

1 Single-cell multi-scale footprinting reveals the modular organization of DNA regulatory 2 elements

3
4 Yan Hu^{1,2}, Sai Ma^{1,2,4}, Vinay K. Kartha^{1,2}, Fabiana M. Duarte^{1,2}, Max Horlbeck^{1,2}, Ruochi Zhang^{1,2}, Rojesh
5 Shrestha^{1,2}, Ajay Labade^{1,2}, Heidi Kletzien^{2,3}, Alia Meliki^{1,2}, Andrew Castillo^{1,2}, Neva Durand¹, Eugenio Mattei¹,
6 Lauren J. Anderson¹, Tristan Tay^{1,2}, Andrew S. Earl^{1,2}, Noam Shoresh¹, Charles B. Epstein¹, Amy Wagers^{2,3},
7 Jason D. Buenrostro^{1,2,#}

8
9 ¹Gene Regulation Observatory, Broad Institute of MIT and Harvard, Cambridge, MA, 02142 USA.

10 ²Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, MA, 02138 USA.

11 ³Paul F. Glenn Center for the Biology of Aging, Harvard Medical School, Boston, MA 02115.

12 ⁴Current address: Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai,
13 New York, NY 10029 USA.

14
15 #Correspondence: jason_buenrostro@harvard.edu

16
17 Keywords: epigenomics, single-cell, gene regulation, transcription factors, aging

18 **Abstract:**

19 *Cis*-regulatory elements control gene expression and are dynamic in their structure, reflecting changes to the
20 composition of diverse effector proteins over time¹⁻³. Here we sought to connect the structural changes at *cis*-
21 regulatory elements to alterations in cellular fate and function. To do this we developed PRINT, a
22 computational method that uses deep learning to correct sequence bias in chromatin accessibility data and
23 identifies multi-scale footprints of DNA-protein interactions. We find that multi-scale footprints enable more
24 accurate inference of TF and nucleosome binding. Using PRINT with single-cell multi-omics, we discover
25 wide-spread changes to the structure and function of candidate *cis*-regulatory elements (cCREs) across
26 hematopoiesis, wherein nucleosomes slide, expose DNA for TF binding, and promote gene expression.
27 Activity segmentation using the co-variance across cell states identifies “sub-cCREs” as modular cCRE
28 subunits of regulatory DNA. We apply this single-cell and PRINT approach to characterize the age-associated
29 alterations to cCREs within hematopoietic stem cells (HSCs). Remarkably, we find a spectrum of aging
30 alterations among HSCs corresponding to a global gain of sub-cCRE activity while preserving cCRE
31 accessibility. Collectively, we reveal the functional importance of cCRE structure across cell states,
32 highlighting changes to gene regulation at single-cell and single-base-pair resolution.
33
34

35 **Introduction:**

36 Through homeostasis, development, and disease, *cis*-regulatory elements change in structure and recruit new
37 regulatory proteins, which define the overall function of the element⁴. In this process, *cis*-regulatory elements
38 act as hubs of gene regulation to establish primed, activated, or repressed genes and determine the overall
39 function and potency of cells^{1,3,5,6}. These structural changes are largely mediated by the competition of
40 nucleosomes, TFs and transcriptional machinery, which dynamically slide, evict and recruit effector proteins^{7,8}.
41 Despite this rich understanding of the biochemical activities occurring on DNA, in genomics individual *cis*-
42 regulatory elements are often studied as discrete functional units, motivating a need for genomic tools that
43 trace chromatin structure at single-base-pair resolution.
44

45 Methods that measure chromatin accessibility have revealed a diverse repertoire of cCREs⁶. Additionally, DNA
46 footprinting methods elucidate TF binding at cCREs by quantifying the protection of DNA from chemical⁹ or
47 enzymatic¹⁰⁻¹³ cleavage, yielding base-pair resolved maps of diverse proteins bound to DNA¹⁴. Using high-
48 throughput DNA sequencing, footprinting is now performed genome-wide, revealing the function of non-coding

49 genetic variation¹⁵ and improving the construction of gene regulatory networks^{16–18}. However, despite best
50 efforts, footprinting methods are afflicted with sequence bias severely limiting accuracy¹⁹. Further,
51 computational methods for footprinting are optimized to detect binding of TFs of typical size, excluding the
52 analysis of nucleosomes or atypical TFs. These limitations preclude our ability to measure intra-cCRE
53 structural dynamics that reflect changes to the composition and function of cCREs over time.

54
55 Here, we develop PRINT (Protein-Regulatory element Interactions at Nucleotide resolution using
56 Transposition), a framework that i) accurately corrects for sequence bias in chromatin accessibility data, ii)
57 computes the interaction of DNA with objects of various sizes (multi-scale footprinting) and iii) leverages
58 single-cell multi-omics to identify the structural changes to cCREs and their impact on gene expression. Using
59 this approach, we show that DNA bound proteins, including TFs and nucleosomes, create unique cleavage
60 patterns and demonstrate that multi-scale footprints enable accurate prediction of TF binding genome-wide.
61 Next, we combine multi-scale footprinting with single-cell multi-omic data across human hematopoiesis to
62 track TF and nucleosome binding dynamics across differentiation. We discover wide-spread restructuring of
63 cCREs during differentiation, wherein nucleosomes slide, expose new sites for TF binding and promote gene
64 expression. We refer to the genomic regions that modularly expand and shrink within cCREs as “sub-cCREs”
65 and show that sub-cCREs can explain changes to gene expression in the absence of overt changes to
66 chromatin accessibility. Finally, we define sub-cCRE dynamics in response to aging of mouse hematopoietic
67 stem cells (HSCs). Here, we find that many age-associated promoter alterations increase TF binding of sub-
68 cCREs, while maintaining the overall accessibility of the element. Overall, we find sub-CREs as regulators of
69 gene expression and cell state, revealing a unique structure of gene regulation at single-cell and single-base-
70 pair resolution.

71 72 **Multi-scale footprinting enables detection of DNA binding by factors of diverse** 73 **sizes**

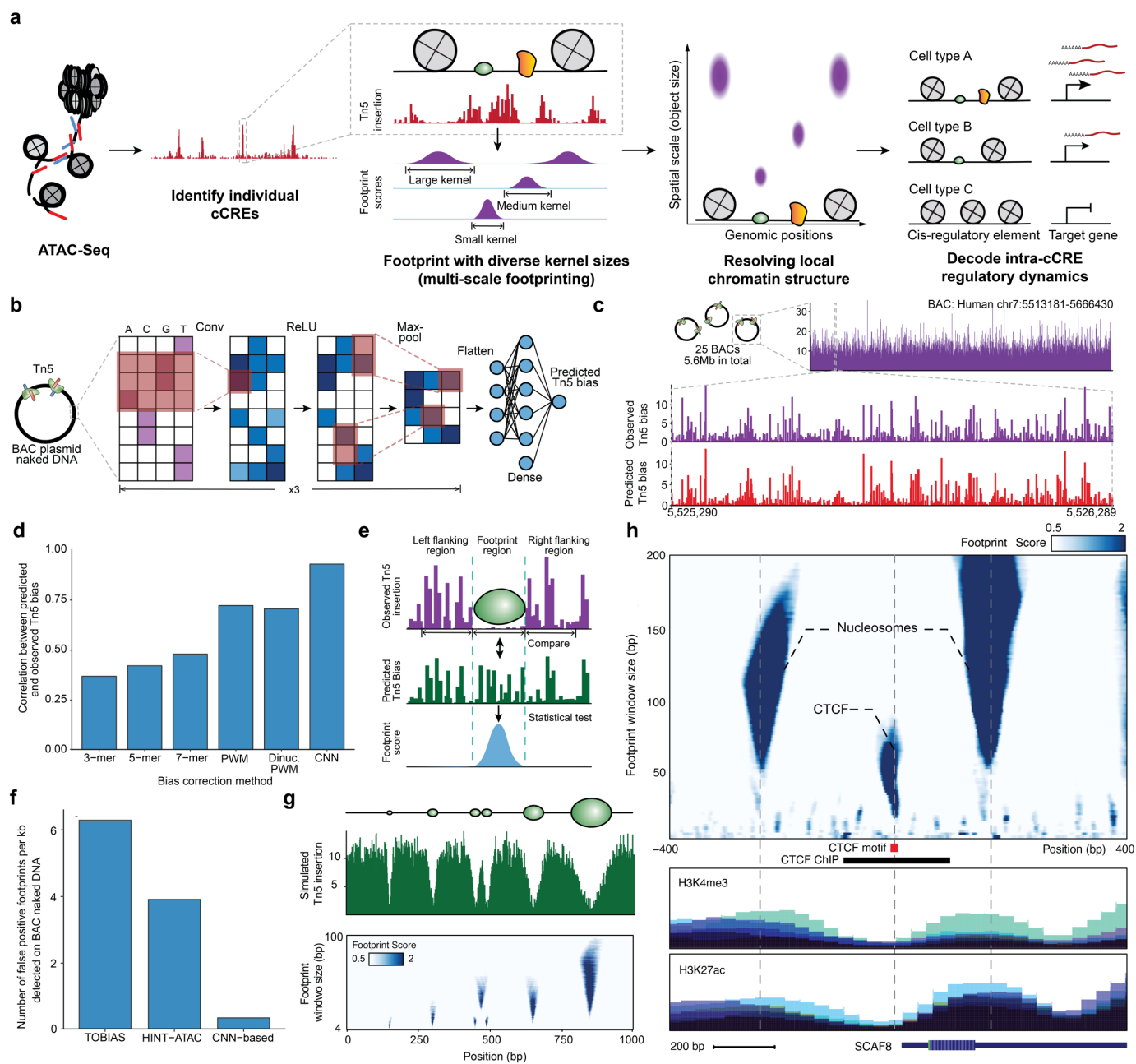
74 To enable tracking of structural changes within cCREs we developed PRINT, a computational approach to
75 footprint proteins of diverse sizes (**Fig. 1a**). To do this, we sought to use bulk or single-cell ATAC-seq data as
76 input. However, Tn5 transposase has a strong sequence preference^{13,20}, which may significantly confound
77 footprint detection. To evaluate and create approaches for modeling Tn5 sequence bias, we generated high-
78 coverage Tn5 insertion data on deproteinized DNA from bacterial artificial chromosomes (BACs) containing a
79 total of 5.6 Mb of the human genome (**Extended Data Table 1**). A total of 193.2 million reads aligned to the
80 BACs, resulting in 34.5 Tn5 insertions per base-pair. We also performed 5 biological replicates and found that
81 the observed Tn5 bias is highly reproducible ($R > 0.97$, **Extended Data Fig. S1a-c**).

82
83 Using the BAC data, we trained a convolutional neural network that takes as input DNA sequence and predicts
84 Tn5 sequence preference (**Fig. 1b**). We found that deep learning achieved a correlation of 0.94 between
85 predicted and observed bias, significantly outperforming k-mer and PWM models (**Fig. 1c, d**) while achieving
86 the highest improvements in regions of high GC-content (**Extended Data Fig. S1d, e**). Exemplifying the utility
87 of modeling Tn5 preference, we provide Tn5 bias prediction for the entire human genome, alongside common
88 model organisms including *Pan troglodytes*, *Mus musculus*, *Drosophila melanogaster*, *Saccharomyces*
89 *cerevisiae*, *Caenorhabditis elegans* and *Danio rerio*, covering a total of ~11B bases of DNA sequence. We
90 also provide a pre-trained deep learning model that can be extended to any new species or applied to personal
91 genomes (see **Data Availability**).

92
93 To call footprints, we developed a statistical approach for footprinting that quantifies the depletion of observed
94 Tn5 insertions relative to the Tn5 sequence bias, resulting in a footprint score representing the statistical

95 significance ($-\log_{10} p$ -value) for each base pair position (**Fig. 1e**, Methods). Using our deproteinized BAC data
96 as a control, we detected little to no footprint signal using our approach on naked DNA (**Extended Data Fig.**
97 **S1f-o**). In contrast, prior footprinting methods^{21,22} report up to 35,262 false positive footprints within the 5.6
98 Mb BAC regions, corresponding to an average false positive rate of 23% across all TFs (**Fig. 1f**). The Tn5
99 bias model and statistical approach described here reduced the number of false positive footprints by
100 approximately one order of magnitude (**Fig. 1f**), demonstrating that bias correction is essential for accurate
101 footprinting.

102
103 Finally, we explored footprinting across spatial scales to detect DNA-bound proteins of different sizes. We
104 performed footprinting, using simulated data (**Fig. 1g**) and ATAC-seq data (**Fig. 1h**), with window sizes ranging
105 between 4-200 base pairs and observed drastically different footprint patterns corresponding to TF and
106 nucleosomes (**Fig. 1h**). Therefore, we reasoned that multi-scale footprinting may fractionate molecular
107 interactions at different scales and outline the local physical structure of chromatin.



108
109
110
111
112
113
114
115
116
117
118
119
120

Figure 1. Multi-scale footprinting detects DNA-protein interactions at various spatial scales. a, Overview of the multi-scale footprinting workflow. **b,** Schematic illustration of the Tn5 bias prediction model. **c,** Single-nucleotide resolution tracks of observed and predicted Tn5 bias on naked DNA in the BAC RP11-93G19. **d,** Bar plot comparing performance of the CNN model with previous bias correction models. **e,** Schematic illustration of footprint score calculation. **f,** Bar plot showing the frequency of calling false positive footprints by previous ATAC-footprinting methods and our method. **g,** Multi-scale footprints with simulated objects. Top: schematic of simulated objects with various sizes. Middle: Simulated single base pair resolution Tn5 insertion tracks based on the above objects. Bottom: Heatmap showing the multi-scale footprints calculated based on the simulated Tn5 insertions. The horizontal axis represents single base pair positions, and the vertical axis represents footprint window sizes. **h,** Multi-scale footprints in the cCRE region chr6:154732871-154733870. Bottom tracks are histone ChIP signals obtained from ENCODE.

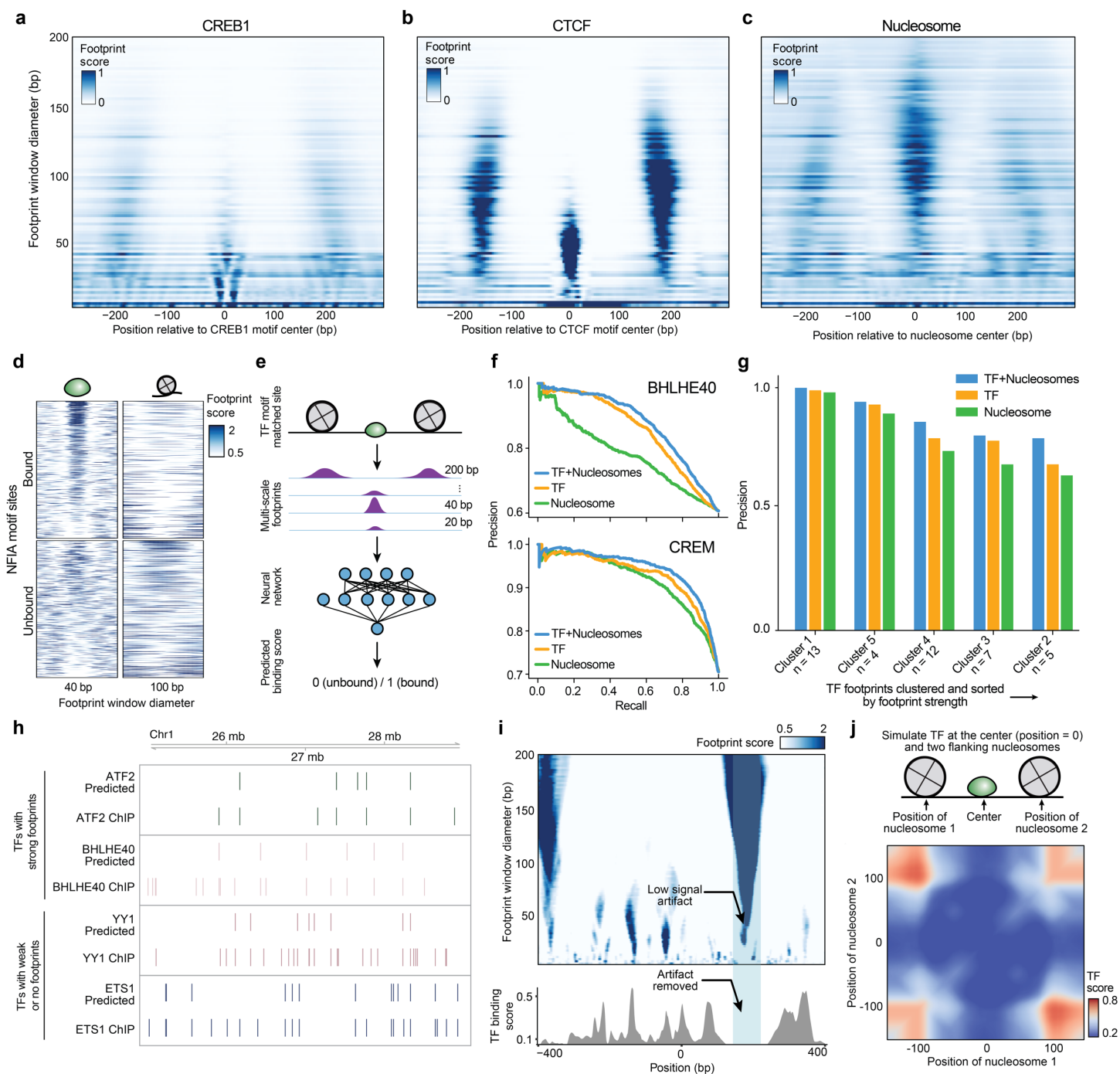
121 **TFs and nucleosomes have signature multi-scale footprints**

122 Inspired by the diversity of structures seen in multi-scale footprints, we sought to categorize proteins by
123 footprint sizes and shapes. We obtained TF ChIP-seq data⁶ and generated aggregate multi-scale footprints
124 to find that TFs may leave small (CREB1) and large (CTCF) footprints (**Fig. 2a-b, Extended Data Fig. S2a-**
125 **c**). We found that TFs (n = 112) clustered into 6 distinct groups based on their size, shape, and footprint
126 strength (**Extended Data Fig. S2d**). We found that the majority of TFs in cluster 1, 4, and 5 leave visible
127 footprints (n = 71) at 20 bp and 40 bp scales, whereas TFs in cluster 2, 3, and 6 (n = 41) leave weak or no
128 footprints at these same size scales. We also validated that footprints of larger sizes (100-140bp) correspond
129 to prior measures of nucleosome position^{23,24} (**Fig. 2c**). In summary, multi-scale footprints reveal diverse DNA-
130 protein interactions, enabling the analysis of both TFs and nucleosomes in one computational approach.

131
132 Motivated by these observations, we trained a neural network classifier that uses multi-scale footprints and
133 motif positions as input to predict TF binding (**Fig. 2d, e and Extended Data Fig. S3a-c**). Provided that cluster
134 1 TFs leave the strongest footprints, we first trained the model predicting TF binding of cluster 1 TFs using
135 multiscale footprints. The model achieved a median precision of 0.71 on held out test ChIP data²⁵,
136 outperforming prior methods (0.65 for HINT-ATAC and 0.62 for TOBIAS when benchmarked at a matched
137 recall, **Fig. 2f and Extended Data Table 2**).

138
139 We next sought to extend this approach to TFs that leave weak or undetectable footprints. We trained a new
140 model using data from all 6 clusters of TFs. As many TFs (37%) do not leave clear footprints, this model further
141 prioritizes nucleosome position for TF binding prediction. As such we refer to this model as the “TF habitation
142 model” and its prediction scores as “TF habitation scores”. The TF habitation model achieved a median
143 precision of 0.76 for cluster 1 TFs and 0.67 across all TFs on held out K562 data (n = 41), again outperforming
144 previous methods (0.58 for HINT-ATAC and 0.59 for TOBIAS, **Extended Data Fig. S3d, e**). We next tested
145 model performance on primary cell samples, expanding the analysis to 91 TF binding data datasets in total⁶.
146 The model achieved a median precision of 0.73 across all TFs while recovering hundreds to thousands of
147 binding sites per TF (**Fig. 2g-h and Extended Data Table 2**). Additionally, we determined a 0.8% false positive
148 rate using the BAC data. In conclusion, using multi-scale footprinting, we developed an approach that
149 accurately predicts protein-DNA interactions at multiple length scales.

150
151 Upon further investigation we find that these models use nearby nucleosomes, together with TF footprints, to
152 improve predictions (**Fig. 2i, j**). As expected, the model uses a high TF footprint signal (40 bp) at the motif
153 center and low nucleosome signal (100-200 bp) surrounding the motif (**Extended Data Fig. S3f**) for prediction.
154 Additionally, we found frequent cases wherein nucleosome footprints span lower (40bp) scales, but are
155 correctly identified as artifacts by the model (**Fig. 2i**). Interestingly, the model also found that high nucleosome
156 signal distal to the motif (~100-150 bp) is predictive of TF binding. To further explore the model, we simulated
157 TF and nucleosome footprint within a 300 bp window to find that TF binding scores significantly decrease as
158 nucleosomes approach the TF motif (**Fig. 2j**) or as nucleosomes become delocalized or “fuzzy” (**Extended**
159 **Data Fig. S3g**). To quantitatively assess improvements, we performed ablation tests wherein TF or
160 nucleosome footprints are removed during training (**Fig. 2f, g**). Using this approach, we observed decreased
161 precision after ablating nucleosomes and found that nucleosomes, without TF footprints, may be highly
162 predictive of TF binding (e.g., CREM). Altogether this indicates that nucleosome position strongly influences
163 TF binding.



164
165
166
167
168
169
170
171
172
173

Figure 2. TFs and nucleosomes binding leave signature multi-scale footprint patterns. **a, b**, Multi-scale aggregate footprints for TFs CREB1 and CTCF. The x-axis represents the position relative to the center of the TF motif, and the y-axis represents footprint scores computed using each footprint window size. **c**, Multi-scale aggregate footprints for nucleosomes. The x-axis represents the position relative to the center of the nucleosome as determined by chemical mapping, and the y-axis represents footprint scores computed using each footprint window size. **d**, Multi-scale footprints around individual bound and unbound NFIA motif sites. Each row represents a single locus with a matched NFIA motif. **e**, Schematic illustration of training TF binding prediction models using multi-scale footprints as input. **f-g** Ablation test results. **f**, Example precision-recall curves of cluster 1-specific models trained without masking, with TF masking, and with nucleosome masking,

174 respectively. **g**, Bar plot showing precision of the TF habitation model when trained without masking, with TF
175 masking, and with nucleosome masking, respectively. **h**, Comparison between predicted and ChIP-detected
176 TF binding sites. Only sites with a matched TF motif are included. **i**, Top: heatmap showing multi-scale
177 footprints within the cCRE at chr11:67629937-67630936. The x-axis represents single base pair positions in
178 the cCRE, and the y-axis represents footprint window size. Bottom: predicted TF binding scores within the
179 same region. **j**, Heatmap showing predicted TF habitation score for different simulated TF and nucleosome
180 configurations. Horizontal and vertical axes represent the distances of the two simulated nucleosomes from
181 the center TF.
182

183 **Emerging modular structures of intra-cCRE dynamics**

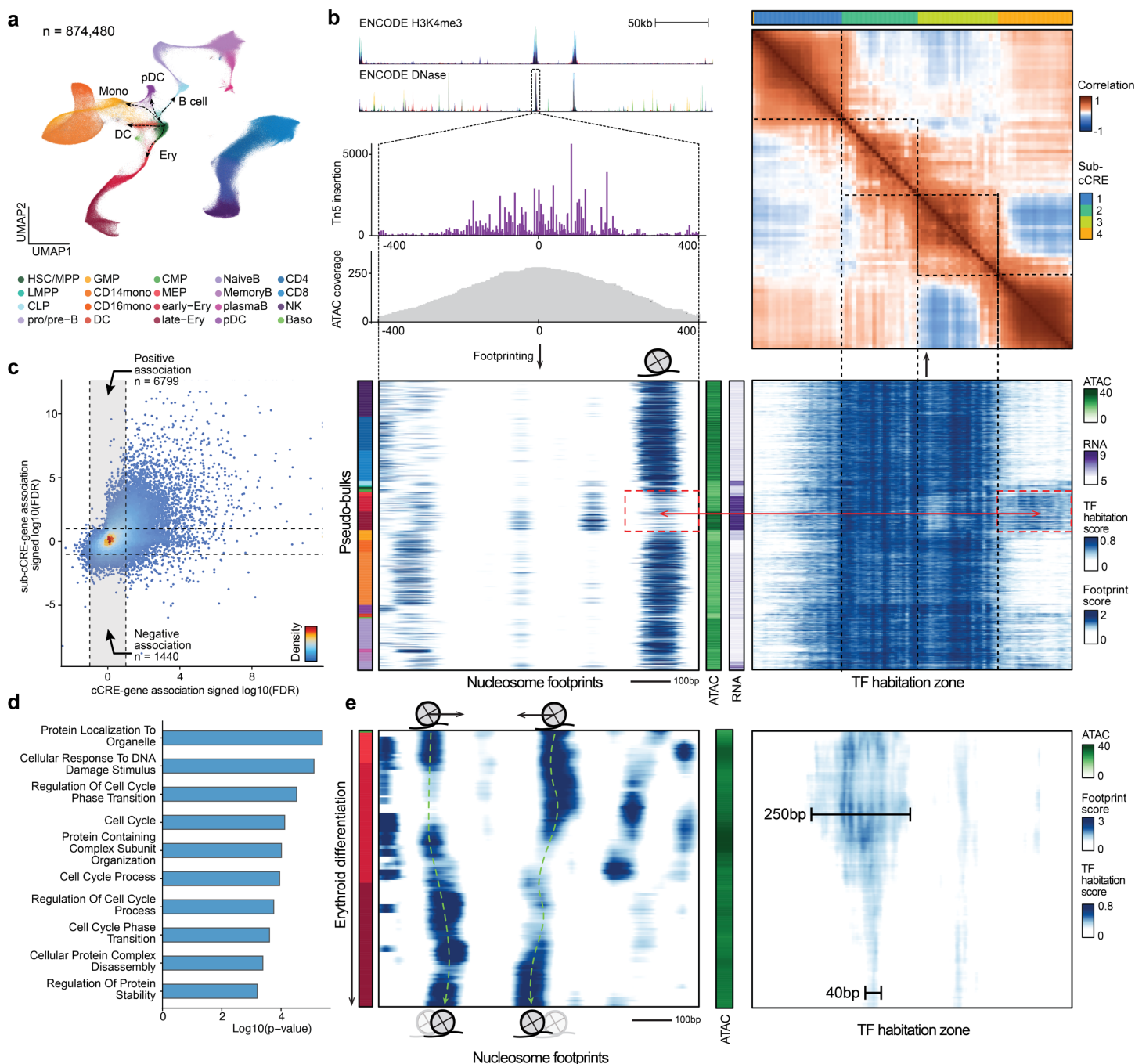
184 We reasoned that a single-cell multi-omic analysis of footprints would enable pseudo-time-resolved tracking
185 of protein-binding and connect these changes to alterations in gene expression. To generate multi-omic data
186 at a throughput and depth needed for footprinting, we used SHARE-seq²⁶ (ATAC and RNA) to profile 874,480
187 total cells from 7 human bone marrow donors. The resulting data represents a total of 935,959,306 nuclear
188 ATAC fragments and 608,148,224 RNA UMIs across all major hematopoietic cell types, including
189 hematopoietic stem cells (HSCs) and differentiated cell types (**Fig. 3a, Extended Data Fig. S4a-e**).

190
191 Using these single-cell data we sought to define the dynamics at a sub-cCRE scale using PRINT. We
192 generated 1,000 pseudo-bulks encompassing all major cell types and major developmental transitions
193 (**Extended Data Fig. S4f-i, Methods**). Next, we applied multi-scale footprinting and our TF habitation model
194 to these hematopoietic pseudo-bulks. Within individual cCREs, we observed modular structures reflecting gain
195 or loss of TF habitation scores across pseudo-bulks, which we refer to as “sub-cCREs” (**Fig 3b, Extended**
196 **Data Fig. S5a, b**). To quantitatively detect such sub-cCREs, we computed the correlation of TF habitation
197 scores between all positions within each cCRE. The results again show modular structures within the cCRE,
198 as exemplified by regions showing strong off-diagonal correlation (**Fig. 3b, top-right panel, Extended Data**
199 **Fig. S5c, d**). Using such intra-cCRE correlation maps as input, we designed an algorithm to segment each
200 cCRE into sub-cCREs with strong self-association. As a result, we detected 265,070 sub-cCREs across
201 human hematopoiesis. We observed a positive association between cCRE accessibility and the number of
202 sub-cCREs detected inside the cCRE. For the top 10,000 accessible cCREs, we detected on average 3.7
203 sub-cCREs within each cCRE and the average size of a sub-cCREs is 211.9 bp, which is approximately the
204 size of a nucleosome flanked with linker DNA. As a result, cCREs do not appear to have a fixed boundary or
205 size, but instead shrink, expand, merge or split as modular sub-cCREs lose or gain activity across
206 differentiation (**Extended Data Fig. S5e, f**).

207
208 We next sought to examine if the activity of independent sub-cCREs is associated with gene expression
209 variation across cell types. For every cCRE, we computed the correlations between the accessibility of the
210 cCRE and the RNA levels of nearby genes (+/-50 kb). We next computed correlations between the activity of
211 each sub-cCRE (as defined by the average TF habitation score with the sub-cCRE) within this cCRE with the
212 same nearby genes. Interestingly, cCRE-gene and sub-cCRE-gene correlations show divergent association
213 (**Fig. 3c**). A total of 8,239 sub-cCREs were significantly correlated to gene-expression while the corresponding
214 cCREs were not (permutation test, FDR < 0.1). In these cases we find that cCREs re-organize TF binding while
215 maintaining overall accessibility of the cCRE (**Fig. 3b, bottom panels, Extended Data Fig. S5g, h**).
216 Furthermore, at these regions we observed strong enrichment in pathways related to cell cycle, proteostasis,
217 and DNA damage response, suggesting a unique mode of regulation for such pathways (**Fig. 3d, Extended**
218 **Data Table 3**).

219

220 Sub-cCREs are approximately ~200 bp in size, similar to the size of nucleosomes plus linker regions. Hence,
 221 we hypothesized that the activity of sub-cCREs is driven by the dynamics of nucleosomes. To further explore
 222 this idea, we tracked nucleosome positioning and sub-cCRE dynamics across pseudo-time along erythroid
 223 differentiation. We observed nucleosome dynamics in the form of binding, eviction, as well as sliding
 224 accompanied by sub-cCRE activation/repression at the same locus (**Fig. 3e, Extended Data Fig. S5i, j**),
 225 providing evidence for nucleosome reorganization during native human hematopoiesis.



226 **Figure 3. Emerging modular structures of intra-cCRE dynamics.** **a**, UMAP of the human bone marrow
 227 SHARE-seq dataset. **b**, Example of sub-cCREs. Top left: Tracks showing chromatin accessibility and single-
 228 base pair resolution Tn5 insertion in the cCRE at chr4:173334022-173335021. Bottom left and bottom right:

229 Heatmap of nucleosome footprints (100 bp scale) and TF habitation scores in the same region across all
230 pseudo-bulks, respectively. Each row corresponds to a single pseudo-bulk, while each column represents a
231 single base pair position in the cCRE. Left color bar shows the cell type annotation of each pseudo-bulk. Color
232 palette is the same as in **a**. Middle color bars show total accessibility within the cCRE and RNA level of the
233 gene *HMGB2* in each pseudo-bulk, respectively. Top right: Heatmap showing correlation of TF habitation
234 scores between any two positions within the cCRE. Top color bar shows results of automatic segmentation of
235 the cCRE into sub-cCREs. **c**, Scatter plot comparing cCRE-gene correlation and sub-cCRE-gene correlation.
236 For each cCRE, the sub-cCRE with the strongest correlation is selected. Dashed lines represent the FDR
237 threshold of 0.1. **d**, Bar plot showing pathway enrichment of genes with significant sub-cCRE-gene correlation
238 but not cCRE-gene correlation (FDR < 0.1). **e**, Nucleosome tracking across erythroid differentiation. Left:
239 Heatmap of nucleosome footprints in the region chr7:99471434-99472433 across pseudo-bulks in the
240 erythroid lineage. Pseudo-bulks are ordered by pseudo-time. Right: heatmap of TF habitation scores in the
241 same region and pseudo-bulks. Left color bar shows the cell type annotation of each pseudo-bulk. Color
242 palette is the same as in **a**. Middle color bar shows total accessibility within the cCRE in each pseudo-bulk.
243

244 **Intra-cCRE dynamics in hematopoietic aging**

245 Aging is a major risk factor for many highly prevalent diseases such as cancer, cardiovascular disease and
246 neurodegeneration²⁷. Extensive previous studies have shown that aging is accompanied by widespread
247 “epigenetic decline”^{28–30}. In particular, HSCs have been shown to be compromised in function during aging,
248 contributing to deficient pathogen- and vaccine-evoked immunity and heightened inflammatory responses^{31–}
249 ³³. Mutation of genes involved in epigenetic and chromatin remodeling has frequently been observed in
250 humans with clonal hematopoiesis of indeterminate potential (CHIP), an age-associated condition
251 characterized by the expansion of somatically mutated hematopoietic cell clones, a process associated with
252 an increased risk of hematopoietic malignancy, cardiovascular disease, stroke and all-cause mortality^{34,35}. As
253 such, prior studies have investigated alterations to DNA methylation, heterochromatin or chromatin
254 accessibility during HSC aging³⁶. Here, we hypothesized that aging HSCs relocalize regulatory proteins to
255 restructure sub-cCREs and alter the expression of aging genes. Thus, we applied PRINT to discover
256 alterations of TF and nucleosome binding in young or aged hematopoietic cells.
257

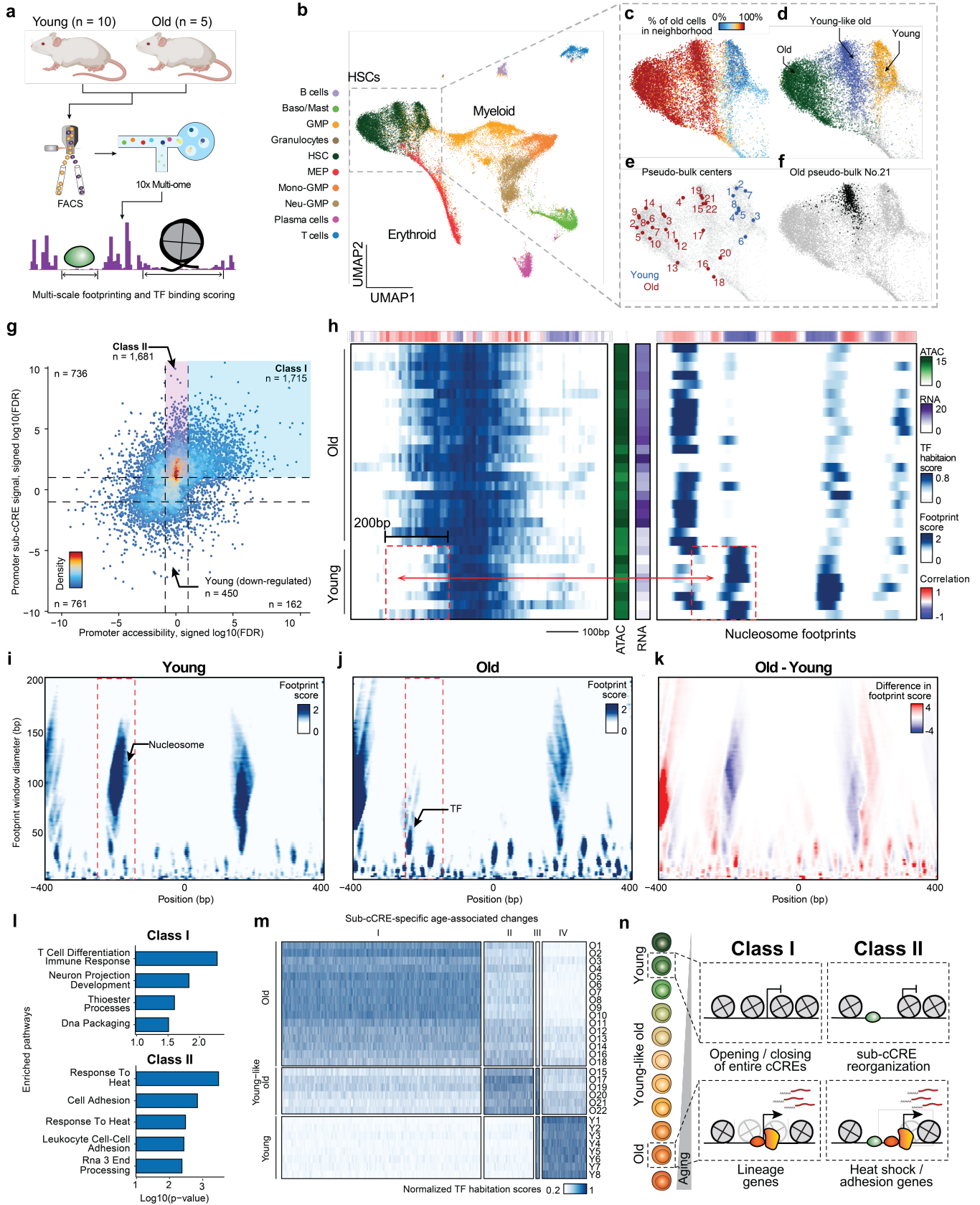
258 We isolated hematopoietic progenitor cells (Lineage⁻) and HSCs (Lineage⁻ Sca-1⁺ c-Kit⁺ Cd48⁻ CD150⁺) from
259 the bone marrow of young (11 weeks old, n = 10) or aged (24 months old, n = 5) mice by FACS. We then
260 obtained joint ATAC-RNA profiling using the 10x platform (**Fig. 4a, Extended Data Fig. S6a-c, Methods**).
261 Consistent with previous studies, we observed an expansion of the HSC compartment during aging³³
262 (**Extended Data Fig. S6d**). After QC filtering, we obtained 48,225 cells covering 14,640 HSCs and 33,585
263 hematopoietic progenitor cells in the mouse bone marrow (**Fig. 4b, Extended Data Fig. S7a**). From the
264 Lineage⁻ single-cell data we confirmed an age-associated increase in HSC frequency (**Extended Data Fig.**
265 **S7b**). Further, HSCs clustered into 3 clusters reflecting points along a continuum of age-related cell states
266 (**Fig. 4c, d**). Validating these clusters, we used gene expression to find age-specific marker genes (*Nupr1*,
267 *Clu*, *Selp*) (**Extended Data Fig. S7c-e**), consistent with findings of previous work^{36,37}. Of the two aging clusters,
268 we found that cluster 6 was more similar to young HSCs (R = 0.95 vs R = 0.91). We therefore refer to the
269 three clusters as young, young-like old, and old HSC states (**Fig. 4d**). Focused on age-related alterations, we
270 used the ATAC profile of single cells to define pseudo-bulks³⁸ revealing representative cell states (**Fig. 4e, f,**
271 **Extended Data Fig. S7f**).
272

273 We first applied PRINT to examine intra-cCRE reorganization in promoters of genes expressed in an age-
274 variant manner. More specifically, we identified promoters of genes with differential expression (**Extended**
275 **Data Fig. S7e, Extended Data Table 4**), and segmented these promoters into sub-cCREs. Applying
276 differential testing among young and old pseudo-bulks, we detected 4,132 old-specific and 1,373 young-

277 specific sub-cCREs (**Fig. 4g**, two-sample t test, FDR < 0.1). We identified age-associated elements with robust
278 increase to the overall accessibility and sub-cCRE activity (“Class I”; n = 1,715). In contrast, we discovered
279 age-associated increases to sub-cCRE activity that are missed when assessing the overall accessibility of
280 cCREs (“Class II”; n=1,681). As one example of Class II promoters, in young cells the promoter of *Cdc25b* is
281 flanked by two nucleosomes. In aged cells, the -1 phased nucleosome is lost, exposing additional DNA for TF
282 binding, which is accompanied by increased gene expression (**Fig. 4h-k**). Strikingly, promoters showing age-
283 related changes only at the sub-cCRE level (i.e., Class II promoters) were enriched for genes in the heat shock
284 pathway, such as *Hsp90ab1*, *Dnaja2* (*Hsp40* member), *Bcl2l11*, and *Ubqln2*, suggesting dysregulated
285 proteostasis during aging (**Fig. 4i**, **Extended Data Fig. S8a**, **Extended Data Table 5**). This suggests that
286 intra-cCRE reorganization might be involved in age-related impairment of proteostasis as reported by previous
287 studies in model organisms³⁹⁻⁴¹ and in HSCs⁴². Additionally, we observed enrichment in pathways related to
288 cell adhesion, involving genes such as *Igf1*, *Grb2*, and *Thy1*, potentially reflecting responses to the altered
289 cell-cell interactions within the HSC niche.

290
291 We next expanded the above analysis to include distal and proximal cCREs to identify sub-cCRE age-
292 associated alterations genome-wide. This analysis revealed 4 clusters (n = 18,166) with 84% of age-
293 associated sub-cCREs gaining activity during aging (**Fig. 4m**), denoting a global widening of cCREs. Further
294 categorizing distal and proximal sub-cCREs as Class I or Class II, revealed Class II elements were more
295 proximal to promoters ($p = 1.39 \times 10^{-9}$). We observed both the significant gain in expression and downstream
296 displacement of nucleosomes occluding motifs associated with AP-1 (*Fosl2*, *Fos*, *Fosb*, *Jund* and *Junb*) and
297 Tcf4 TFs (**Extended Data Fig. S8b**), which has been reported by previous studies to be involved in HSC
298 aging^{36,43}. Similarly, we also observed down-regulation of TFs such as *Arnt* and *Atf7*^{36,43}. Analysis of Class II
299 elements revealed *Hif1a* (heat shock), *Smad3* (cell adhesion), and *Ybx1/3* (proteostasis) regulators. In
300 contrast, Class I elements reflected alterations to TFs such as *Hox*, *Rorc*, *Maf* and *Runx* factors.

301
302 Overall, we find that aging is accompanied by widespread widening of cCREs to expose new TF binding sites.
303 These sub-cCRE changes are particularly enriched at loci encoding genes involved in regulating proteostasis
304 and cell adhesion in HSCs, and constitute a different class of regulation (Class II) than the Class I regulation
305 that has commonly been seen in development and cell fate decisions, which is driven by opening and closing
306 of entire cCREs (**Fig. 4n**). These data thus point to a new mechanism underlying age-dependent alterations
307 in gene expression, and help to explain why certain HSC functions, including protein quality control, cell
308 adhesion and RNA processing, are particularly vulnerable to age-dependent decline.



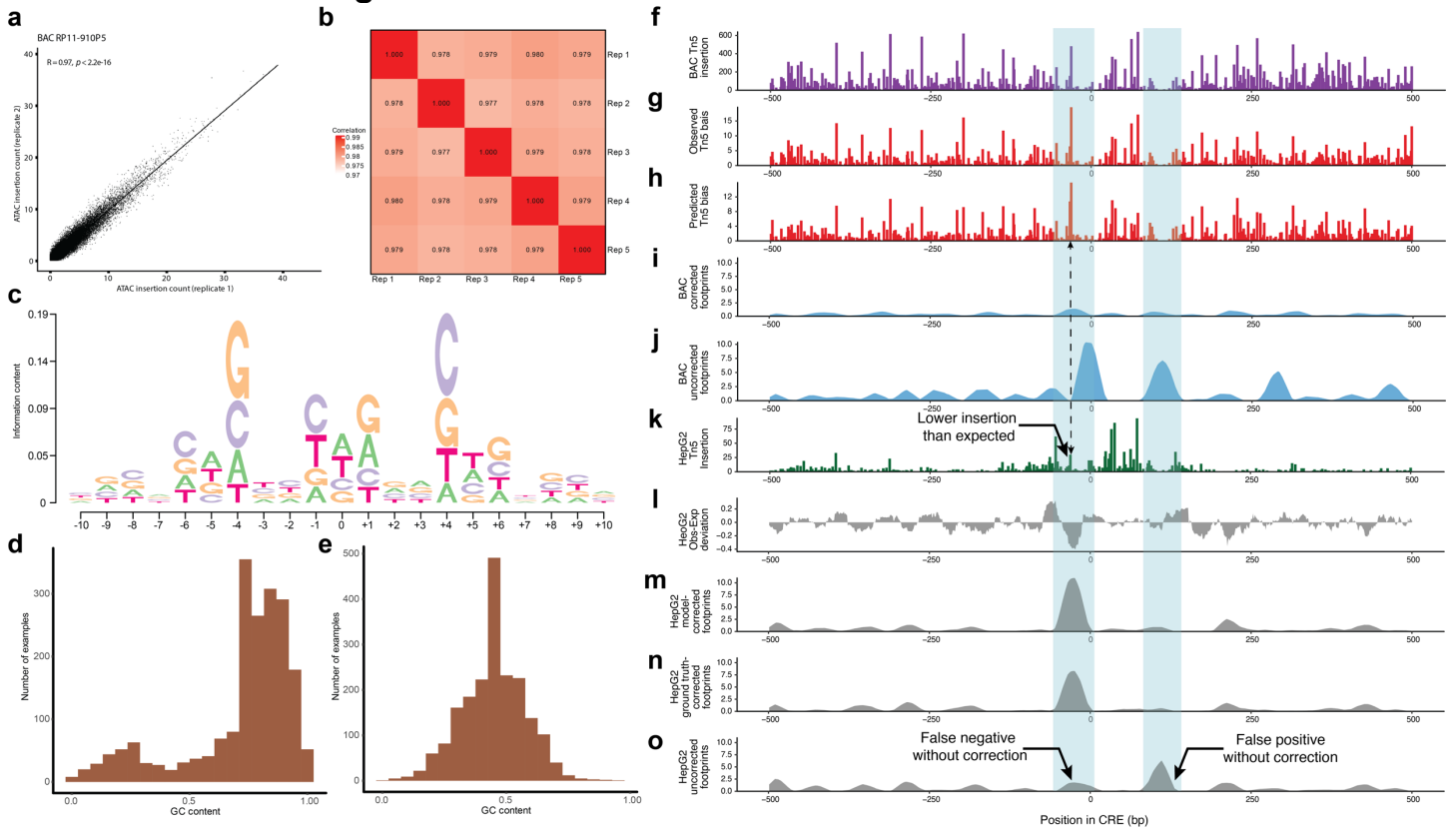
310 **Figure 4. Intra-cCRE dynamics in hematopoietic aging.** **a**, Schematic illustration of dataset generation and
311 analysis. **b-f**, UMAP of HSC and progenitor cells. **b**, Cell type annotation. **c**, Percentage of old cells in the 100-
312 cell nearest neighborhood. **d**, Young, young-like old, and old HSC clusters. **e**, Representative cell states
313 detected by SEACells. **f**, Example pseudo-bulk. Black dots represent member cells in old pseudo-bulk 21. **g**,
314 Scatter plot comparing differential cCRE testing and differential sub-cCRE testing results for promoters of
315 differentially expressed genes. Dashed lines represent the FDR = 0.1 threshold. **h**, Heatmaps of TF habitation
316 scores and nucleosome footprints (100 bp scale) within the promoter of *Cdc25b* at chr2:131186436-
317 131187435. Each row corresponds to a single pseudo-bulk, while each column represents a single base pair
318 position in the cCRE. Middle color bars show total accessibility within the cCRE and RNA of *Cdc25b* in each
319 pseudo-bulk, respectively. **i-k**, Heatmaps showing the multi-scale footprints within the *Cdc25b* promoter
320 across age groups. The horizontal axis represents single base pair positions, and the vertical axis represents
321 footprint window sizes. **i**, Young. **j**, Old. **k**, difference between young and old. **l**, Bar plot of pathway enrichment
322 Amy using either Class I or Class II as foreground and the other category as background. **m**, Heatmap showing
323 activity of age-related differential sub-cCREs across pseudo-bulks. Rows correspond to pseudo-bulks and
324 columns represent sub-cCREs. **n**, Schematic illustrating contrasting two classes of age-related cCRE changes
325 (modulation of overall cCRE accessibility and intra-cCRE reorganization).
326

327 Discussion

328 Our results highlight limitations of treating cCREs as digital, indivisible units. The observation that cCREs
329 shrink, expand, and merge as cells modulate the activity of sub-cCREs argues for a model wherein cCREs
330 dynamically recruit new effector proteins to alter their function over time. Prior studies argued that TF binding
331 is determined by wholesale opening or closing of cCREs instead of differential binding of TFs within the same
332 cCRE⁴⁴. In contrast, our study describes structural changes to cCREs, mediated by the repositioning of
333 nucleosomes and exposure of previously inaccessible DNA for TF binding. This difference likely arises from
334 the increased resolution with which we were able to examine cCRE structure, including the ability to footprint
335 objects of various sizes along a continuous trajectory of cell differentiation. In support of this model, studies
336 mapping TF binding by ChIP-seq report that TFs switch in development^{3,45,46}. Further, prior studies using high
337 resolution ChIP-seq find that nucleosomes are in active competition with transcriptional machinery⁷.
338

339 We find that cCREs may be divided into sub-cCREs according to their change across single-cell data to
340 significantly improve mapping of chromatin accessibility. In hematopoiesis, we find that individual cCREs
341 change in structure, exposing DNA for TF binding, and altering gene expression. In aging, we find that most
342 age-associated promoter changes alter the structure of cCREs, while fewer alter overall accessibility of the
343 element. Our approach for footprinting is generalizable, and may reveal sub-cCREs in previously published
344 bulk or single-cell ATAC-seq data, creating immediate opportunities across diverse studies of healthy and
345 disease biology. From this vast repertoire of regulatory diversity, we anticipate discovering functions for not
346 yet appreciated chromatin remodelers that slide or evict nucleosomes from regulatory DNA. Parsing cCREs
347 into sub-cCREs may also ascribe new functions to disease-causing genetic variation previously overlooked
348 by peak-based analyses. Taken together, our approach reveals the dynamics and functional importance of
349 cCRE structure, providing new insights into gene expression and highlighting functional DNA at single-cell
350 and single-base-pair resolution.

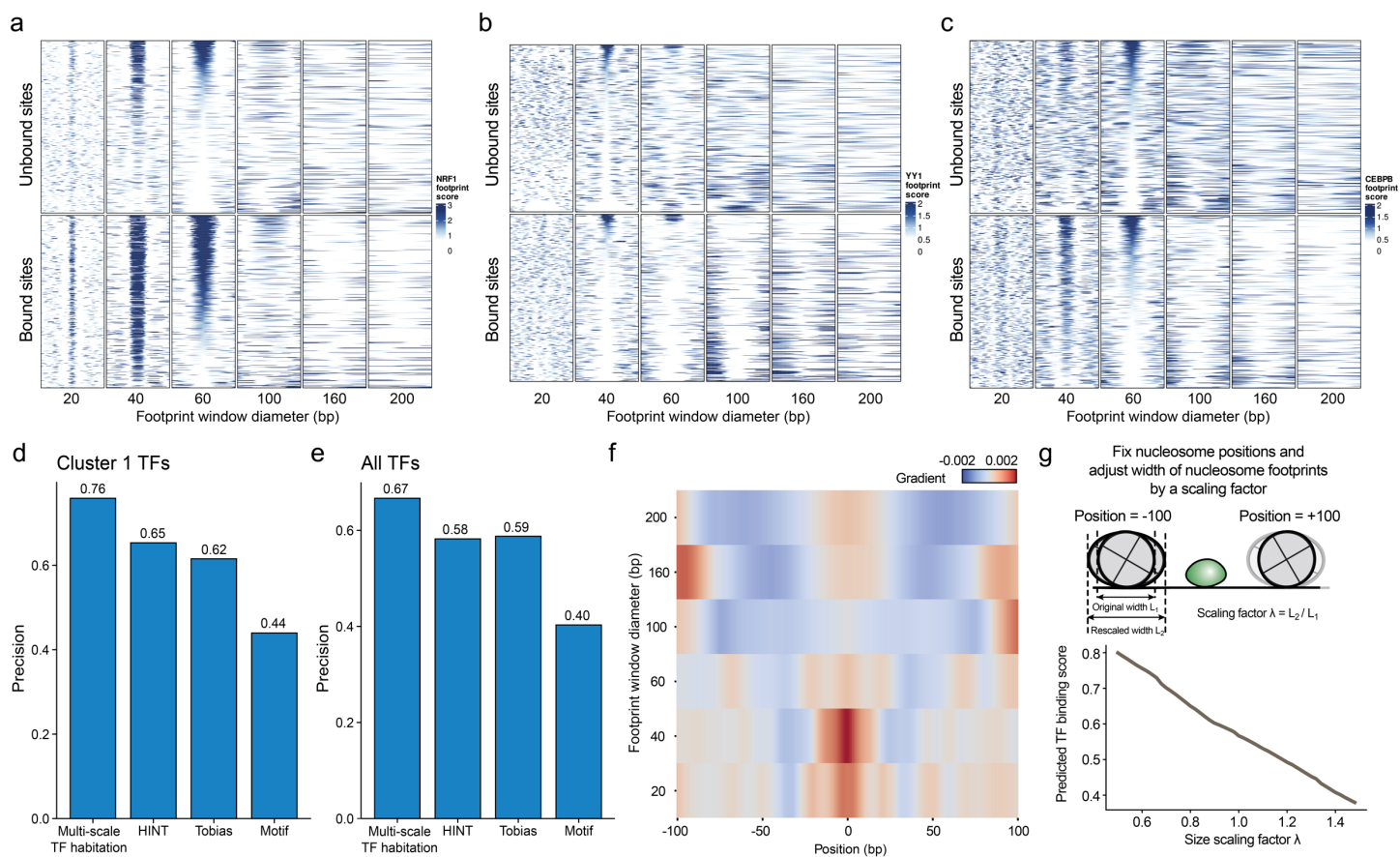
351 **Extended Data Figures:**



352 **Extended Data Figure S1. Tn5 bias modeling and footprinting**

353 **a**, Scatter plot comparing single nucleotide observed Tn5 insertion bias on BAC RP11-910P5 from replicate
 354 1 and 2. **b**, Heatmap showing Pearson correlation of observed Tn5 on all BACs among replicates. **c**, Motif plot
 355 of Tn5 sequence bias. **d**, Histogram of local GC-content in a +/- 10 bp window for top 2000 genomic positions
 356 where the neural network Tn5 bias model achieved the highest improvement in prediction error compared to
 357 the PWM bias model. **e**, Histogram of local GC-content in a +/- 10 bp window for bottom 2000 genomic
 358 positions where the neural network Tn5 bias model achieved the least improvement in prediction error
 359 compared to the PWM bias model. **f-o**, Testing our footprinting framework in an example cCRE region. **f-j**,
 360 Results for BAC naked DNA. **f**, Observed raw Tn5 insertion counts. **g**, Observed Tn5 bias. **h**, Tn5 bias
 361 predicted by the convolutional neural network. **i**, Footprint scores with bias correction. **j**, Footprint scores
 362 without bias correction. **k-o**, Results for HepG2 chromatin ATAC-seq. **k**, Observed raw Tn5 insertion counts.
 363 **l**, Observed-expected deviation of center / (center + flank) insertion ratio. **m**, Footprint scores with model-
 364 based bias correction. **n**, Footprint scores with bias correction using ground truth bias in **g**. **o**, Footprint scores
 365 without bias correction.

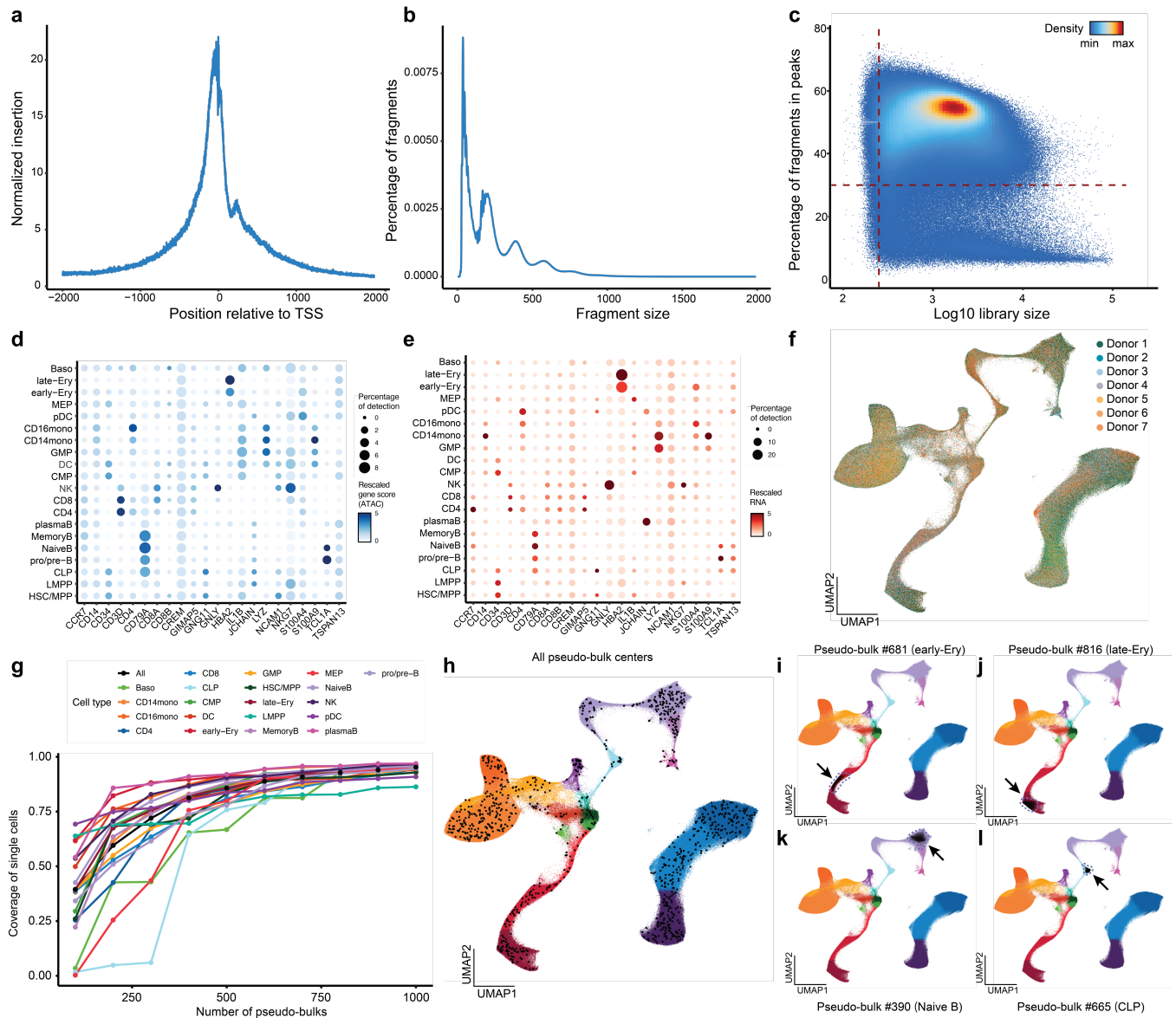
368 **a-c**, Multi-scale footprints for example TFs. **a**, NFIA. **b**, ATF2. **c**, TFE3. **d**, Heatmap showing clustering of
369 multi-scale aggregate footprints of different TFs. Each row is the multi-scale aggregate footprints of a specific
370 TF. Left color bar shows the cluster each TF is in. Right color bar shows the TF family each TF belongs to.



371
372
373
374
375
376
377
378
379
380

Extended Data Figure S3. Predicting TF binding using multi-scale footprints

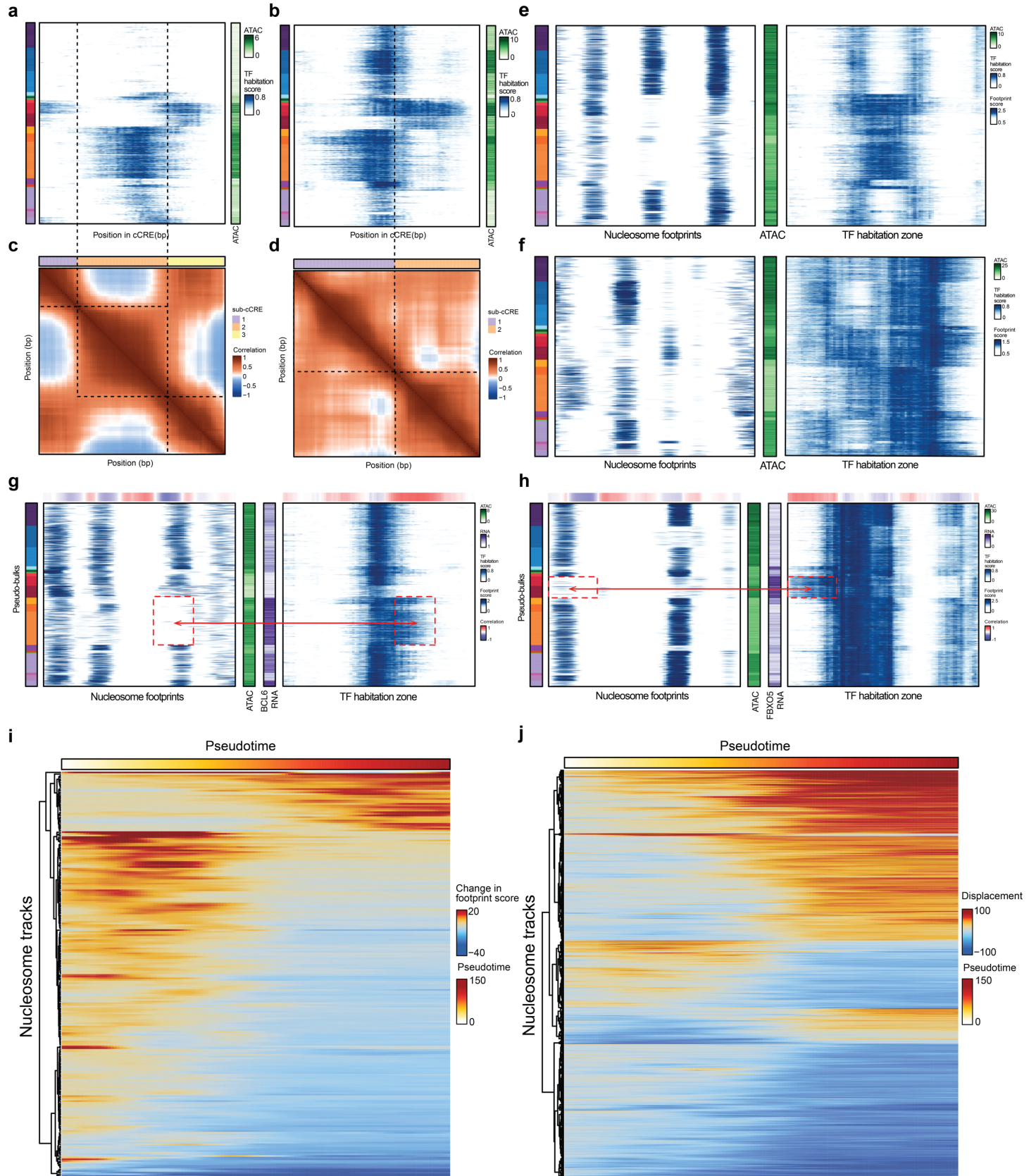
a-c, Multi-scale footprints around individual bound and unbound TF motif sites similar to Figure 2D. **a**, NRF1. **b**, YY1. **c**, CEBPB. **d**, Bar plot showing performance of different methods when benchmarked on cluster 1 TFs. **e**, Bar plot showing performance of different methods when benchmarked on TFs from all clusters. **f**, Heatmap showing gradients of predicted TF binding score with respect to input multi-scale footprints. Rows correspond to different footprint scales and columns represent single base pair positions within a +/- 100 bp range from the center. **g**, Effect of changing nucleosome footprint width on predicted TF binding scores. The two nucleosomes are fixed at +/- 100 bp positions, respectively and their widths are scaled by a scaling factor λ .



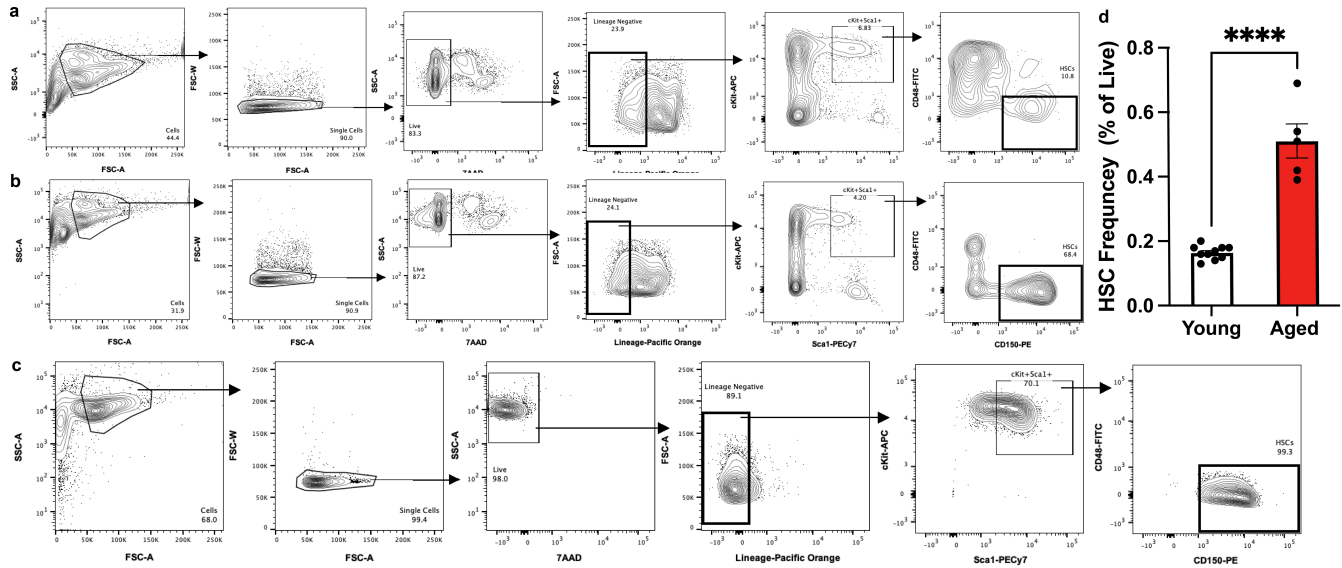
381
382
383
384
385
386
387
388
389

Extended Data Figure S4. Quality control and pseudo-bulking of the human bone marrow dataset.

a, Tn5 insertion enrichment around TSSs. **b**, Fragment size distribution. **c**, Scatter plot showing library size and fraction of reads in peaks (FRIP) of single cells. **d**, Dot plot showing gene scores (ATAC signal within a region around promoter) of marker genes across cell types. **e**, Dot plot showing of RNA levels of marker genes across cell types. **f**, UMAP showing donor origin of single cells. **g**, Line plot showing single cell coverage for each cell type as a function of the number of pseudo-bulks. **h**, UMAP showing the positions of pseudo-bulk centers for all 1000 pseudo-bulks we generated. **i-l**, Example pseudo-bulks. Black dots represent member cells within the pseudo-bulk.

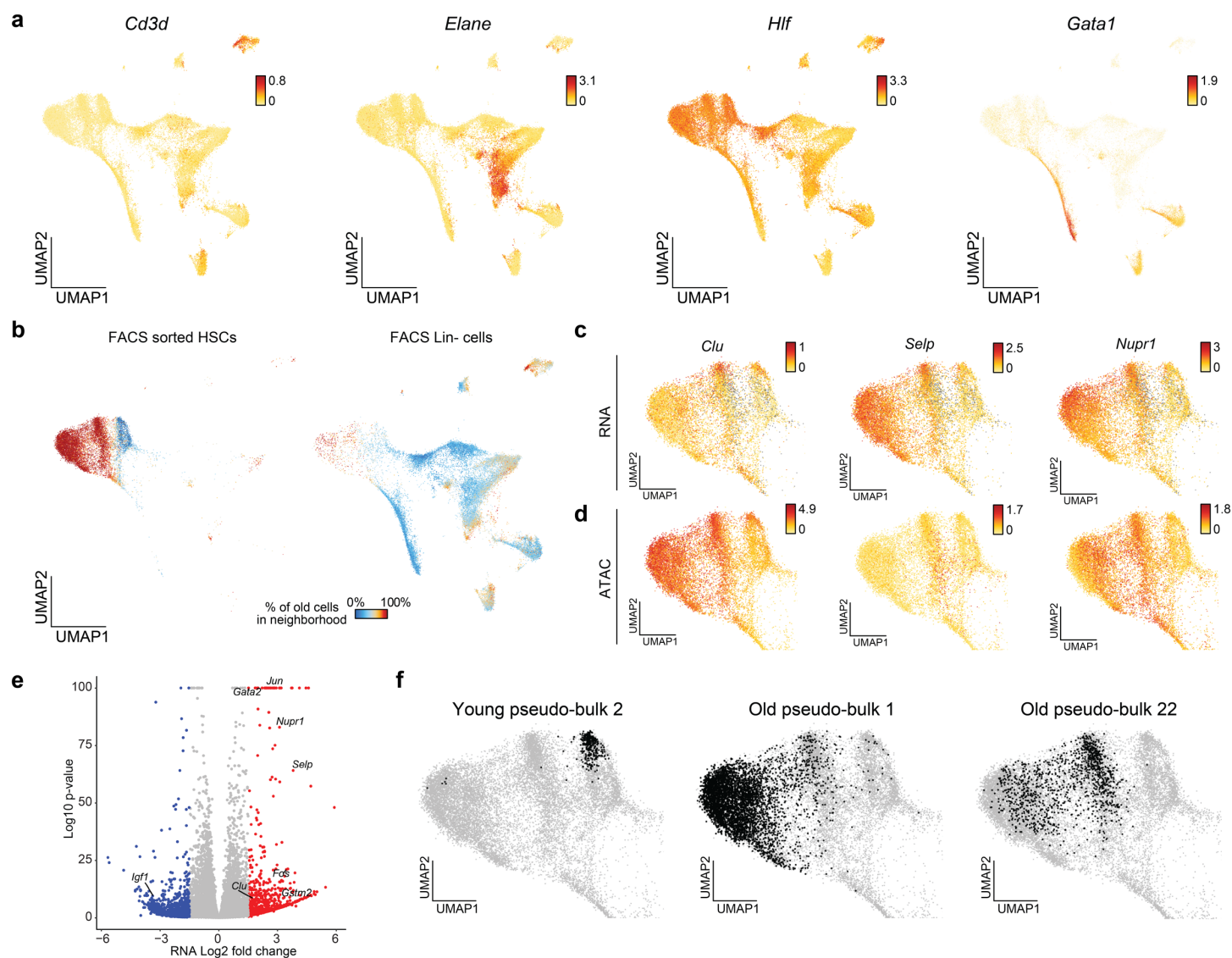


391 **Extended Data Figure S5. sub-cCRE and nucleosome dynamics.**
392 **a-d**, Defining sub-cCREs. **a-b**, Heatmap of predicted TF habitation scores in the cCRE at chr17:40329944-
393 40330943 and chr20:56411666-56412665, respectively. Each row corresponds to a single pseudo-bulk, while
394 each column represents a single base pair position in the cCRE. Left color bar shows the cell type labels of
395 pseudo-bulks, and the colormap is the same as in Figure 3a. Right color bar shows total accessibility of the
396 cCRE across the pseudo-bulks. **c-d**, Heatmap showing pairwise correlation of TF habitation scores among
397 individual base pair positions within the same cCRE as in **a-b**. Top color bar shows results of sub-cCRE
398 segmentation. **e-f**, Examples of cCRE merging and unmerging. **e**, Example cCRE at chr11:64810186-
399 64811185. **f**, Example cCRE at chr1:84690049-84691048. Left heatmaps show the nucleosome footprints
400 (footprint scores calculated at 100 bp scale), while right heatmaps show the TF habitation scores in the same
401 region. Each row corresponds to a single pseudo-bulk, while each column represents a single base pair
402 position in the cCRE. Middle color bar shows total accessibility of the cCRE across the pseudo-bulks. **g-h**,
403 Example cCREs with strong sub-cCRE-gene correlation and weak cCRE-gene correlation. **g**, Example cCRE
404 at region chr3:187739617-187740616. **h**, Example cCRE at region chr6:152982810-152983809. Middle color
405 bars show total accessibility of the cCRE and the RNA of the target gene, respectively. **i-j**, Tracking
406 nucleosome dynamics across erythroid differentiation. **i**, Heatmap showing nucleosome binding/eviction
407 dynamics across pseudotime during erythroid differentiation. Rows are individual nucleosome tracks. Color
408 represents the change in footprint score compared to $t = 0$. **j**, Heatmap showing nucleosome sliding dynamics
409 across pseudotime during erythroid differentiation. Rows are individual nucleosome tracks. Color represents
410 the displacement (in bp) compared to starting position at $t = 0$. Negative values represent sliding towards
411 upstream regions and vice versa.



412
413
414
415
416
417
418
419

Extended Data Figure S6. FACSSorted Hematopoietic Cells from Aging Male Mice. **a-b**, Flow cytometry gating strategy for isolation of hematopoietic progenitor cells (Lineage Negative, gate bolded; Live Lin⁻) and hematopoietic stem cells (HSCs, gate bolded; Live Lin⁻ Sca1⁺ cKit⁺ CD48⁻ CD150⁺) from the bone marrow (BM) of young (**a**, n = 10) and aged (**b**, n = 5) male C57BL6/J mice. Representative FACS plots shown from one young and one aged mouse. For individual FACS plots from each mouse, see Extended Data File 1. **c**, Purity of resorted HSCs was greater than 99%. **d**, Frequency of FACSSorted HSCs in young and aged mice (two-tailed t-test; $t_{13} = 9.283$, $p < 0.0001$).



420

421

Extended Data Figure S7. Characterizing age-related changes in mouse HSCs

422

a, UMAP showing gene scores of cell type marker genes *Cd3d*, *Elane*, *Hlf*, and *Gata1*, respectively. **b**,

423

Percentage of old cells in the 100-cell nearest neighborhood for FACS sorted HSCs (left) or Lineage- cells

424

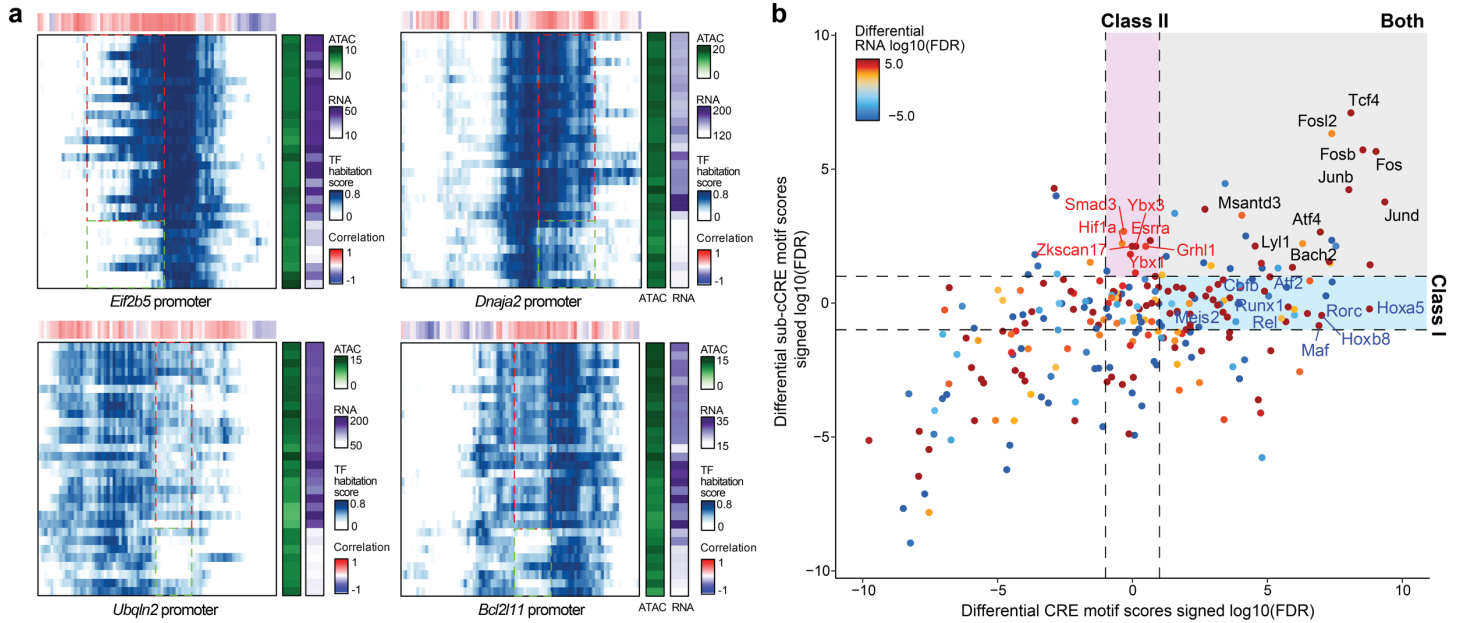
(right). **c-d**, UMAP showing (c) RNA and (d) ATAC levels of aging marker genes (*Clu*, *Selp*, *Nupr1*) in HSCs.

425

e, Volcano plot of differential RNA testing (old-versus-young). **f**, UMAP showing example pseudo-bulks. Black

426

dots represent member cells within each pseudo-bulk.



427

428

Extended Data Figure S8. Hallmarks of aging-associated sub-cCRE alterations

429

a, Heatmaps of predicted TF habitation scores in the promoters of *Eif2b5*, *Dnaja2*, *Ubqln2*, and *Bcl2l11*, respectively. Each row corresponds to a single pseudo-bulk, while each column represents a single base-pair position in the cCRE. Right color bar shows total accessibility of the cCRE and RNA level of the corresponding gene across the pseudo-bulks. Top color bar shows correlation of TF habitation score with RNA level at each base-pair position. **b**, Scatter plot comparing differential TF motif score testing results using cCREs and sub-cCREs as input features, respectively.

430

431

432

433

434

435 **Methods:**

436 **EXPERIMENTAL METHODS**

437 Cell culture

438 HepG2 cells were cultured in Dulbecco's Modified Eagle Medium (DMEM, 11965092, ThermoFisher) with the
439 addition of 10% FBS and 1% of penicillin-streptomycin. Cells were incubated at 37°C in 5% CO₂ and
440 maintained at the exponential phase. Cells were digested with TrypLE express (12604013, ThermoFisher) for
441 preparing single-cell suspension.

442 V6.5 mouse embryonic stem cells were cultured in Glasgow Minimum Essential Medium (GMEM)
443 supplemented with 10% FBS, 2 mM L-glutamine, 1% Pen/Strep, 1 mM sodium pyruvate, 2000 units/mL
444 (10ng/mL) Leukemia Inhibitory Factor (LIF, Millipore), 1x Minimum Essential Medium Non-Essential Amino
445 Acids (MEM NEAA, Invitrogen) and 50 uM β-Mercaptoethanol. Tissue culture plates were coated with 0.2%
446 gelatin and 0.25 mg/mL laminin for 2 hours at 37C before seeding. Media was changed every other day, and
447 cells were split every 3-4 days.

448

449 BMMC sample processing

450 Frozen human Bone Marrow Mononuclear Cells (BMMCs, Allcells) were thawed in a 37 °C water bath for 1
451 min and transferred to a 15 mL centrifuge tube. 10 mL of pre-warmed DMEM with 10% FBS was added to
452 cells drop-wisely. The cells were spun at 400g for 3 min at room temperature. After removing supernatant, the
453 cells were washed twice in 0.5 mL PBS with 0.04% BSA. To deplete neutrophils, the cells were resuspended
454 in 100 µl chilled DPBS with 0.2% BSA and 10 µl of human TrueStain FcX (BioLegend, 422302) and incubated
455 on ice for 10 min to reduce non-specific labeling. The cells were then incubated on ice for another 30 min after
456 adding 0.5 µl of biotin anti-human CD15 antibody (BioLegend, 301913). After immunostaining, 25 µl of MyOne
457 T1 beads were added to the sample to capture the neutrophils for 5 min at room temperature. We then added
458 900 µl of DPBS with 0.2% BSA to dilute the sample. The sample was placed on a magnet for 3 min and 1 ml
459 of the sample was transferred to a new tube while the sample was on the magnet. The cells were ready for
460 fixation and SHARE-seq experiment.

461

462 Fixation

463 Cells were centrifuged at 300g for 5 minutes and resuspended to 1 million cells/ml in PBSI. Cells were fixed
464 by adding formaldehyde (28906, ThermoFisher) to a final concentration of 1% and incubated at room
465 temperature for 5 minutes. The fixation was stopped by adding 56.1 µl of 2.5M glycine, 50 µl of 1M Tris-HCl
466 pH 8.0, and 13.3 µl of 7.5% BSA on ice. The sample was incubated at room temperature for 5 minutes and
467 then centrifuged at 500g for 5 minutes to remove supernatant. All centrifugations were performed on a swing
468 bucket centrifuge. The cell pellet was washed twice with 1ml of PBSI, and centrifuged at 500g for 5 minutes
469 between washings. The cells were resuspended in PBS with 0.1U/µl Enzymatics RNase Inhibitor and
470 aliquoted for transposition.

471

472 SHARE-seq

473 Following fixation SHARE-seq was performed as previously described²⁶, with the following modifications. To
474 improve transposition, transposition was performed using pre-assembled Tn5 (seqWell, Tagify(TM) SHARE-
475 seq Reagent). To improve RNA capture, we added polyA to transcripts prior to reverse transcription. To do
476 this, transposed cells (60 µl) were mixed 240 µl of poly(A) mix (final concentration of 1× Maxima RT buffer,
477 0.25 U/µl Enzymatics RNase Inhibitor, 0.25 U/µl SUPERase RI, 0.018 U/µl E.coli poly-A enzyme (M0276L), 1
478 mM rATP). The sample was aliquoted to 50 µl per PCR tube and incubated at 37 °C for 15 minutes.

479

480 Quantification and sequencing

481 Both scATAC-seq and scRNA-seq libraries were quantified with the KAPA Library Quantification Kit and
482 pooled for sequencing. Single cell libraries were sequenced on the Nova-seq platform (Illumina) using a 200-
483 cycle kit (Read 1: 50 cycles, Index 1: 99 cycles, Index 2: 8 cycles, Read 2: 50 cycles). Bulk libraries were

484 sequenced on the Nova-seq platform (Illumina) using a 100-cycle kit (Read 1: 50 cycles, Index 1: 8 cycles,
485 Index 2: 8 cycles, Read 2: 50 cycles).

486

487 SHARE-seq data pre-processing

488 SHARE-seq data were processed using the SHARE-seqV2 alignment pipeline
489 (<https://github.com/masai1116/SHARE-seq-alignmentV2/>) and aligned to hg38. Open chromatin region peaks
490 were called on individual samples using MACS2 peak caller⁴⁷ with the following parameters: --nomodel –
491 nolambda –keep-dup -call-summits. Peaks from all samples were merged and peaks overlapping with
492 ENCODE blacklisted regions (<https://sites.google.com/site/anshulkundaje/projects/blacklists>) were filtered out.
493 Peak summits were extended by 150 bp on each side and defined as accessible regions (for footprinting
494 analyses, these peaks were later resized to 1000 bp in width). Peaks were annotated to genes using Homer⁴⁸.
495 The fragment counts in peaks and TF scores were calculated using chromVAR⁴⁹. Cell barcodes with less than
496 30% reads in peaks (FRiP) or 250 unique fragments were removed. The aligned reads were then intersected
497 with peak window regions, producing a matrix of chromatin accessibility counts in peaks (rows) by cells
498 (columns). To examine the cell identity, the cisTopic (50 topics)⁵⁰ were used for dimension reduction, followed
499 by Louvain clustering. The progenitor populations were sub-clustered to obtain finer cell identity. The data
500 were projected into 2D space by UMAP⁵¹. Seurat V3⁵² was used to scale the DGE matrix by total UMI counts,
501 multiplied by the mean number of transcripts, and values were log transformed.

502

503 Generation of BAC naked DNA data

504 We selected 25 chromatin regions based on overlap with a manually selected set of key transcription factors
505 and differentiation related genes. The BAC clones (BACPAC Resources) were cultured in LB for 14 hours.
506 The BAC DNA was extracted using ZR BAC DNA Miniprep Kit (Zymo, D4048) following manufacturer's
507 instructions. The purified DNA was quantified using Qubit (ThermoFisher). The BAC DNA were tagmented
508 similar to the SHARE-seq ATAC-seq experiment. Briefly, 50 ng of BAC DNA from multiple clones were pooled
509 for tagmentation following the SHARE-seq transposition condition. The tagmented DNA was purified using a
510 Qiagen Minelute PCR clean up kit and then amplified for 7 cycles by PCR. To minimize batch effect, we
511 generated 5 replicates and pooled all the materials for sequencing. The library was sequenced on a Nova
512 platform (Illumina) using a 100-cycle kit (Read 1: 50 cycles, Index 1: 8 cycles, Index 2: 8 cycles, Read 2: 50
513 cycles). The sequencing data was processed the same way as SHARE-seq ATAC-seq data.

514

515 Aging Multi-ome experiment

516 Mouse experiments were approved and performed in compliance with Harvard University's Institutional Animal
517 Care and Use Committee. C57BL6 mice were obtained from either Jackson Laboratory or the National Institute
518 on Aging Aged Rodent Colony (Charles River Laboratory), and housed at a density of 2-5 mice per cage in
519 standard ventilated racks and provided food and water *ad libitum* in a pathogen-specific free facility accredited
520 by the Association and Accreditation of Laboratory Animal Committee (AALAC). Mouse cages contained
521 Anderson's Bed o Cob bedding (The Anderson, Inc), two nestlets (Ancare, 2x2" compressed cotton square),
522 and a red mouse hut (Bioserv). For HSC isolation and flow cytometry. Cells from the bone marrow of long
523 bones (2 femurs and 2 tibias per mouse) from young (n = 10; 11 weeks old) and aged (n = 5; 24 mo. old) male
524 C57BL/6 mice were flushed with a 21-gauge needle into staining media (HBSS/2% fetal bovine serum),
525 pelleted, and resuspended in ACK lysis buffer for 5 min on ice. Cells were then washed with staining media,
526 filtered through a 40mm cell strainer, pelleted, and incubated with the following cocktail of rat anti-mouse,
527 biotin conjugated lineage antibodies on ice for 30 min: CD3 clone C145-2c11 (Biolegend, 1000304; 1:100),
528 CD4 clone GK15 (Biolegend, 1000404; 1:400), CD5 clone 53-7.3 (eBioscience, 13-0051-85; 1:400), CD8
529 clone 53-6.7 (Biolegend, 100704; 1:400), CD19 clone 6D5 (Biolegend, 115504; 1:400), B220 clone RA3-6B2
530 (Biolegend, 103204; 1:200), GR1 (Ly6-G/Ly6-C) clone RB6-8C5 (eBioscience, 13-5931-82; 1:400),
531 Mac1/CD11b clone M1/70 (Biolegend, 101204; 1:800), and Terr119 clone TERR-119 (Biolegend, 116204;
532 1:100). Cells were then washed in staining media, with a small aliquot reserved for each sample to serve as
533 a non-depleted control, and lineage depleted using sheep anti-rat Dynabeads (Invitrogen, 1135) on a magnet.
534 Cells were washed, pelleted, and incubated with the following cocktail of anti-mouse antibodies on ice for 45
535 min. to identify hematopoietic stem cells (HSC): Pacific Orange Streptavidin (Invitrogen, S32365; 1:500),
536 PE/Cy7 Sca1(Ly-6a/E) clone D7 (eBioscience, 25-5981-82; 1:200), APC cKit clone 2B8 (BD Pharmingen,

537 553356; 1:200), FITC CD48 clone HM48-1 (Biolegend, 103403; 1:200), and PE CD150 clone Tc15-12F12.2
538 (Biolegend, 115904; 1:200). Following incubation, cells were washed and resuspended in staining media, and
539 7-AAD (BD Pharmingen, 559925; 1:50) added immediately prior to flow cytometry. Cell sorting of HSCs (Live
540 Lin⁻ Sca1⁺ cKit⁺ CD48⁻ CD150⁺) was performed on a BD FACS Aria II, and data analysis performed using BD
541 FACS Diva and FlowJo software. Data processing was performed using CellRanger.

542
543 After sorting, nuclei were isolated following 10x Genomics' demonstrated protocol "Low Cell Input Nuclei
544 Isolation", which is described in the CG000365 User Guide. Nuclei were then processed using the Chromium
545 Single Cell Multiome ATAC + Gene Expression kit (10x Genomics), following manufacturer's instructions, to
546 obtain between 2,000 and 10,000 cells per sample. Libraries were sequenced on an Illumina Nextseq system
547 using the following sequencing formats: Read 1 - 28, i7 index - 10, i5 index - 10, Read 2 - 44 (scRNA-seq),
548 Read 1 - 30, i7 index - 8, i5 index - 24, Read 2 - 30 (scATAC-seq). Data processing was performed using the
549 CellRanger software from 10x Genomics.

550
551 Tn5 sequence bias modeling

552
553 *Getting Tn5 insertion counts*

554 The ends of the fragments files are shifted by +/-4 (in 1-based indexing system) to obtain the center of the 9
555 bp staggered end created by Tn5 transposition. The number of insertions at each single base-pair position
556 within each cCRE from each sample is then quantified and stored in a sample-by-cCRE-by-position 3D tensor
557 for fast data retrieval.

558
559 *Data preprocessing*

560 The model takes local DNA sequence context as input and predicts single-base pair resolution Tn5 bias. To
561 this end, the +/- 50 bp DNA sequence surrounding each position of interest is encoded by one-hot encoding
562 into a 101-by-4 matrix and used as model input. For the prediction target, we use local relative Tn5 bias as
563 the target value. More specifically, the raw Tn5 insertion count at each position is divided by the average Tn5
564 insertion count within a +/-50 bp window. Positions with low local coverage (< 20 insertions per bp) were
565 removed to guarantee quality of training data. To facilitate model training, the resulting observed Tn5 bias
566 values are log₁₀-transformed and rescaled. For dataset partition, we randomly split all the BACs into 80%,
567 10%, and 10% for training, validation, and test sets. In other words, all data originating from the same BAC
568 belong to the same partition. This is to prevent overlapping local sequence contexts ending up in both training
569 and testing datasets, which might lead to overestimation of performance. To guarantee equal coverage of
570 examples with different bias levels, we binned all training examples into 5 bins based on their Tn5 bias values,
571 and up-sampled each bin so that all bins end up with the same number of examples. Additionally, given the
572 symmetric nature of Tn5 insertion, we generated reverse complement sequences of the training examples as
573 data augmentation. The original and reverse complement data were combined, shuffled, and then used for
574 model training.

575
576 *Model architecture*

577 The convolutional network consists of three convolution & max-pooling layers and two fully connected layers.
578 Each convolution and max-pooling layer performs convolution, ReLU nonlinear activation⁵³, and max pooling
579 sequentially. We used 32 filters of width 5 for each layer, along with "same" padding mode and stride size of
580 1. The two following fully connected layers have output dimensions of 32 and 1, respectively. ReLU activation
581 is used by the first fully connected layer and linear activation is used by the second layer (i.e., the final output
582 layer).

583
584 *Model training and evaluation*

585 The model was trained on the training set, and hyperparameters were optimized based on performance on
586 the validation set. Final performance of the frozen model was evaluated on the test set. The model was
587 implemented using Keras⁵⁴, trained with mean square error as loss function and optimized using the Adam
588 optimizer⁵⁵ with default parameters. Training was performed with a batch size of 64 and early stopping based
589 on model loss on the validation set.

590

591 *Benchmarking with other Tn5 bias models*

592 Methods including k-mer models (k = 3, 5, 7) and PWM methods (single nucleotide and dinucleotide) were
593 included in benchmarking. For k-mer methods, the foreground and background frequencies for all possible k-
594 mer sequences were quantified. The foreground frequency / background frequency ratio was used as the
595 estimated Tn5 bias for the corresponding k-mer. For single nucleotide PWM, we calculated foreground and
596 background base frequencies within a +/- 10 bp window (total length = 21) and computed the PWM of Tn5
597 insertion. Dinucleotide PWM scores were calculated using TOBIAS²¹ with default settings.

598

599 *Genome-wide Tn5 bias reference tracks*

600 Sequences of reference genomes for *Homo sapiens* (hg38), *Mus musculus* (mm10), *Drosophila melanogaster*
601 (dm6), *Saccharomyces cerevisiae* (sacCer3), *Caenorhabditis elegans* (ce11), *Danio rerio* (danRer11), and
602 *Pan troglodytes* (panTro6) are downloaded from the UCSC genome browser website⁵⁶
603 <https://hgdownload.soe.ucsc.edu/goldenPath/>. The aforementioned Tn5 bias neural network model was
604 applied to each position in the reference genomes to generate genome-wide Tn5 bias tracks.

605

606 *Computing footprint scores*

607 To detect DNA-protein interactions at different scales within cCREs, we implemented a framework for
608 computing footprint scores for each base pair position in the cCRE. In short, for each single bp position, we
609 define a center footprint window and flanking windows on both sides (Figure 1e). Then we calculate the
610 observed ratio of center / (center + flanking) Tn5 insertion counts. The foreground observed ratio is compared
611 to a background distribution to calculate statistical significance, which is then converted to a footprint score.

612

613 *Estimation of background dispersion*

614 Given a specific combination of center bias, flanking bias, and local coverage, we expect a certain distribution
615 of center / (center + flanking) insertion ratio when no protein is bound. This is defined as the background
616 distribution. Such background distribution can be estimated using BAC naked DNA Tn5 insertion data. To this
617 end, we first randomly sampled 100,000 positions from the BAC dataset, and retrieved their local coverage
618 (defined as the total insertion number in center and flanking areas), center bias, as well as flanking bias. Then
619 for each sampled position A, we identified 500 nearest neighbor positions NN₁-NN₅₀₀ in the 3-dimensional
620 space of (center bias, flanking bias, local coverage). To make sure each dimension is weighed equally, the
621 values of each dimension were first normalized to zero mean and unit variance. The 500 nearest neighbor
622 observations can be considered as background observations with nearly identical bias and coverage, and the
623 center / (center + flanking) ratio of NN₁-NN₅₀₀ forms the background distribution of position A. Therefore, for
624 each of the 100,000 sampled positions, we can calculate the mean and standard deviation of its background
625 ratio distribution. This allows us to train a background dispersion model that takes the tuple (center bias,
626 flanking bias, local coverage) as input and predicts the mean and standard deviation of the background
627 distribution very efficiently. To make sure the model is exposed to training examples with a wide range of local
628 coverage, we down-sampled the BAC dataset to 50%, 20%, 10%, 5%, and 1% of the original sequencing
629 depth. Finally, we trained a neural network with a single hidden layer (32 nodes, ReLU activation⁵³) and linear
630 output layer activation. The dataset was randomly split into 80% training, 10% validation, and 10% test. The
631 model was implemented using Keras⁵⁴, and trained on the training dataset with mean squared error loss using
632 the Adam optimizer^{54,55}. Early stopping was determined using loss on the validation set, and performance of
633 the final model was evaluated on the test set. Additionally, we trained separate models for each footprint radius
634 due to the drastic differences in total center or flank bias when footprint radius varies. For details, see
635 Supplementary Notes.

636

637 *Calculating footprint scores*

638 For each position in the cCRE, we define a center footprint window and flanking windows on both sides. We
639 first calculate the foreground observed center / (center + flanking) ratio of Tn5 insertion counts. Then we apply
640 the pre-trained background dispersion model to calculate the mean and standard deviation of its background
641 distribution. We next use a lower-tailed z-test to calculate the p-value for footprinting. If the observed ratio is
642 significantly lower than the background distribution, then this position is likely to be bound by a protein. More

643 specifically, to avoid calling footprints at positions where only one flanking side shows higher Tn5 insertion
644 than the center window but not the other, we perform center-versus-left and center-versus-right tests
645 separately and keep the larger p -value (See Supplementary Notes for details). The $-\log_{10}(p\text{-values})$ are
646 smoothed by running-max and running-mean smoothing and then used as the final footprint scores.

647

648 *Aggregate footprinting*

649 To calculate aggregate footprints, Tn5 insertions surrounding TF or nucleosome binding sites across the
650 genome are first aggregated and then used to calculate footprint scores. For TFs, we selected sites with a
651 matched TF motif using `motifmatchr`⁵⁷ ($p\text{-cutoff} = 1e-5$) and overlapping with a ChIP-seq peak of the
652 corresponding TF. For motif matches on the minus strand, the Tn5 insertion profile surrounding the motif is
653 inverted so the insertions for different sites are aligned in the same direction. For nucleosomes, we
654 downloaded a previously published list of chemically mapped nucleosome positions in mouse embryonic stem
655 cells (mESCs)²⁴ and used these positions for aggregating nucleosome footprinting with mESC single cell
656 ATAC-seq data.

657

658 Predicting TF binding

659 *Input data*

660 To predict the landscape of TF binding, we trained a binary classifier that predicts whether any TF motif site
661 is bound by the corresponding TF. Motif sites are identified by the `matchMotifs` function in the `motifmatchr`
662 package⁵⁷. All sites with a matching p -value below $5e-5$ are kept. For any TF motif site, we use multi-scale
663 (20 bp, 40 bp, 60 bp, 100 bp, 160 bp, 200 bp in diameter) footprints within a ± 100 bp local area centered
664 around the motif, as well as a motif match score as input to the model. The motif match score returned by the
665 `matchMotifs` function is quantile-transformed to uniform distribution. As a result, by combining the 201-
666 dimensional footprint vectors from 6 different scales with a single motif match score, we end up with a 1207-
667 dimensional vector as the final model input. The first 1206 dimensions of footprint scores are standardized
668 individually to zero mean and unit variance. For the prediction target, we assign a label of 1 to all sites
669 overlapping with a ChIP peak of the same TF, and a label of 0 to sites not overlapping with ChIP.

670

671 In total, we trained two separate models. (1) The first model was trained using only data of cluster 1 TFs. This
672 model was trained to predict binding of TFs that leave strong footprints. Some TFs were found to have a very
673 low percentage of motif sites overlapping with ChIP ($< 25\%$), potentially due to low quality of the motif or the
674 ChIP dataset. Such TFs are removed from model training and testing. We also added an equal number of
675 random negative examples as well as reverse-complement examples for data augmentation. (2) The second
676 model, referred to as the TF habitation model, was trained on TFs from all clusters (to include more training
677 data, we were keeping TFs with $> 20\%$ of motifs overlapping with ChIP data). The model was trained to infer
678 binding for both strong-footprinting and weak-footprinting TFs. Similarly, we added reverse complement
679 examples for data augmentation.

680

681 For data partition, we used HepG2 SHARE-seq data and GM12878 SHARE-seq data (GM12878 data is
682 previously published in the original SHARE-seq paper²⁶) for model training and validation, and test the model
683 on K562 Biorad single cell ATAC data, as well as three cell types (naive B cells, CD14 monocytes, and late-
684 erythroid cells) in the human BMDC SHARE-seq dataset. In particular, for the cluster I-specific model, TFs in
685 cluster I were used as training data and other TFs were used as validation. For the TF habitation model,
686 HepG2 data was used as training data and GM12878 data was used as validation. After fixing model
687 hyperparameters, HepG2 and GM12878 data were combined to train a final TF habitation model for
688 performance testing.

689

690 *Model architecture and training*

691 The TF binding prediction model is a neural network model with two hidden layers (32 + 16 nodes for cluster-
692 I specific model, and 128 + 16 nodes for the TF habitation model). ReLU activation⁵³ is used by both hidden
693 layers and sigmoid activation is used by the final output layer. The model was implemented using Keras⁵⁴.
694 The model was trained on the training dataset with a batch size of 128 using the Adam optimizer⁵⁸. Binary
695 cross entropy is used as the loss function. Early stopping was used based on model loss on the validation set.

696

697 *ChIP validation and benchmarking with previous methods*

698 To evaluate model performance, we used ChIP-seq as ground truth and validated predicted binding events.
699 HepG2 and GM12878 data for model training were downloaded from ENCODE⁶. ChIP-seq for BMMC cell
700 types were downloaded from cistromeDB⁵⁹. For benchmarking with previous methods, to make sure we only
701 include high quality TF binding sites, we downloaded K562 ChIP-based TF binding data from unibind²⁵
702 (<https://unibind.uio.no/search>). For ENCODE datasets, we removed those with the two most severe levels of
703 audit categories. For cistromeDB datasets, we applied QC filters as specified on the cistromeDB website
704 <http://cistrome.org/db/#/about>. More specifically, we included the below filters: FRiP ≥ 0.01 , FastQC ≥ 0.25 ,
705 uniquely mapped ratio ≥ 0.6 , peaks with fold change above 10 ≥ 500 , peaks union DHS ratio ≥ 0.7 , and
706 PBC ≥ 0.8 . Datasets with the below cell type labels are included: “Monocyte”, “B Lymphocyte”, “Erythroid
707 cell”, “Erythroid Progenitor Cell”, and “Erythroid progenitor”.

708

709 The K562 datasets from unibind were used for benchmarking with previous methods, including HINT-ATAC
710 and TOBIAS. In short, the same ATAC-seq data was used as input to all three methods. To guarantee fair
711 comparison, we first took the intersection of candidate TF binding sites from all three methods. Then for each
712 method, we ranked the remaining candidate sites by predicted binding score, and evaluated precision of
713 prediction using the top 10% sites. Only TFs with at least 10% of motifs overlapping with unibind validated TF
714 binding sites were included. Visualization of predicted and ground truth binding sites was done with the Gviz
715 package⁶⁰. Furthermore, to evaluate the false positive rate of each model, we also tested all three models on
716 our BAC naked DNA data. The same data was used as input to each model and the number of predicted
717 binding events are used to represent the false positive predictions.

718

719 *Model Interpretation*

720 To interpret how the TF binding model makes predictions, we first calculated gradients of the output with
721 respect to input features using GradientTape from tensorflow⁶¹. Gradients were computed for each individual
722 motif site and then the gradients for all sites were averaged and smoothed to generate the final gradient map.
723 Additionally, we performed ablation tests to evaluate the contribution of TF and nucleosome footprints. Here,
724 we masked TF footprints (20 bp, 40 bp, and 60 bp footprints) or nucleosome footprints (100 bp, 160 bp, and
725 200 bp footprints) separately during training by setting the corresponding features to zeros. The performances
726 of the models trained on masked data were then compared to the model trained on unmasked data.
727 Furthermore, we performed simulation analysis to study the impact of nucleosome positioning and width on
728 model prediction. To simulate TF and nucleosome footprints that reflect their real sizes, we first generated a
729 gaussian signal to represent the initial guess for their size (30 bp for TFs, 50 bp for nucleosomes). Next, we
730 went through footprints called on the HepG2 dataset at the corresponding scale and found examples that
731 highly correlates (Pearson correlation > 0.8) with our initial guess profile. The matched data was eventually
732 averaged to get the realistic width of footprints at each scale. We next simulated a TF signal at the center of
733 the motif, as well as signals of two flanking nucleosomes. By changing either the positioning or width of the
734 flanking nucleosomes, we were able to evaluate their impact on TF binding by observing the changes in the
735 predicted TF habitation scores.

736

737 Segmentation of sub-cCREs

738 cCREs are segmented into sub-cCREs using an approach conceptually similar to segmentation algorithms
739 for topologically associating domains (TADs)⁶². In brief, each cCRE is first divided into 10 bp intervals as
740 candidate binding sites. Suppose there are k intervals in a specific cCRE. We first compute TF habitation
741 scores for these k candidate binding sites across all n pseudo-bulks, resulting in a k -by- n matrix S of TF
742 binding scores. We then calculate the pairwise correlation among the rows, obtaining a k -by- k correlation
743 matrix M . Then for each TF site, we assign a score to it indicating whether it should be a sub-cCRE boundary.
744 More specifically, for each TF site, we first calculate 3 separate scores representing the average correlation
745 (1) within the upstream neighboring TF sites, (2) across the current TF site, and (3) within the downstream TF
746 sites. With a local neighborhood of radius r , we have:

747

748

$$ScoreA = \sum_{i=\max(1,x-r+1)}^x \sum_{j=\max(1,x-r+1)}^x \frac{M_{i,j}}{\min(x,r)^2}$$

$$ScoreB = \sum_{i=\max(1,x-r+1)}^x \sum_{j=x+1}^{\min(k,x+r)} \frac{M_{i,j}}{\min(x,r) \times \min(r,k-x)}$$

$$ScoreC = \sum_{i=x+1}^{\min(k,x+r)} \sum_{j=x+1}^{\min(k,x+r)} \frac{M_{i,j}}{\min(r,k-x)^2}$$

The boundary score is defined as

$$boundary\ score = \max(ScoreA, ScoreC) - ScoreB$$

We then smooth the boundary score and identify local maximum positions as the final sub-cCRE boundaries. Furthermore, the average predicted TF habitation scores within a sub-cCRE is used as the activity of the latter. For details on cCRE segmentation, see Supplementary Notes.

Calculating cCRE-gene correlation and sub-cCRE-gene correlation

To investigate whether sub-cCREs can be associated with gene expression independent from the cCRE they reside in, we compared cCRE-gene correlation and sub-cCRE-gene correlation. For each gene, we identified cCREs within a 50 kb radius from the gene TSS. We then calculated correlation between cCRE accessibility (as quantified by ATAC signal within the 1 kb window) and RNA level of the gene across pseudo-bulks. Meanwhile, we also took sub-cCREs detected within the same cCRE and correlated their activities with RNA of the same gene. Low signal sub-cCREs with an activity < 0.3 were removed. To also assign statistical significance to each cCRE-gene pair or sub-cCRE-gene pair that we correlated, we constructed a background distribution of correlation values. In the case of cCREs, we first randomly selected 100,000 cCRE-gene pairs and calculated their observed correlation. For each cCRE-gene pair, we also recorded average gene expression across pseudo-bulks, average cCRE accessibility across pseudo-bulks, and GC content of the cCRE as the 3 main features of the cCRE-gene pair. Therefore, each pair will have a unique coordinate in the 3-dimensional feature space (i.e., accessibility, expression, GC content). Next, for each cCRE-gene pair that we wish to assign significance to, we find its 100 nearest neighbors in the 3-dimensional feature space. The correlation values of the nearest neighbor cCRE-gene pairs are used as the background distribution. A z-test is performed to obtain a *p*-value for the pair of interest. In the case of sub-cCREs, a similar procedure is conducted. The only difference is that we use mean activity of the sub-cCRE instead of mean accessibility as a main feature. Again for every sub-cCRE-gene pair 100 background pairs are found in the 3-dimensional feature space and a z-test is performed to get the *p*-value.

Tracking TF binding dynamics across human hematopoiesis

Generation of pseudo-bulks

Single cells in the human BMMC dataset were first embedded into lower dimensional space using cisTopic⁵⁰, and then grouped into 1000 pseudo-bulks based on their spatial proximity in the cisTopic space. More specifically, we first sample 1000 cells as pseudo-bulk centers, and then identify *k*-nearest neighbors (*k* = 5000) of each center cell in the cisTopic space as other members of the same pseudo-bulk. We reasoned that sampling center cells with low local connectivity can help increase coverage of the state space by preventing over-sampling of densely connected local neighborhoods. Therefore, we first randomly sampled 10,000 scaffold cells and used them to construct a KNN graph (*k* = 10). Then we selected the 1000 cells with the lowest in-degree in the KNN graph as pseudo-bulk centers.

Computing pseudo-time

Pseudo-time along human hematopoietic lineages was computed using the Palantir package⁶³. To reduce computing time, we randomly sampled 100,000 cells from the human BMMC dataset as scaffold cells. The cisTopic embedding of the scaffold cells as well as pseudo-bulk center cells are used as input to Palantir.

Pathway enrichment analysis

Gene set annotations used for pathway enrichment analyses are obtained using the msigdb_gsets function from the R package hypeR⁶⁴. For human data, we used the "Homo sapiens", "C5", "BP" gene sets, while for

801 mouse data, we used the "Mus musculus","C5","BP" gene sets. Pathway enrichment was calculated using
802 Fisher's exact test.

803
804 Tracking nucleosome binding dynamics across human hematopoiesis

805 To characterize nucleosome reorganization during human hematopoiesis, we implemented a custom script
806 for automatic tracking of nucleosome footprints. Given any lineage of interest, we ordered all pseudo-bulks in
807 the lineage by pseudo-time. Due to the sparsity of data, we applied a sliding pseudo-time window of 10
808 pseudo-bulks and combined the data in each window before footprinting. Nucleosome footprints (100 bp scale
809 footprints) were then called at each pseudo-time point. Next, we aggregated the data from all pseudo-bulks
810 and called nucleosome footprints. The aggregate nucleosome footprint centers were used as rough mapping
811 of the position for each nucleosome. For fine-mapping of nucleosome position, we defined a window of 100
812 bp in diameter centered at the aggregate nucleosome footprint position. At each pseudo-time point, we found
813 the position of maximum footprint score within this 100 bp window as the instantaneous position of the
814 nucleosome. The instantaneous position as well as footprint intensity of the nucleosome were recorded for
815 each nucleosome and pseudo-time point. Eventually, the position and footprint intensity of the same
816 nucleosome were compared across pseudo-time to analyze sliding / binding / eviction of the nucleosome.

817
818 Characterizing age-related intra-cCRE dynamics

819 *Data preprocessing*

820 Cells with fraction of reads in peaks (FRIP) < 0.3 and depth < 300 were first removed. Additionally, we used
821 ArchR⁶⁵ to calculate doublet scores for each single cell and removed cells with top 5% doublet scores. The
822 remaining cells were then processed with the Seurat package⁶⁶. Cells were embedded into lower dimensional
823 space using latent semantic indexing (LSI)⁶⁷ and then clustered. Seurat clusters corresponding to HSCs were
824 selected for pseudo-bulking and downstream differential testing. Cells with the "LinNeg" FACS sort label were
825 excluded for HSC-specific analyses. To identify representative cell states, we used SEACells³⁸ to identify 30
826 representative cell states across HSCs. The representative cells are used as centers to form pseudo-bulks.
827 Each pseudo-bulk is generated by serially including nearest neighbor cells from the center cell in an order of
828 increasing distance until we reach a total of 5 million reads.

829
830 *Differential testing*

831 Differential RNA testing was performed using DESeq2⁶⁸. We first quantified total RNA read counts for each
832 gene in each pseudo-bulk, and used DESeq2 to identify significant differential genes with age as the covariate.
833 Additionally, we applied the TF habitation model to all cCREs and then segmented cCREs into sub-cCREs
834 using the method mentioned above. After filtering out low signal sub-cCREs (< 0.3 activity), we performed
835 differential cCRE and sub-cCRE testing with two-tailed unequal variance t-test (Welch's test) using the cCRE
836 accessibility and sub-cCRE activity, respectively.

837
838 *Motif score analysis*

839 Motif scores were calculated using chromVAR⁴⁹. Unlike the standard practice using the cCRE-by-sample
840 count matrix, here we used sub-cCRE-by-pseudo-bulk activity matrix. Only sub-cCREs with differential sub-
841 cCRE FDR < 0.1 as well as differential cCRE FDR > 0.1 are included for motif scoring. We performed
842 differential motif score analysis using two-tailed unequal variance t-test (Welch's test) between young and old
843 age groups.

844 **Data Availability**

845 Additional data such as pre-trained machine learning models and pre-computed Tn5 bias tracks can be
846 accessed on Zenodo at <https://zenodo.org/record/7121027#.ZCLo0ezMI8M>. Interactive visualization using
847 Shinyapps can be found at <https://buenrostromlab.shinyapps.io/ACAMShiny/> (human bone marrow) and
848 <https://buenrostromlab.shinyapps.io/aging/> (mouse HSC aging). Raw and processed sequencing data can be
849 found on Gene Expression Omnibus (GEO) with the accession number GSE216464.

850

851 **Code Availability**

852 All code used in this study, including tutorial for running PRINT, can be found at
853 <https://github.com/HYsxe/PRINT>.

854

855 **Acknowledgements**

856 We thank members of the Buenrostrom lab and the Wagers Lab for useful discussions and critical assessment
857 of this work. J.D.B. and the Buenrostrom lab acknowledge support from the Gene Regulation Observatory at
858 the Broad Institute of MIT & Harvard, the Chan Zuckerberg Initiative, the NHGRI IGVF consortium (UM1
859 HG011986) and the NIH New Innovator Award (DP2 HL151353). A.J.W. and the Wagers Lab acknowledge
860 support by grants from NIH (DP1 OD025432) and the Glenn Foundation for Medical Research (to A.J.W.) and
861 NIH F32 AG071208 to H.K. We thank J. LaVecchio N. Kheramand at the HSCI/HSCRB Flow Core for
862 assistance with flow cytometry and FACS. We thank A. Brack and C. Epstein for assistance in data generation.

863

864 **Author contributions**

865 Y. Hu led all computational developments and analyses described in this work with contributions from S. Ma,
866 V. Kartha, R. Zhang, A. Meliki, A. Castillo, N. Durand, E. Mattei, and N. Shores. S. Ma, F. Duarte, M. Horlbeck,
867 R. Shrestha, A. Labade, H. Kletzien, L. J. Anderson, T. Tay, and A. S. Earl generated the data with supervision
868 by C. B. Epstein, A. Wagers, and J. Buenrostrom. J. Buenrostrom supervised all aspects of this work. Y. Hu and
869 J. Buenrostrom wrote the manuscript with input from all authors.

870

871 **Declaration of Interests**

872 J. Buenrostrom holds patents related to ATAC-seq and is an SAB member of Camp4 and seqWell. J. Buenrostrom
873 and S. Ma holds a patent based on SHARE-seq. A.J.W. is a scientific advisor for Frequency Therapeutics and
874 Kate Therapeutics. A.J.W. is also a co-founder and scientific advisory board member and holds private equity
875 in Elevian, Inc., a company that aims to develop medicines to restore regenerative capacity. Elevian also
876 provides sponsored research to the Wagers lab.

877 **References**

- 878 1. Shlyueva, D., Stampfel, G. & Stark, A. Transcriptional enhancers: from properties to genome-wide
879 predictions. *Nat. Rev. Genet.* **15**, 272–286 (2014).
- 880 2. Gasperini, M., Tome, J. M. & Shendure, J. Towards a comprehensive catalogue of validated and target-
881 linked human enhancers. *Nat. Rev. Genet.* **21**, 292–310 (2020).
- 882 3. François Spitz & Eileen E. M. Furlong. Transcription factors: from enhancer binding to developmental control.
883 *Nat. Rev. Genet.* (2012) doi:10.1038/nrg3207.
- 884 4. Segal, E. & Widom, J. From DNA sequence to transcriptional behaviour: a quantitative approach. *Nat. Rev.*
885 *Genet.* **10**, 443–456 (2009).
- 886 5. Calo, E. & Wysocka, J. Modification of enhancer chromatin: what, how, and why? *Mol. Cell* **49**, 825–837
887 (2013).
- 888 6. ENCODE Project Consortium *et al.* Expanded encyclopaedias of DNA elements in the human and mouse
889 genomes. *Nature* **583**, 699–710 (2020).
- 890 7. de Dieuleveult, M. *et al.* Genome-wide nucleosome specificity and function of chromatin remodellers in ES
891 cells. *Nature* **530**, 113–116 (2016).
- 892 8. Lai, W. K. M. & Pugh, B. F. Understanding nucleosome dynamics and their links to gene expression and
893 DNA replication. *Nat. Rev. Mol. Cell Biol.* **18**, 548–562 (2017).
- 894 9. Jain, S. S. & Tullius, T. D. Footprinting protein-DNA complexes using the hydroxyl radical. *Nat. Protoc.* **3**,
895 1092–1100 (2008).
- 896 10. Galas, D. J. & Schmitz, A. DNase footprinting: a simple method for the detection of protein-DNA binding
897 specificity. *Nucleic Acids Res.* **5**, 3157–3170 (1978).
- 898 11. Hesselberth, J. R. *et al.* Global mapping of protein-DNA interactions in vivo by digital genomic footprinting.
899 *Nat. Methods* **6**, 283–289 (2009).
- 900 12. Neph, S. *et al.* An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* **489**,
901 83–90 (2012).
- 902 13. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native
903 chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and
904 nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).

- 905 14. Zentner, G. E. & Henikoff, S. High-resolution digital profiling of the epigenome. *Nat. Rev. Genet.* **15**, 814–827
906 (2014).
- 907 15. Degner, J. F. *et al.* DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature*
908 **482**, 390–394 (2012).
- 909 16. Stergachis, A. B. *et al.* Developmental fate and cellular maturity encoded in human regulatory DNA
910 landscapes. *Cell* **154**, 888–903 (2013).
- 911 17. Neph, S. *et al.* Circuitry and dynamics of human transcription factor regulatory networks. *Cell* **150**, 1274–
912 1286 (2012).
- 913 18. Stergachis, A. B. *et al.* Conservation of trans-acting circuitry during mammalian regulatory evolution. *Nature*
914 **515**, 365–370 (2014).
- 915 19. He, H. H. *et al.* Refined DNase-seq protocol and data analysis reveals intrinsic bias in transcription factor
916 footprint identification. *Nat. Methods* **11**, 73–78 (2014).
- 917 20. Adey, A. *et al.* Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro
918 transposition. *Genome Biol.* **11**, R119 (2010).
- 919 21. Bentsen, M. *et al.* ATAC-seq footprinting unravels kinetics of transcription factor binding during zygotic
920 genome activation. *Nat. Commun.* **11**, 4267 (2020).
- 921 22. Li, Z. *et al.* Identification of transcription factor binding sites using ATAC-seq. *Genome Biol.* **20**, 45 (2019).
- 922 23. Schep, A. N. *et al.* Structured nucleosome fingerprints enable high-resolution mapping of chromatin
923 architecture within regulatory regions. *Genome Res.* **25**, 1757–1770 (2015).
- 924 24. Voong, L. N. *et al.* Insights into Nucleosome Organization in Mouse Embryonic Stem Cells through Chemical
925 Mapping. *Cell* **167**, 1555–1570.e15 (2016).
- 926 25. Puig, R. R., Boddie, P., Khan, A., Castro-Mondragon, J. A. & Mathelier, A. UniBind: maps of high-confidence
927 direct TF-DNA interactions across nine species. *BMC Genomics* **22**, 482 (2021).
- 928 26. Ma, S. *et al.* Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin. *Cell* **183**,
929 1103–1116.e20 (2020).
- 930 27. Niccoli, T. & Partridge, L. Ageing as a risk factor for disease. *Curr. Biol.* **22**, R741–52 (2012).
- 931 28. Pal, S. & Tyler, J. K. Epigenetics and aging. *Sci Adv* **2**, e1600584 (2016).
- 932 29. Zhang, W., Qu, J., Liu, G.-H. & Belmonte, J. C. I. The ageing epigenome and its rejuvenation. *Nat. Rev. Mol.*

- 933 *Cell Biol.* **21**, 137–150 (2020).
- 934 30. Riedel, C. G. *et al.* DAF-16 employs the chromatin remodeller SWI/SNF to promote stress resistance and
935 longevity. *Nat. Cell Biol.* **15**, 491–501 (2013).
- 936 31. Morrison, S. J., Wandycz, A. M., Akashi, K., Globerson, A. & Weissman, I. L. The aging of hematopoietic
937 stem cells. *Nat. Med.* **2**, 1011–1016 (1996).
- 938 32. Chambers, S. M. *et al.* Aging hematopoietic stem cells decline in function and exhibit epigenetic
939 dysregulation. *PLoS Biol.* **5**, e201 (2007).
- 940 33. Rossi, D. J. *et al.* Cell intrinsic alterations underlie hematopoietic stem cell aging. *Proc. Natl. Acad. Sci. U. S.*
941 *A.* **102**, 9194–9199 (2005).
- 942 34. Evans, M. A. & Walsh, K. Clonal Hematopoiesis, Somatic Mosaicism, and Age-Associated Disease. *Physiol.*
943 *Rev.* (2022) doi:10.1152/physrev.00004.2022.
- 944 35. Jaiswal, S. & Ebert, B. L. Clonal hematopoiesis in human aging and disease. *Science* **366**, (2019).
- 945 36. Sun, D. *et al.* Epigenomic Profiling of Young and Aged HSCs Reveals Concerted Changes during Aging that
946 Reinforce Self-Renewal. *Cell Stem Cell* vol. 14 673–688 Preprint at
947 <https://doi.org/10.1016/j.stem.2014.03.002> (2014).
- 948 37. Mansell, E. *et al.* Mitochondrial Potentiation Ameliorates Age-Related Heterogeneity in Hematopoietic Stem
949 Cell Function. *Cell Stem Cell* **28**, 241–256.e6 (2021).
- 950 38. Persad, S. *et al.* SEACells: Inference of transcriptional and epigenomic cellular states from single-cell
951 genomics data. *bioRxiv* 2022.04.02.486748 (2022) doi:10.1101/2022.04.02.486748.
- 952 39. Hsu, A.-L., Murphy, C. T. & Kenyon, C. Regulation of aging and age-related disease by DAF-16 and heat-
953 shock factor. *Science* **300**, 1142–1145 (2003).
- 954 40. Labbadia, J. *et al.* Mitochondrial Stress Restores the Heat Shock Response and Prevents Proteostasis
955 Collapse during Aging. *Cell Rep.* **21**, 1481–1494 (2017).
- 956 41. Moll, L. *et al.* The insulin/IGF signaling cascade modulates SUMOylation to regulate aging and proteostasis
957 in *Caenorhabditis elegans*. *Elife* **7**, (2018).
- 958 42. Ho, T. T. *et al.* Autophagy maintains the metabolism and function of young and old stem cells. *Nature* **543**,
959 205–210 (2017).
- 960 43. Itokawa, N. *et al.* Epigenetic traits inscribed in chromatin accessibility in aged hematopoietic stem cells. *Nat.*

- 961 *Commun.* **13**, 2691 (2022).
- 962 44. Vierstra, J. *et al.* Global reference mapping of human transcription factor footprints. *Nature* **583**, 729–736
963 (2020).
- 964 45. Bergsland, M. *et al.* Sequentially acting Sox transcription factors in neural lineage development. *Genes Dev.*
965 **25**, 2453–2464 (2011).
- 966 46. Jing, H. *et al.* Exchange of GATA factors mediates transitions in looped chromatin organization at a
967 developmentally regulated gene locus. *Mol. Cell* **29**, 232–242 (2008).
- 968 47. Zhang, Y. *et al.* Model-based analysis of ChIP-seq (MACS). *Genome Biol.* **9**, R137 (2008).
- 969 48. Heinz, S. *et al.* Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory
970 Elements Required for Macrophage and B Cell Identities. *Molecular Cell* vol. 38 576–589 Preprint at
971 <https://doi.org/10.1016/j.molcel.2010.05.004> (2010).
- 972 49. Schep, A. N., Wu, B., Buenrostro, J. D. & Greenleaf, W. J. chromVAR: inferring transcription-factor-
973 associated accessibility from single-cell epigenomic data. *Nat. Methods* **14**, 975–978 (2017).
- 974 50. Bravo González-Blas, C. *et al.* cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data. *Nat.*
975 *Methods* **16**, 397–400 (2019).
- 976 51. McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: Uniform Manifold Approximation and Projection.
977 *Journal of Open Source Software* vol. 3 861 Preprint at <https://doi.org/10.21105/joss.00861> (2018).
- 978 52. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888–1902.e21 (2019).
- 979 53. Nair & Hinton. Rectified linear units improve restricted boltzmann machines. *icml* (2010).
- 980 54. Chollet, F. & Others. Keras. <https://keras.io> (2015).
- 981 55. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. Preprint at
982 <https://doi.org/10.48550/ARXIV.1412.6980> (2014).
- 983 56. Kent, W. J. The Human Genome Browser at UCSC. *Genome Research* vol. 12 996–1006 Preprint at
984 <https://doi.org/10.1101/gr.229102>. (2002).
- 985 57. Schep, A. *motifmatchr: Fast Motif Matching in R.* (2022).
- 986 58. Li, Z. *et al.* Chromatin-accessibility estimation from single-cell ATAC data with scOpen. *bioRxiv* 865931
987 (2021) doi:10.1101/865931.
- 988 59. Mei, S. *et al.* Cistrome Data Browser: a data portal for ChIP-seq and chromatin accessibility data in human

- 989 and mouse. *Nucleic Acids Res.* **45**, D658–D662 (2017).
- 990 60. Hahne, F. & Ivanek, R. Visualizing Genomic Data Using Gviz and Bioconductor. *Methods Mol. Biol.* **1418**,
991 335–351 (2016).
- 992 61. Martín Abadi *et al.* TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. Preprint at
993 <https://www.tensorflow.org/> (2015).
- 994 62. Crane, E. *et al.* Condensin-driven remodelling of X chromosome topology during dosage compensation.
995 *Nature* **523**, 240–244 (2015).
- 996 63. Setty, M. *et al.* Characterization of cell fate probabilities in single-cell data with Palantir. *Nat. Biotechnol.* **37**,
997 451–460 (2019).
- 998 64. Federico, A. & Monti, S. hypeR: an R package for geneset enrichment workflows. *Bioinformatics* **36**, 1307–
999 1308 (2020).
- 1000 65. Granja, J. M. *et al.* ArchR is a scalable software package for integrative single-cell chromatin accessibility
1001 analysis. *Nat. Genet.* **53**, 403–411 (2021).
- 1002 66. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene
1003 expression data. *Nat. Biotechnol.* **33**, 495–502 (2015).
- 1004 67. Deerwester, Dumais, Furnas, Landauer & Harshman. Indexing by latent semantic analysis. *J. Am. Soc. Inf.*
1005 *Sci.*
- 1006 68. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data
1007 with DESeq2. *Genome Biol.* **15**, 550 (2014).