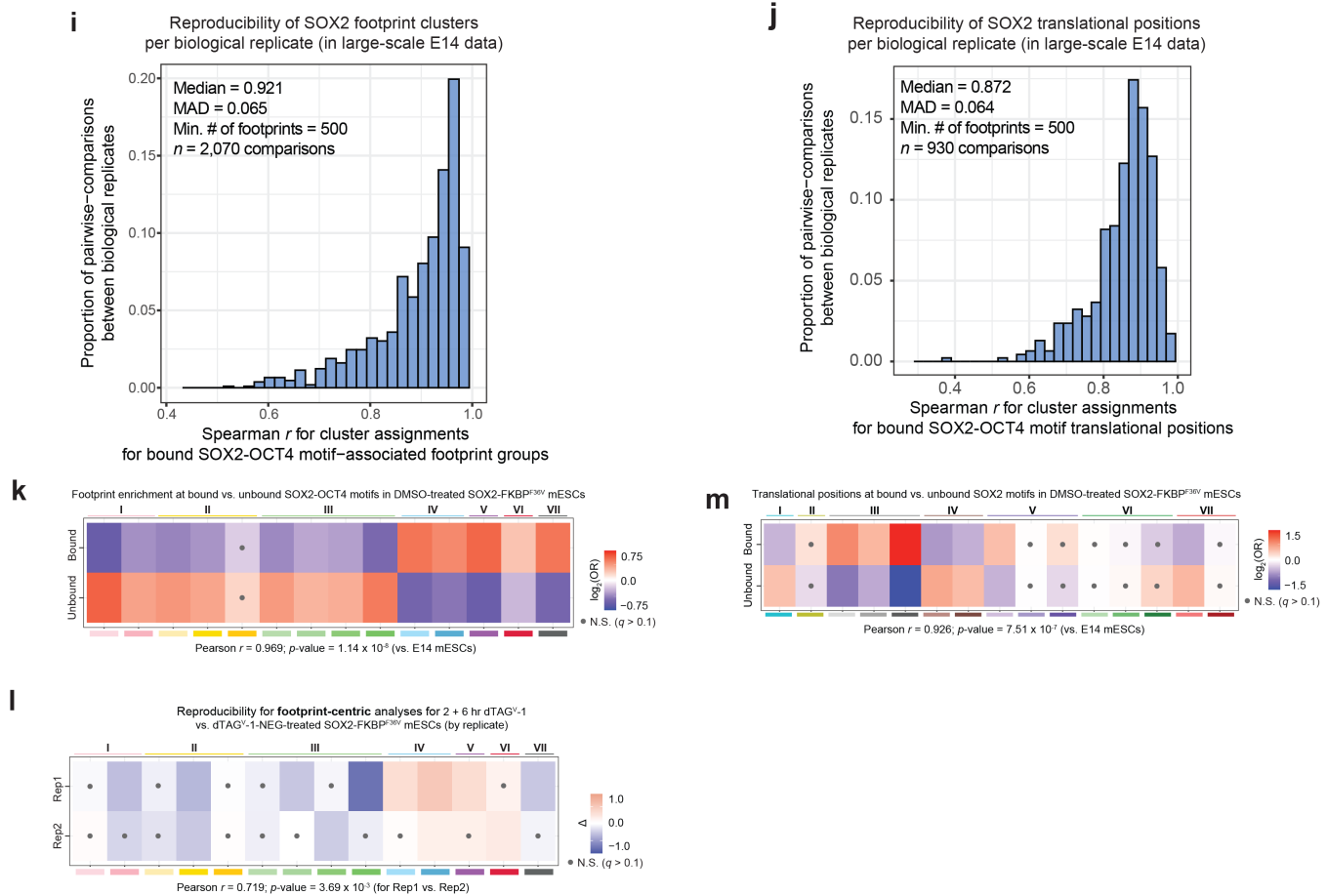
1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19



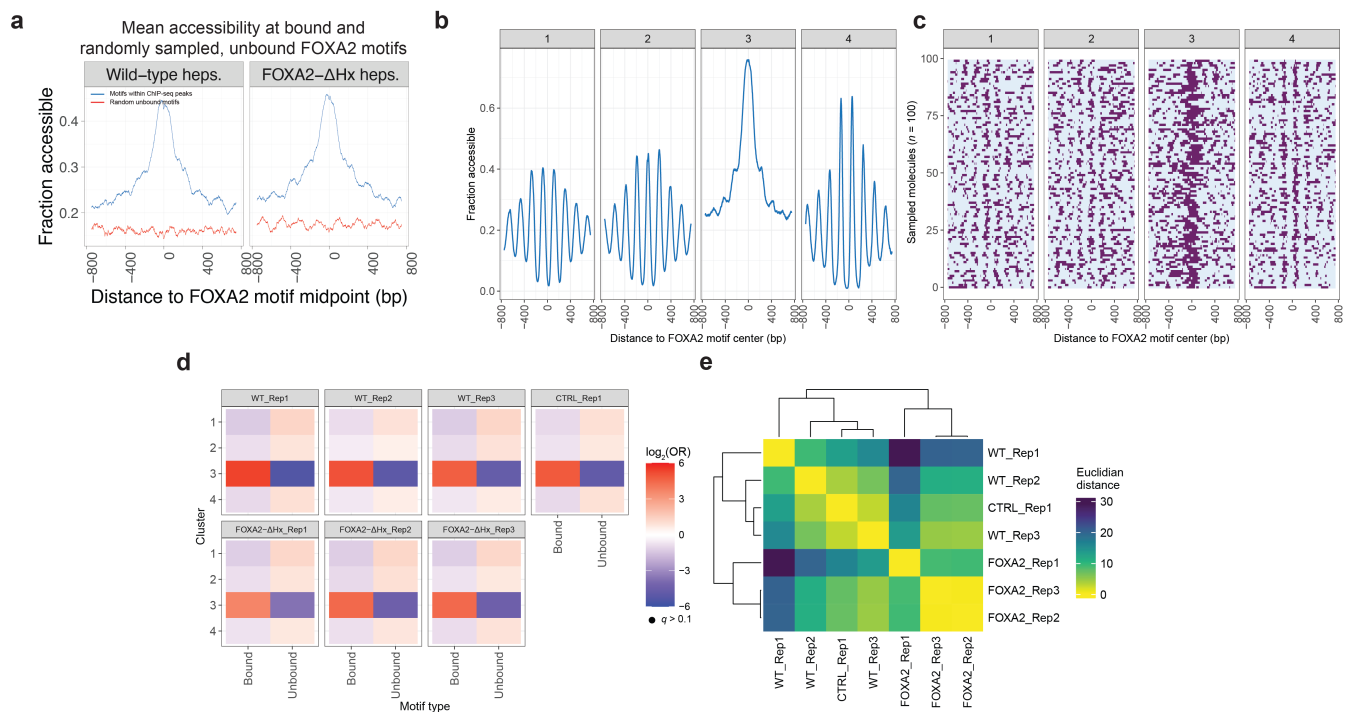


**Extended Data Figure 5: (continued). G.)** Analysis of individual Leiden-defined clusters of nucleosome types at SOX2-OCT4 composite motifs. **Top:** Histograms of footprint sizes on a per-cluster basis. **Middle:** Mean accessibility traces centered on footprint midpoints on a per-cluster basis. **Bottom:** ‘Horizon plots’ for subfootprint sizes versus distance from footprint midpoints to assess subfootprint enrichment at particular locations on a per-cluster basis. z-scores are capped such that the maximum shown in the heatmap are the 95<sup>th</sup> percentile values within  $\pm 70$  bp of the respective footprint midpoints. **H.)** As in (G), but for SOX2-associated translational positions. Accordingly, mean accessibility and ‘horizon plots’ are instead plotted relative to motif positions.

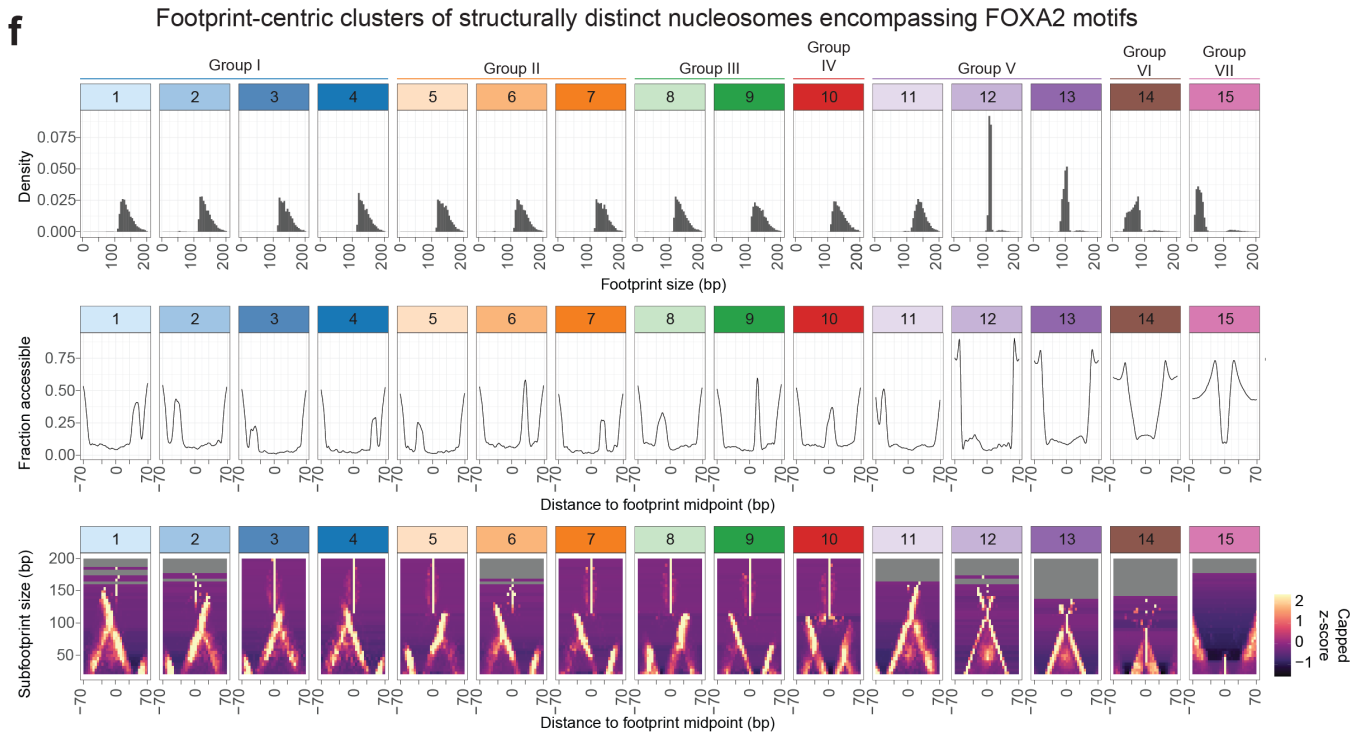
1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11



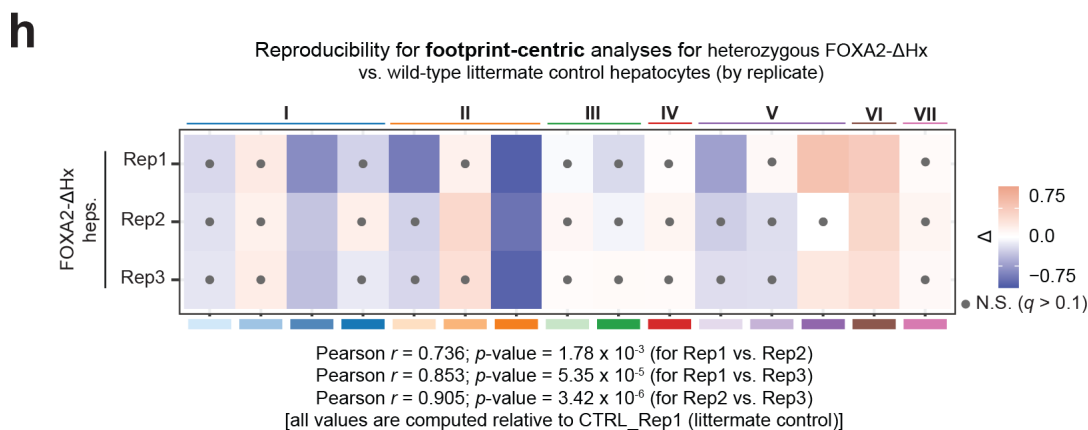
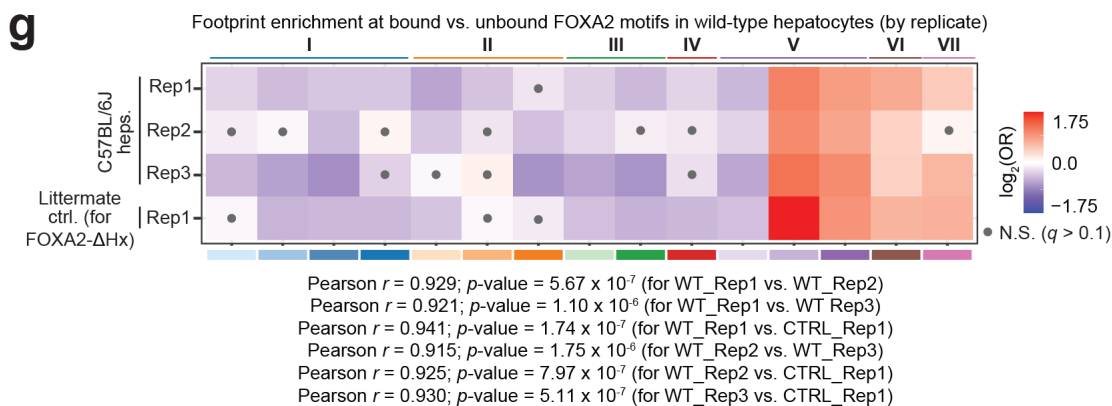
**Extended Data Figure 5: (continued). I.)** Distribution of Spearman's  $r$  values for pairwise comparisons of cluster abundances for SOX2-OCT4 motif-associated distortion patterns in large-scale E14 mESC dataset. A minimum of  $n = 500$  footprints was required in order to exclude lowly-sequenced libraries from this correlation analysis. **J.)** As in **(I)**, but for SOX2-associated translational positions. **K.)** Heatmap of  $\log_2$ -transformed odds ratios (ORs) for assessing enrichment / depletion of different nucleosome types at bound vs. randomly sampled SOX2-OCT4 composite motifs in dTAG<sup>V-1</sup>-NEG-treated SOX2-dTAG mESCs. ORs designated with a grey dot indicate those from Fisher's exact test that are not statistically significant (*i.e.* Storey  $q$ -value  $> 0.1$ ). Correlation is computed relative to E14 mESC data (data from **Figure 4C**). **L.)** Heatmap of effect sizes ( $\Delta$ ) for nucleosome types at bound SOX2-OCT4 motifs in dTAG<sup>V-1</sup>-NEG vs. dTAG<sup>V-1</sup>-treated (for 2 and 6 hr) SOX2-dTAG mESCs, stratified by biological replicate. Effect sizes designated with a grey dot indicate those from Fisher's exact test that are not statistically significant (*i.e.* Storey  $q$ -value  $> 0.1$ ). **M.)** As in **(K)**, but for SOX2-associated translational positions.



1  
2  
3 **Extended Data Figure 6: Validation and reproducibility of FOXA2-associated distortion patterns and**  
4 **translational positions, related to Figure 5. A.)** Fraction accessible across  $\pm 750$  bp window centered on bound  
5 (blue) and random (red) FOXA2 motifs. Comparison of data from wild-type and heterozygous FOXA2-ΔHx  
6 hepatocytes indicates that deletion of a FOXA2 histone-interacting domain is associated with minimal changes  
7 in mean accessibility at cognate FOXA2 binding sites. **B.)** Line plot representation of mean fraction accessible  
8 for Leiden-defined clusters at FOXA2 motifs in wild-type and heterozygous FOXA2-ΔHx hepatocytes. **C.)**  
9 Sampled single molecules underlying the averages shown as lineplots for each cluster. Each line represents 1,500  
10 nucleotides extracted from an individual molecule, centered at the FOXA2 motif. **D.)** Heatmap representation of  
11 enrichment (red) or depletion (blue) of clusters comparing fibers at bound vs. unbound FOXA2 motifs. Data are  
12 shown for wild-type (samples designated as WT are from C57BL/6J mice; sample designated as CTRL serves as  
13 littermate control for FOXA2-ΔHx samples) and heterozygous FOXA2-ΔHx hepatocyte samples, stratified by  
14 biological replicate. Black dots mark Fisher's exact tests where  $q > 0.1$  (not significant). **E.)** Heatmap of Euclidian  
15 distances computed from enrichment data in (D); replicates cluster by genotype, suggesting that accessibility  
16 patterns are more similar within wild-type samples than to helical-domain-mutant samples, and vice versa.  
17

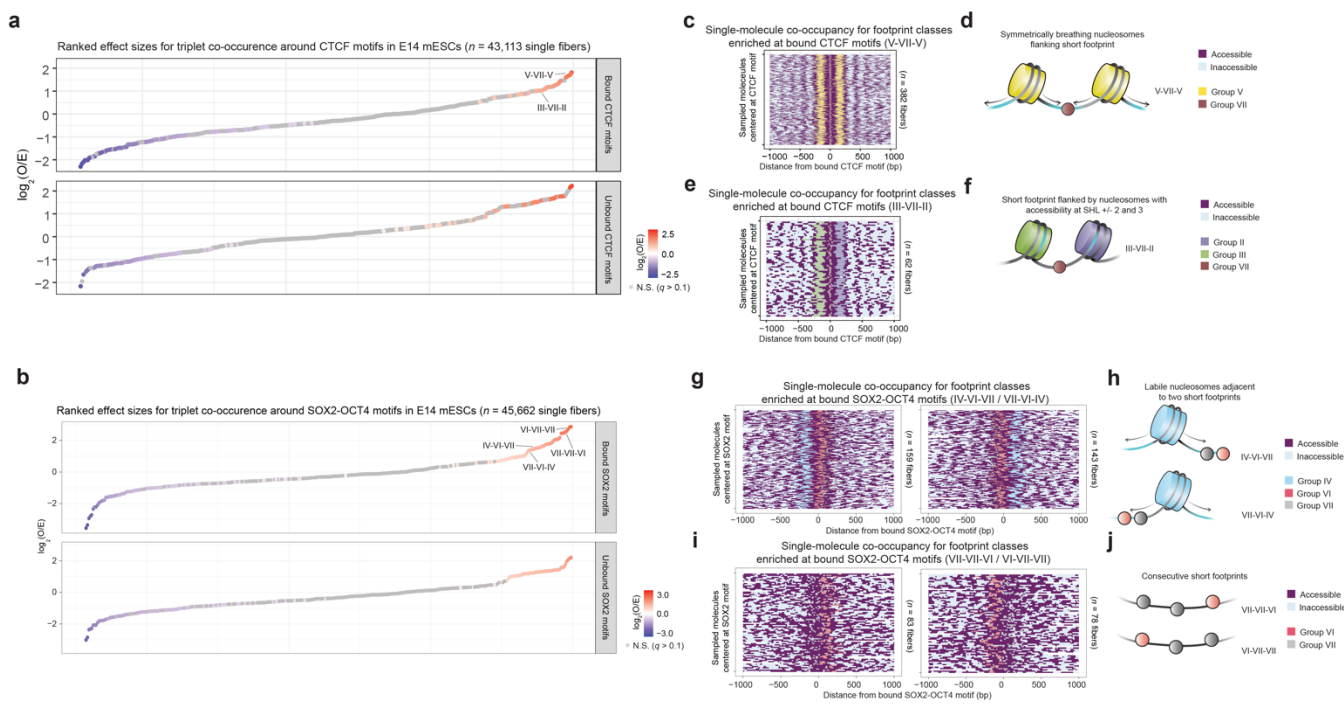


1  
2  
3 **Extended Data Figure 6: (continued). F.)** Analysis of individual Leiden-defined clusters of nucleosome types  
4 at FOXA2 motifs. **Top:** Histograms of footprint sizes on a per-cluster basis. **Middle:** Mean accessibility traces  
5 centered on footprint midpoints on a per-cluster basis. **Bottom:** ‘Horizon plots’ for subfootprint sizes versus  
6 distance from footprint midpoints to assess subfootprint enrichment at particular locations on a per-cluster basis.  
7 z-scores are capped such that the maximum shown in the heatmap are the 95<sup>th</sup> percentile values within  $\pm 70$  bp of  
8 the respective footprint midpoints.  
9



1  
2  
3  
4  
5  
6  
7  
8  
9

**Extended Data Figure 6: (continued).** **G.)** Heatmap of  $\log_2$ -transformed odds ratios (ORs) for assessing enrichment / depletion of different nucleosome types at bound vs. randomly sampled FOXA2 motifs in wild-type hepatocytes, stratified by biological replicate. ORs designated with a grey dot indicate those from Fisher's exact test that are not statistically significant (*i.e.* Storey  $q$ -value  $> 0.1$ ). **H.)** Heatmap of effect sizes ( $\Delta$ ) for nucleosome types at bound FOXA2 motifs in heterozygous FOXA2- $\Delta$ Hx hepatocytes vs. littermate control, stratified by biological replicate. Effect sizes designated with a grey dot indicate those from Fisher's exact test that are not statistically significant (*i.e.* Storey  $q$ -value  $> 0.1$ ).



**Extended Data Figure 7: Co-occupancy analysis at CTCF and SOX2-OCT4 composite motifs, related to**

**Figure 3-4. A.)** Waterfall plots of ranked effect sizes for all possible 'triplets' at bound and random CTCF motifs. Grey dots used to indicate effect sizes for 'triplets' with  $q > 0.1$ . Red and blue dots are used to indicate 'triplets' that are enriched and depleted within a given locus, respectively. Specific 'triplets' enriched at bound, but not at unbound, CTCF motifs used for single-molecule visualization are indicated. **B.)** As in **(A)**, but for 'triplets' at bound and random SOX2-OCT4 composite motifs. **C.)** Single-molecule visualization of bound CTCF-associated chromatin fibers with V-VII-V 'triplets', wherein each row represents a unique sequenced fiber (total number of molecules indicated on right side of each plot). Accessibility patterns across a 2-kb window centered on domain midpoints of interest are also depicted (purple for EcoGII-accessible and teal for EcoGII-inaccessible DNA, respectively). Footprints from different groups are colored as described in associated legend; relative widths depict the number of nucleotides protected per footprint. **D.)** Schematic for 'triplets' with symmetrically breathing nucleosomes flanking short footprint (V-VII-V), enriched at bound CTCF motifs. **E-F.)** As in **(C-D)**, but for 'triplets' comprising short footprint flanked by nucleosomes with accessibility at SHL  $\pm 2$  and  $\pm 3$  (III-VI-II), enriched at bound CTCF motifs. **G-H.)** As in **(C-D)**, but for 'triplets' comprising labile nucleosomes adjacent to two shorter footprints (IV-VI-VII / VII-VI-IV), enriched at bound SOX2-OCT4 motifs. **I-J.)** As in **(C-D)**, but for 'triplets' comprising three consecutive short footprints (VII-VII-VI / VI-VII-VII), enriched at bound SOX2-OCT4 motifs.

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20

## METHODS

### SAMOSA from E14 mESCs

We compiled a large compendium of SAMOSA data from wild-type E14 mESCs, comprising both newly generated datasets and published datasets from our laboratory, including: (1) our study on the function of ISWI-family remodeling complexes<sup>29</sup>; (2) samples prepared with our SMRT-Tag protocol, which utilizes a transposase-based strategy for PacBio library preparation from lower input amounts of footprinted DNA<sup>27</sup>; and (3) mESCs labeled with the halogenated thymidine analog, BrdU, for varying timepoints in our study of single-fiber accessibility patterns in newly-replicated chromatin<sup>33</sup>. Only samples with a mononucleosome-dinucleosome ratio (MDR; calculation detailed below in “Data analysis and visualization”) > 10 were included for analysis.

#### *Cell culture*

For newly generated libraries, E14 mESCs were grown on gelatin-coated (0.1% solution in 1x PBS) tissue culture plates. Cells were cultured in mESC media (DMEM with GlutaMAX (ThermoFisher 10566-016)), supplemented with: 15% FBS (ThermoFisher SH30071.03), 14.2 mM 2-Mercaptoethanol (Bio-Rad 1610710), 1x NEAA (ThermoFisher 11140-50), 1 mM sodium pyruvate (ThermoFisher 11360-070), and 1x LIF (purified by the laboratory of Barbara Panning (University of California, San Francisco)). Some samples were cultured with MEK (1  $\mu$ M PD0325901 (Apex Bio / Fischer A3013-25)) and GSK3 $\beta$  (3  $\mu$ M CHIR99021 (Apex Bio / Fischer A3011-100)) inhibitors (2i). Media was changed daily and cells were passaged with 1x TrypLE (ThermoFisher 12605010) when confluent. To harvest mESCs, cells were rinsed with 1x PBS (ThermoFisher 10010023), dissociated with 1x TrypLE, quenched with mESC media, and centrifuged at 500 g for 5 min. All mESC lines used in this study were regularly tested for mycoplasma contamination (Lonza LT07-318).

#### *Methyltransferase footprinting and library preparation*

Live cell counts were estimated (Countess 3) to ensure equal numbers of nuclei per unit methyltransferase were used in each footprinting reaction. Nuclei were isolated by incubating 1 x 10<sup>6</sup> live mESCs in 1 mL ice-cold NE1 Buffer (20 mM HEPES, 10 mM KCl, 1 mM MgCl<sub>2</sub>, 0.1% Triton X-100, 20% glycerol, 1x Protease Inhibitor (Roche 4693132001)) on ice for 5 min. Nuclei were pelleted by centrifugation at 500 g for 5 min at 4°C. Nuclei were rinsed in 1 mL Buffer M (15 mM Tris-HCl pH 8.0, 15 mM NaCl, 60 mM KCl, 0.5 mM spermidine) and re-pelleted by centrifugation at 500 g for 5 min at 4°C. For footprinting, nuclei were resuspended in 200  $\mu$ L Buffer M with 1 mM SAM (NEB B9003S) and 10  $\mu$ L high-concentration EcoGII (custom order from NEB). Nuclei were incubated at 37°C for 30 min, with a spike-in of 1  $\mu$ L 32 mM SAM stock halfway through the footprinting reaction. For light MNase digest to liberate chromatin fragments, footprinted nuclei were pelleted by centrifugation at 500 g for 5 min at 4°C, resuspended in 200  $\mu$ L Buffer M with 1 mM CaCl<sub>2</sub> and 0.02 U MNase (Sigma N3755), and incubated on ice for 30 min. To quench the MNase reaction, 2 mM EGTA was added.

To purify DNA, samples were treated with 10  $\mu$ L RNaseA (ThermoFisher AM2270) for 10 min at 37°C, followed by the addition of 2  $\mu$ L 10% SDS and 2  $\mu$ L 20 mg/mL Proteinase K (ThermoFisher AM2548) for >2 hr (up to overnight) at 65°C. To extract DNA, an equal volume (~250  $\mu$ L) phenol-chloroform-isoamyl alcohol was added, samples were mixed by brief vortexing, and samples were centrifuged at max speed for 2 min. The aqueous phase was transferred to a fresh tube along with 0.1x volumes 3M NaOAc, 2x volumes ice-cold 100% ethanol, and 1  $\mu$ L GlycoBlue co-precipitant (Thermo AM9516). Samples were mixed by inverting tubes and incubated overnight at -20°C, centrifuged at max speed for 30 min at 4°C, washed in 500  $\mu$ L ice-cold 70% ethanol, air dried, and resuspended in 30  $\mu$ L Buffer EB. DNA concentrations were measured by Qubit dsDNA High Sensitivity Quantification Kit (ThermoFisher Q32851). Up to 1  $\mu$ g DNA was prepared into PacBio SMRT libraries with the SMRTbell Express Template Prep Kit 2.0 per manufacturer’s directions. Fragment size distributions were estimated by Femto Pulse (Agilent). Libraries were sequenced on PacBio Sequel II 8M SMRTcells.

### SAMOSA from SOX2-dTAG mESCs

1 Degron-tagged *Sox2<sup>FKBP-F36V</sup>* (hereafter referred to as SOX2-dTAG) mESCs were provided by the laboratory of  
2 Elzo de Wit (Netherlands Cancer Institute)<sup>54</sup>. SOX2-dTAG mESCs were cultured in the same mESC media  
3 described above (including 2i) and were plated at a density of  $5 \times 10^5$  per well in 6-well plates at 24 hr before  
4 harvesting. For degradation, SOX2-dTAG mESCs were treated with  $0.5 \mu\text{M}$  dTAG<sup>V</sup>-1 (Tocris 6914) for varying  
5 timepoints (30 min, 2 hr, or 6 hr). As control, SOX2-dTAG mESC samples were treated with  $0.5 \mu\text{M}$  dTAG<sup>V</sup>-1-  
6 NEG (Tocris 6915) for 6 hr before harvesting. All experiments were performed in biological duplicate.  
7 Footprinting and downstream steps were performed as detailed in the section above, except in lieu of the MNase  
8 digest step, we sheared purified genomic DNA with Covaris g-TUBEs (Covaris 520079; 4,600 g for 6 passes).

#### 9 *Western blot for SOX2 protein in nuclear extracts*

10 To validate levels of SOX2 knockdown, mESC nuclei were isolated in NE1 Buffer, as done for matched  
11 methyltransferase-footprinted samples. Pelleted nuclei were resuspended in 100  $\mu\text{L}$  ice-cold RIPA buffer (150  
12 mM NaCl, 1% NP-40, 0.5% sodium deoxycholate, 0.1% SDS, 50 mM Tris pH 7.4, 1x Protease Inhibitor (Roche  
13 4693132001)). Samples were incubated for 30 min on ice and sonicated briefly to solubilize chromatin. Protein  
14 extracts were clarified by centrifugation at 12,000 g for 20 min at 4°C. Soluble material was collected, protein  
15 concentrations were measured by BCA (ThermoFisher 23227), and samples were diluted in NuPAGE LDS  
16 Sample Buffer (ThermoFisher NP0007) with 5% BME. Samples were reduced by boiling for 10 min at 95°C,  
17 resolved on 4-12% Bis-Tris gels (ThermoFisher NP0322), and transferred to PVDF membranes. Membranes were  
18 incubated overnight with the following primary antibodies: anti-SOX2 (CST 23064S; 1:1,000 dilution) and anti-  
19 Vinculin (Sigma V9264; 1:1,000 dilution; used as loading control). After washing, membranes were incubated  
20 with secondary antibodies conjugated to IRdye 700 or 800 (1:10,000 dilution) and imaged with the LiCor  
21 Odyssey.

#### 22 **SAMOSA from CTCF-AID mESCs**

23  
24  
25  
26 CTCF-AID mESCs (EN52.9.1) were generated by the laboratory of Elphège Nora (University of California, San  
27 Francisco)<sup>47</sup>. Raw data from CTCF-AID samples used in this study were previously published by our group<sup>33</sup>,  
28 and samples were required to have a mononucleosome-dinucleosome ratio (MDR; calculation detailed below) >  
29 10 to be included in analyses. This criteria resulted in: (1)  $n = 10$  biological replicates of auxin-treated (for 6 hr)  
30 CTCF-AID mESCs; and (2)  $n = 6$  biological replicates of DMSO-treated CTCF-AID mESCs (as controls).

#### 31 **SAMOSA from primary mouse hepatocytes**

##### 32 *Mouse husbandry and genotyping*

33  
34 All animal experiments were performed under the supervision and approval of the Institutional Animal Care and  
35 Uses Committee (IACUC) and the University of California, San Francisco (Protocol # AN179718-03F). 6-week  
36 old, C57BL/6J mice (Jackson Laboratory #000664) were used as wild-type hepatocyte samples in our dataset ( $n$   
37 = 3 biological replicates). *Foxa2<sup>ΔHx-TagRFP</sup>*-expressing mice (hereafter referred to as FOXA2-ΔHx) were provided  
38 by the laboratory of Kenneth Zaret (University of Pennsylvania)<sup>58</sup>. Samples in this study included heterozygous  
39 FOXA2-ΔHx mice ( $n = 3$ ; 1 male and 2 females) and a littermate wild-type control male mouse.

40  
41  
42 Genotyping for the presence (~230 bp) or absence (~200 bp) of the associated Tag-RFP fluorophore was  
43 performed with the following primer pair:

44 lox2272\_F: AGTGTTGTCTTCTGCCTTTGAG  
45 lox2272\_R: GCTTACCTTAGTCTCGGTCTTGG

46  
47 Genotyping for the presence (~300 bp) or absence (~270 bp) of the helical domain in FOXA2 was performed  
48 with the following primer pair:

49 TagRFP\_F: GCTCTTCGCCCTTAGACACC  
50 TagRFP\_R: ATCAGCCCCACAAAATGGAC



## 1 *Isolation of primary mouse hepatocytes*

2 Hepatocyte isolation was performed by the Liver Cell Isolation, Analysis, and Immunology Core of the UCSF  
3 Liver Center. After flushing the liver free of blood, the organ was perfused for 3 min at 4 mL/min with Liver  
4 Perfusion Medium (ThermoFisher 17701038), followed by an additional 7 min at 4 mL/min with Liver Digest  
5 Medium (ThermoFisher 17703034). The liver was then removed, diced with scissors, and suspended in DME-  
6 H21 with insulin-transferrin-selenium (ThermoFisher 41400045), penicillin-streptomycin, and 5% FBS. The  
7 crude suspension was strained through a 100- $\mu$ m filter and pelleted twice at 60 g with interval washing.  
8 Hepatocytes were purified by suspension of the pellet and centrifugation through 45% Percoll (Cytiva 17089102)  
9 at 130 g for 15 min. Purified hepatocytes, which pelleted through the Percoll, were suspended in DME-H21 as  
10 above. Final hepatocyte suspensions had a viability > 90%.

## 11 *Methyltransferase footprinting for hepatocyte samples*

12 Footprinting reaction for samples denoted HepWT\_Rep1 and HepWT\_Rep2 was performed in  $1 \times 10^6$  purified  
13 hepatocyte nuclei per library. Footprinting and downstream steps were performed as described above, except  
14 primary hepatocytes were initially dounced in NE1 Buffer (10x with loose pestle, 10x with tight pestle) to  
15 facilitate nuclear isolation.

16  
17  
18 The footprinting reaction for all remaining hepatocyte samples was performed with  $1 \times 10^6$  digitonin-  
19 permeabilized hepatocytes. Footprinting and downstream steps were performed as described above, with two  
20 notable exceptions. First, in lieu of isolating nuclei with NE1 Buffer, we directly permeabilized cells by adding  
21 0.05% digitonin (ThermoFisher BN2006) to Buffer M when performing the EcoGII footprinting reaction. Second,  
22 instead of treating nuclei with MNase to liberate chromatin fragments, we sonicated purified genomic DNA with  
23 the Megaruptor 3 (concentration = 10 ng/ $\mu$ L, volume = 100  $\mu$ L, speed = 031) to achieve more consistently-sized  
24 and longer molecules (>10 kb).

## 25 **Preprocessing PacBio data for single-molecule accessibility**

26  
27  
28 To preprocess PacBio data, we utilized software from Pacific Biosciences and a custom script  
29 (*hmm\_output\_t\_values.py*) to successively run our NN-HMM<sup>29</sup> across a range of user-specified transition  
30 probabilities. These resulted in the following output per sample: (1) an alignment of CCS reads to the mouse  
31 reference genome (mm10); (2) an accessibility prediction per CCS molecule (Viterbi path of HMM component  
32 of model); and (3) an atlas of footprint locations (*i.e.* start / end positions) and sizes on a per-molecule basis. Files  
33 for (2) and (3) were generated for  $t = 1, 31, 51, 71, \text{ and } 101 / 1,000$  for each sequencing library.

## 34 **Data analysis and visualization**

### 35 *Quantifying extent of methyltransferase footprinting per sequencing library*

36  
37 We utilized the ratio of mononucleosome-sized to dinucleosome-sized footprints (MDR) as a proxy for measuring  
38 the extent of EcoGII footprinting on a per-sample basis. Poorly methylated samples exhibit a higher fraction of  
39 footprints longer than the expected length of DNA protected by a mononucleosome (*i.e.* lower MDR values),  
40 which can be ascribed to the failure of EcoGII to efficiently methylate in linker DNA. We employed a custom  
41 script (*compute\_mdr\_per\_sample.py*), which uses the *find\_peaks* function (from *scipy*) and computes the ratio of  
42 maximum peak heights in footprint length histograms for mononucleosome- and dinucleosome-sized footprints  
43 per sequencing library.

### 44 *Visualizing single-molecule data across different t-values at single genomic loci*

45  
46 To visualize SAMOSA data at loci of interest, data were processed with a custom script  
47 (*generate\_bam\_for\_igv.py*) to encode single-molecule accessibility patterns (for any given  $t$ -value) as an  
48 additional flag in aligned *.bam* files. These *.bam* files were concatenated across E14 mESC samples with MDR >  
49 10 and visualized at the *Sox2* locus (chr3: 34,756,929 – 34,759,161; including the *Sox2* promoter, gene body, and  
50 downstream SCR) using a custom version of IGV (<https://github.com/RamaniLab/SMRT-Tag/tree/main/igv-vis>).

1 Bulk ATAC-seq data from mESCs were obtained as processed .bw files (GSE98390) and visualized at identical  
2 coordinates on the UCSC Genome Browser.

### 3 4 *Identifying fibers that overlap annotated epigenomic domains in mESCs*

5 A custom script from our group ([https://github.com/RamaniLab/SAMOSAS-ChAAT/blob/main/scripts/zmw\\_selector.py](https://github.com/RamaniLab/SAMOSAS-ChAAT/blob/main/scripts/zmw_selector.py)) was adapted to identify sequencing reads where a portion of the read  
6 falls within  $\pm 1$  kb of midpoints of histone post-translational modification-defined epigenomic domains in mESCs  
7 (*i.e.* H3K4me1, H3K4me3, H3K9me3, H3K27me3, and H3K36me3 ChIP-seq peaks; derived from ENCODE  
8 data).  
9

### 10 11 *Identifying fibers that overlap repeat elements in mouse genome*

12 A custom script from our group ([https://github.com/RamaniLab/SAMOSAS-ChAAT/blob/main/scripts/blast\\_bam.sh](https://github.com/RamaniLab/SAMOSAS-ChAAT/blob/main/scripts/blast_bam.sh)) was used to identify sequencing reads with one or more matches to  
13 repeat elements of interest. In brief, we ran BLAST on CCS reads from a database consisting of mouse major  
14 satellite sequence (M32564.1), minor satellite sequence (X14462.1), and telomeric DNA sequence (pentameric  
15 repeats of TTAGGG). For fibers with multiple repeat matches, only the repeat sequence with the lowest E-value  
16 was considered for downstream analyses.  
17

### 18 19 *Identifying fibers that overlap bound and unbound transcription factor motif matches*

20 To define instances of TF-bound motifs, we referenced the following datasets: ChIP-nexus data for SOX2 in  
21 mESCs (GSE137193)<sup>53</sup>, ChIP-seq data for CTCF in mESCs (ENCFF508CKL), and ChIP-seq data for FOXA2 in  
22 mouse hepatocytes (GSE157452)<sup>59</sup>. For SOX2 motifs, we retained only the subset of peaks that contained a  
23 SOX2-OCT4 composite motif match as defined by this dataset<sup>68</sup>. For CTCF data, we utilized a curated set of  
24 ChIP-seq backed motifs initially processed in this study<sup>45</sup>. For FOXA2 data, we retained only the subset of peaks  
25 that contained a PWM match (MA0047.2) for the FOXA2 motif<sup>69</sup>. For control sites, we randomly sampled an  
26 equal number of motif matches per factor of interest from a previously published list of motifs across the mouse  
27 genome<sup>68</sup>, requiring that these motifs fall at least 1 kb away from the nearest ChIP-annotated peak. As above with  
28 epigenomic domains, we utilized a custom script to identify sequencing reads where a portion of the read falls  
29 within  $\pm 1$  kb of motif midpoints.  
30

### 31 32 *Clustering single-molecule accessibility patterns with respect to footprint midpoints*

33 To examine patterns of nucleosomal distortion, we clustered accessibility patterns across a range of  $t$ -values with  
34 respect to footprint midpoints. Footprints  $< 200$  nt in length (*i.e.* intended to capture both nucleosomal and  
35 subnucleosomal species) within  $\pm 500$  bp of midpoints of epigenomic domains, repeat sequences, or TF-binding  
36 motifs were selected. On a per-footprint basis, accessibility data (from  $t = 31, 51, \text{ and } 71 / 1,000$ ) for a 140-bp  
37 window centered on each footprint midpoint (*i.e.* 420 bp in total per footprint) was used as input for Leiden  
38 clustering ( $res = 0.5, 0.6, 0.6, \text{ and } 0.6$  for domains / repeats, SOX2, CTCF, and FOXA2 motifs, respectively).  
39 Clusters were filtered such that the least abundant clusters that collectively summed up to  $< 10\%$  of each dataset  
40 were removed. It should be noted that: (1) accessibility data was oriented accounting for sequence feature strand  
41 before clustering to preserve any potential directional effects associated with non-palindromic TF-binding motifs;  
42 (2) this approach was intended to enable clustering on accessibility information across multiple  $t$ -values per  
43 footprint; and (3) this window size was selected to prioritize signal from SHL -7 to SHL +7 for nucleosomal  
44 footprints. Representative code for performing this analysis is provided as a custom script  
45 (*process\_footprints\_nucleosomal\_distortion.py*).

### 46 47 *Clustering single-molecule accessibility patterns with respect to sequence features of interest*

48 To define translational settings, we clustered accessibility patterns across a range of  $t$ -values with respect to  
49 midpoints of a given sequence feature of interest. Footprints  $< 200$  nt in length wherein the absolute distance  
50 between the footprint midpoint and sequence feature was less than half the total length of the footprint in question  
51 (*e.g.* within  $\pm 100$  bp for a 200 nt footprint) were selected. On a per-footprint basis, accessibility data (from  $t = 31,$   
51 51, and 71 / 1,000) for a 140-bp window centered on each sequence feature (*i.e.* 420 bp in total per footprint) was

1 used as input for Leiden clustering ( $res = 0.3, 0.8, 0.8, \text{ and } 0.8$  for repeats, SOX2, CTCF, and FOXA2 motifs,  
2 respectively). Clusters were filtered such that the least abundant clusters that collectively summed up to  $<10\%$  of  
3 each dataset were removed. Representative code for performing this analysis is provided as a custom script  
4 (*process\_footprints\_translational\_positions.py*).

#### 6 *Plotting footprint size distributions, mean accessibility patterns, and 'horizon plots' per cluster*

7 To assign putative biological function to clusters of nucleosomal distortion patterns or translational positions, we  
8 visualized the following features. First, we plotted the distribution of footprint sizes on a per-cluster basis for data  
9 from the lowest  $t$ -value examined ( $t = 1 / 1,000$ ). Second, we plotted mean accessibility as a function of distance  
10 to respective footprint / sequence feature midpoints from the highest  $t$ -value used for Leiden clustering ( $t = 71 /$   
11  $1,000$ ). Third, we generated 'horizon plots' (analogous to commonly used 'V plots' in MNase-seq data) by  
12 computing z-scores for enrichment of footprints of particular size (20 – 200 nt) at specific positions ( $\pm 70$  bp from  
13 footprint / sequence feature midpoints) for each individual cluster. z-scores are summed per size-position  
14 combination across the range of  $t$ -values previously used for clustering ( $t = 31, 51, \text{ and } 71 / 1,000$ ). Color scale  
15 limits for the resultant heatmap are capped by the 0th (lower) and 95th (upper) percentile z-scores for positions  
16 within  $\pm 70$  bp of the footprint / sequence feature midpoint. Clusters of entirely unmethylated footprints were  
17 manually filtered out before downstream quantification. Representative code for performing this analysis is  
18 provided as a custom script (*plot\_footlen\_meanacc\_horizondata.py*).

#### 20 *Identifying groups of clusters with shared nucleosomal distortion patterns / translational positions*

21 To group Leiden assignments for nucleosomal distortion patterns or translational positions in an unbiased manner,  
22 we performed hierarchical clustering of 'horizon plot' data as follows. z-scores per cluster were filtered such that  
23 only those corresponding to footprints  $<100$  nt and  $<50$  nt in length were retained for defining nucleosomal  
24 distortion and translational position groups, respectively. This choice was intended to prioritize clustering of  
25 'horizon plot' data specifically within subfootprint size ranges that capture subnucleosomal organization. Data  
26 were manually inspected and z-scores flipped with respect to footprint / sequence feature midpoints to account  
27 for the nucleosomal dyad as an axis of symmetry. The distance matrix was computed with the *dist* function (in  
28 R), hierarchical clustering was performed with *hclust* (from *factoextra* package; *method = ward.D2*), and output  
29 was visualized as a dendrogram (with *fviz\_dend*).

#### 31 *UMAP visualization of footprint types within epigenomic domains and at repeat sequences*

32 From accessibility data for footprints within histone modification-defined domains and at mouse repeat elements,  
33 we employed Scanpy (v1.9.3) for PCA-based dimensionality reduction, construction of a  $k$ -nearest neighbors  
34 graph (*metric = correlation*, *n\_neighbors = 15*), and UMAP visualization (*min\_dist = 0.5*). Clusters were colored  
35 according to Leiden assignments and proportions of each Leiden-defined cluster were represented as a pie chart  
36 surrounding the UMAP visualization.

#### 38 *Structural visualization of nucleosomal accessibility patterns*

39 For structural visualization of mean accessibility patterns from defined clusters, we utilized a custom script  
40 (*accessibility\_to\_pdb\_structure.py*) to convert our mean single-molecule accessibility data per cluster of interest  
41 from the highest  $t$ -value used for Leiden clustering ( $t = 71 / 1,000$ ) into a format compatible with visualization  
42 via ChimeraX (v1.7.1; *attribute = percentAcc*, *match mode = 1-to-1*, and *recipient = residues*)<sup>70</sup>. Accessibility  
43 values were overlaid on the following PDB structure: 7KBD<sup>71</sup>.

#### 45 *Computing enrichment of nucleosomal distortion patterns / translational positions across distinct genomic loci*

46 To compute odds-ratios for Leiden-defined clusters across epigenomic domains and repeat elements, or for bound  
47 vs. randomly-sampled TF motifs, Fisher's exact tests were carried out with *scipy* (in Python). All computed  $p$ -  
48 values were corrected with a Storey  $q$ -value correction, using the *qvalue* package (in R).

#### 50 *Computing effect sizes associated with TF knockdown or across genotypes*

1 To compute effect sizes upon degnon-mediated TF knockdown or across mouse genotypes, Fisher's exact tests  
2 were carried out for footprints specifically at TF-bound motifs (e.g. footprints from dTAG-V1 vs. DMSO-treated  
3 SOX2-dTAG mESCs) with *scipy* (in Python). All computed *p*-values were corrected with a Storey *q*-value  
4 correction, using the *qvalue* package (in R).

#### 6 *Reproducibility of nucleosomal distortion patterns or translational positions across replicates*

7 To assess reproducibility in our large-scale E14 mESC dataset, we asked if the rank-order of clusters of  
8 nucleosomal distortion patterns or translational positions is preserved across sequencing libraries. To do this, the  
9 relative proportion of each Leiden-defined cluster at domains / repeats or at TF-bound motifs was computed on a  
10 per-sample basis and analyzed using Spearman's correlation (with *spearmanr* from *scipy*). Footprints from all  
11 sequencing libraries with MDR > 10 were included in Leiden clustering of accessibility patterns; this choice was  
12 intended to include as many footprints as possible to increase statistical power. However, for these correlation  
13 analyses, we imposed a minimum sequencing depth for a biological replicate to be included, with the goal of  
14 assessing reproducibility in sufficiently-sequenced samples (minimum number of footprints indicated on each  
15 respective plot).

17 To evaluate reproducibility in SOX2-dTAG mESCs, CTCF-AID mESCs, and FOXA2-ΔHx hepatocytes, we  
18 calculated the Pearson's correlation (with *pearsonr* from *scipy*) for effect sizes computed on a per-sample basis.  
19 This choice was intended to capture potential quantitative changes in cluster proportions between degnon  
20 conditions / genotypes across replicates.

#### 22 *Identifying single-molecule co-occupancy patterns for different footprint groups*

23 Individual fibers with at least three Leiden-classified footprints within ±500 bp of repeat elements / TF motifs of  
24 interest were selected. Fibers with footprints >200 nt and/or with footprints in the least abundant 10% of clusters  
25 were excluded from further analysis. Consecutive 'triplet' footprints were tabulated by computing midpoints and  
26 lengths for: (1) the central-most footprint nearest the repeat element / TF-binding motif; (2) the most proximal  
27 footprint upstream of the central-most footprint; and (3) the most proximal footprint downstream of the central-  
28 most footprint. Per-molecule accessibility data was computed for a 2-kb window centered on the repeat element  
29 / TF-binding motif of interest for visualization alongside footprint positions, accounting for sequence feature  
30 strand appropriately. Representative code for performing this analysis is provided (*process\_triplet\_footprints.py*).

#### 32 *Computing enrichment of consecutive 'triplets' on single chromatin fibers*

33 Individual clusters of nucleosomal distortion patterns were combined based on groups defined by shared  
34 subfootprint organization from hierarchical clustering, as described above. Expected frequencies of footprint  
35 'triplets' were computed as the product of the independent probabilities of observing each group within a given  
36 dataset. These values were used to calculate log<sub>2</sub>(O/E) ratios for each 'triplet' across epigenomic domains, repeat  
37 elements, or TF-binding motifs. Enrichment was assessed using Fisher's exact test (*fisher.test* function in R), with  
38 FDR correction applied for multiple comparisons. The results were visualized as a ranked-order scatterplot,  
39 reflecting the effect size for each 'triplet'.

#### 41 *Visualizing mean accessibility patterns at bound vs. random transcription factor motifs*

42 Previously-published scripts from our group were used to compute and cluster mean accessibility observed within  
43 ±750 bp of TF-binding motifs across different conditions (i.e. before / after TF knockdown; across genotypes)<sup>33</sup>.  
44 Euclidian distances for resultant odds-ratios for bound vs. unbound motifs per sample are plotted to visualize their  
45 similarity across conditions and replicates.

## 47 **DATA AND CODE AVAILABILITY**

48 Raw and processed data will be made available at GEO accession GSEXXXXXX. Custom scripts used for  
49 analyses are available at <https://github.com/RamaniLab/IDL1>.

## 51 **ACKNOWLEDGEMENTS**

1 We thank members of the Ramani Lab, Siva Kasinathan (Stanford), Hiten Madhani (UCSF), Geeta Narlikar  
2 (UCSF), and Srinivas Ramachandran (University of Colorado) for helpful discussions and comments on the  
3 manuscript. We acknowledge Kenneth Zaret (University of Pennsylvania) for sharing Foxa2- $\Delta$ Hx mice, Jackson  
4 Goudreau and Michael Mui for assistance with mouse husbandry, Tami Tolpa for generating schematics used in  
5 our figures, and support from core facilities within the UCSF Liver Center (P30-DK026743). MGY was supported  
6 by NIH T32-DK060414 and by the Gladstone-CIRM Scholars Program. This work was partially supported by  
7 NIH grants U01-DK127421 and DP2-HG012442 to VR. VR acknowledges generous support from the Searle  
8 Scholars Award and the W.M. Keck Foundation.

## 10 REFERENCES

- 11 1. van Holde, K. E. *Chromatin*. (Springer, New York, NY, 2012).
- 12 2. Arimura, Y., Tachiwana, H., Oda, T., Sato, M. & Kurumizaka, H. Structural analysis of the hexasome, lacking  
13 one histone H2A/H2B dimer from the conventional nucleosome. *Biochemistry* **51**, 3302–3309 (2012).
- 14 3. Rhee, H. S., Bataille, A. R., Zhang, L. & Pugh, B. F. Subnucleosomal structures and nucleosome asymmetry  
15 across a genome. *Cell* **159**, 1377–1388 (2014).
- 16 4. Ramachandran, S., Ahmad, K. & Henikoff, S. Transcription and Remodeling Produce Asymmetrically  
17 Unwrapped Nucleosomal Intermediates. *Mol. Cell* **68**, 1038-1053.e4 (2017).
- 18 5. Engholm, M. *et al.* Nucleosomes can invade DNA territories occupied by their neighbors. *Nat. Struct. Mol.*  
19 *Biol.* **16**, 151–158 (2009).
- 20 6. Kato, D. *et al.* Crystal structure of the overlapping dinucleosome composed of hexasome and octasome.  
21 *Science* **356**, 205–208 (2017).
- 22 7. Klein, D. C., Troy, K., Tripplehorn, S. A. & Hainer, S. J. The esBAF and ISWI nucleosome remodeling  
23 complexes influence occupancy of overlapping dinucleosomes and fragile nucleosomes in murine embryonic  
24 stem cells. *BMC Genomics* **24**, 201 (2023).
- 25 8. Nishimura, M. *et al.* Genome-wide mapping and cryo-EM structural analyses of the overlapping tri-  
26 nucleosome composed of hexasome-hexasome-octasome moieties. *Commun. Biol.* **7**, 61 (2024).
- 27 9. Lowary, P. T. & Widom, J. New DNA sequence rules for high affinity binding to histone octamer and  
28 sequence-directed nucleosome positioning. *J. Mol. Biol.* **276**, 19–42 (1998).
- 29 10. Segal, E. *et al.* A genomic code for nucleosome positioning. *Nature* **442**, 772–778 (2006).
- 30 11. Kaplan, N. *et al.* The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature* **458**, 362–366  
31 (2009).
- 32 12. Sinha, K. K., Gross, J. D. & Narlikar, G. J. Distortion of histone octamer core promotes nucleosome  
33 mobilization by a chromatin remodeler. *Science* **355**, eaaa3761 (2017).
- 34 13. Hsieh, L. J. *et al.* A hexasome is the preferred substrate for the INO80 chromatin remodeling complex,  
35 allowing versatility of function. *Mol. Cell* **82**, 2098-2112.e4 (2022).
- 36 14. Wu, H. *et al.* Reorientation of INO80 on hexasomes reveals basis for mechanistic versatility. *Science* **381**,  
37 319–324 (2023).
- 38 15. Isbel, L., Grand, R. S. & Schübeler, D. Generating specificity in genome regulation through transcription  
39 factor sensitivity to chromatin. *Nat. Rev. Genet.* **23**, 728–740 (2022).
- 40 16. Ricci, M. A., Manzo, C., Garcia-Parajo, M. F., Lakadamyali, M. & Cosma, M. P. Chromatin fibers are formed  
41 by heterogeneous groups of nucleosomes in vivo. *Cell* **160**, 1145–1158 (2015).
- 42 17. Ou, H. D. *et al.* ChromEMT: Visualizing 3D chromatin structure and compaction in interphase and mitotic  
43 cells. *Science* **357**, eaag0025 (2017).
- 44 18. Pownall, M. E. *et al.* Chromatin expansion microscopy reveals nanoscale organization of transcription and  
45 chromatin. *Science* **381**, 92–100 (2023).
- 46 19. Henikoff, J. G., Belsky, J. A., Krassovsky, K., MacAlpine, D. M. & Henikoff, S. Epigenome characterization  
47 at single base-pair resolution. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 18318–18323 (2011).
- 48 20. Schep, A. N. *et al.* Structured nucleosome fingerprints enable high-resolution mapping of chromatin  
49 architecture within regulatory regions. *Genome Res.* **25**, 1757–1770 (2015).
- 50 21. Poirier, M. G., Bussiek, M., Langowski, J. & Widom, J. Spontaneous access to DNA target sites in folded  
51 chromatin fibers. *J. Mol. Biol.* **379**, 772–786 (2008).

- 1 22. Poirier, M. G., Oh, E., Tims, H. S. & Widom, J. Dynamics and function of compact nucleosome arrays. *Nat. Struct. Mol. Biol.* **16**, 938–944 (2009).
- 2
- 3 23. Kelly, T. K. *et al.* Genome-wide mapping of nucleosome positioning and DNA methylation within individual  
4 DNA molecules. *Genome Res.* **22**, 2497–2506 (2012).
- 5 24. Kilgore, J. A., Hoose, S. A., Gustafson, T. L., Porter, W. & Kladde, M. P. Single-molecule and population  
6 probing of chromatin structure using DNA methyltransferases. *Methods* **41**, 320–332 (2007).
- 7 25. Krebs, A. R. *et al.* Genome-wide single-molecule footprinting reveals high RNA polymerase II turnover at  
8 paused promoters. *Mol. Cell* **67**, 411–422.e4 (2017).
- 9 26. Abdulhay, N. J. *et al.* Massively multiplex single-molecule oligonucleosome footprinting. *Elife* **9**, (2020).
- 10 27. Nanda, A. S. *et al.* Direct transposition of native DNA for sensitive multimodal single-molecule sequencing.  
11 *Nat. Genet.* (2024) doi:10.1038/s41588-024-01748-0.
- 12 28. Stergachis, A. B., Debo, B. M., Haugen, E., Churchman, L. S. & Stamatoyannopoulos, J. A. Single-molecule  
13 regulatory architectures captured by chromatin fiber sequencing. *Science* **368**, 1449–1454 (2020).
- 14 29. Abdulhay, N. J. *et al.* Nucleosome density shapes kilobase-scale regulation by a mammalian chromatin  
15 remodeler. *Nat. Struct. Mol. Biol.* (2023) doi:10.1038/s41594-023-01093-6.
- 16 30. Moore, C. *et al.* ATP-dependent remodeling of chromatin condensates uncovers distinct mesoscale effects  
17 of two remodelers. *bioRxiv* (2024) doi:10.1101/2024.09.10.611504.
- 18 31. Ostrowski, M. S. *et al.* The single-molecule accessibility landscape of newly replicated mammalian  
19 chromatin. *Cell* (2024) doi:10.1016/j.cell.2024.10.039.
- 20 32. Polach, K. J. & Widom, J. Mechanism of protein access to specific DNA sequences in chromatin: a dynamic  
21 equilibrium model for gene regulation. *J. Mol. Biol.* **254**, 130–149 (1995).
- 22 33. Ostrowski, M. S. *et al.* The single-molecule accessibility landscape of newly replicated mammalian  
23 chromatin. *bioRxiv* 2023.10.09.561582 (2023) doi:10.1101/2023.10.09.561582.
- 24 34. King, H. W., Fursova, N. A., Blackledge, N. P. & Klose, R. J. Polycomb repressive complex 1 shapes the  
25 nucleosome landscape but not accessibility at target genes. *Genome Res.* **28**, 1494–1507 (2018).
- 26 35. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected  
27 communities. *Sci. Rep.* **9**, 5233 (2019).
- 28 36. Becht, E. *et al.* Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* **37**,  
29 38–44 (2018).
- 30 37. Armache, J. P. *et al.* Cryo-EM structures of remodeler-nucleosome intermediates suggest allosteric control  
31 through the nucleosome. *Elife* **8**, (2019).
- 32 38. Saunders, H. S. *et al.* HMGB1 restores a dynamic chromatin environment in the presence of linker histone  
33 by deforming nucleosomal DNA. *bioRxiv* 2024.08.23.609244 (2024) doi:10.1101/2024.08.23.609244.
- 34 39. Rao, S., Ahmad, K. & Ramachandran, S. Cooperative binding between distant transcription factors is a  
35 hallmark of active enhancers. *Mol. Cell* **81**, 1651–1665.e4 (2021).
- 36 40. Sönmezer, C. *et al.* Molecular co-occupancy identifies transcription factor binding cooperativity in vivo.  
37 *Mol. Cell* **81**, 255–267.e6 (2021).
- 38 41. Zaret, K. S. Pioneer transcription factors initiating gene network changes. *Annu. Rev. Genet.* **54**, 367–385  
39 (2020).
- 40 42. Hansen, J. L., Loell, K. J. & Cohen, B. A. A test of the pioneer factor hypothesis using ectopic liver gene  
41 activation. *Elife* **11**, e73358 (2022).
- 42 43. de Wit, E. & Nora, E. P. New insights into genome folding by loop extrusion from inducible degron  
43 technologies. *Nat. Rev. Genet.* **24**, 73–85 (2023).
- 44 44. Hansen, A. S., Pustova, I., Cattoglio, C., Tjian, R. & Darzacq, X. CTCF and cohesin regulate chromatin loop  
45 stability with distinct dynamics. *Elife* **6**, 2848 (2017).
- 46 45. Ramani, V., Qiu, R. & Shendure, J. High sensitivity profiling of chromatin structure by MNase-SSP. *Cell*  
47 *Rep.* **26**, 2465–2476.e4 (2019).
- 48 46. Yue, F. *et al.* A comparative encyclopedia of DNA elements in the mouse genome. *Nature* **515**, 355–364  
49 (2014).
- 50 47. Nora, E. P. *et al.* Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from  
51 Genomic Compartmentalization. *Cell* **169**, 930–944.e22 (2017).

- 1 48. Nakahashi, H. *et al.* A genome-wide map of CTCF multivalency redefines the CTCF code. *Cell Rep.* **3**,  
2 1678–1689 (2013).
- 3 49. Grand, R. S. *et al.* Genome access is transcription factor-specific and defined by nucleosome position. *Mol.*  
4 *Cell* (2024) doi:10.1016/j.molcel.2024.08.009.
- 5 50. Soufi, A. *et al.* Pioneer transcription factors target partial DNA motifs on nucleosomes to initiate  
6 reprogramming. *Cell* **161**, 555–568 (2015).
- 7 51. Dodonova, S. O., Zhu, F., Dienemann, C., Taipale, J. & Cramer, P. Nucleosome-bound SOX2 and SOX11  
8 structures elucidate pioneer factor function. *Nature* **580**, 669–672 (2020).
- 9 52. Michael, A. K. *et al.* Mechanisms of OCT4-SOX2 motif readout on nucleosomes. *Science* **368**, 1460–1465  
10 (2020).
- 11 53. Avsec, Ž. *et al.* Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nat. Genet.*  
12 **53**, 354–366 (2021).
- 13 54. Maresca, M. *et al.* Pioneer activity distinguishes activating from non-activating SOX2 binding sites. *EMBO*  
14 *J.* **42**, e113150 (2023).
- 15 55. Nocente, M. C. *et al.* cBAF generates subnucleosomes that expand OCT4 binding and function beyond DNA  
16 motifs at enhancers. *Nat. Struct. Mol. Biol.* (2024) doi:10.1038/s41594-024-01344-0.
- 17 56. Barisic, D., Stadler, M. B., Iurlaro, M. & Schübeler, D. Mammalian ISWI and SWI/SNF selectively mediate  
18 binding of distinct transcription factors. *Nature* **569**, 136–140 (2019).
- 19 57. Iurlaro, M. *et al.* Mammalian SWI/SNF continuously restores local accessibility to chromatin. *Nat. Genet.*  
20 **53**, 279–287 (2021).
- 21 58. Iwafuchi, M. *et al.* Gene network transitions in embryos depend upon interactions between a pioneer  
22 transcription factor and core histones. *Nat. Genet.* **52**, 418–427 (2020).
- 23 59. Qu, M., Qu, H., Jia, Z. & Kay, S. A. HNF4A defines tissue-specific circadian rhythms by beaconing  
24 BMAL1::CLOCK chromatin binding and shaping the rhythmic chromatin landscape. *Nat. Commun.* **12**,  
25 6350 (2021).
- 26 60. Whitney, P. H. & Lionnet, T. The method in the madness: Transcriptional control from stochastic action at  
27 the single-molecule scale. *Curr. Opin. Struct. Biol.* **87**, 102873 (2024).
- 28 61. Gilpatrick, T. *et al.* Targeted nanopore sequencing with Cas9-guided adapter ligation. *Nat. Biotechnol.* **38**,  
29 433–438 (2020).
- 30 62. Kreibich, E., Kleinendorst, R., Barzaghi, G., Kaspar, S. & Krebs, A. R. Single-molecule footprinting  
31 identifies context-dependent regulation of enhancers by DNA methylation. *Mol. Cell* **83**, 787-802.e9 (2023).
- 32 63. Martin, B. J. E. *et al.* Global identification of SWI/SNF targets reveals compensation by EP400. *Cell* **186**,  
33 5290-5307.e26 (2023).
- 34 64. Meers, M. P., Janssens, D. H. & Henikoff, S. Pioneer factor-nucleosome binding events during differentiation  
35 are motif encoded. *Mol. Cell* **75**, 562-575.e5 (2019).
- 36 65. Zaret, K. S. & Mango, S. E. Pioneer transcription factors, chromatin dynamics, and cell fate control. *Curr.*  
37 *Opin. Genet. Dev.* **37**, 76–81 (2016).
- 38 66. Tullius, T. W. *et al.* RNA polymerases reshape chromatin architecture and couple transcription on individual  
39 fibers. *Mol. Cell* **84**, 3209-3222.e5 (2024).
- 40 67. Zhou, H. *et al.* Quantitative spatial analysis of chromatin biomolecular condensates using cryo-electron  
41 tomography. *bioRxiv* (2024) doi:10.1101/2024.12.01.626131.
- 42 68. Vierstra, J. *et al.* Global reference mapping of human transcription factor footprints. *Nature* **583**, 729–736  
43 (2020).
- 44 69. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics*  
45 **27**, 1017–1018 (2011).
- 46 70. Pettersen, E. F. *et al.* UCSF ChimeraX: Structure visualization for researchers, educators, and developers.  
47 *Protein Sci.* **30**, 70–82 (2020).
- 48 71. Arimura, Y., Shih, R. M., Froom, R. & Funabiki, H. Structural features of nucleosomes in interphase and  
49 metaphase chromosomes. *Mol. Cell* **81**, 4377-4397.e12 (2021).