

1 **MACMIC Reveals Dual Role of CTCF in Epigenetic Regulation of Cell Identity**
2 **Genes**

3
4 Guangyu Wang^{1, 2, 3, 4, #}, Bo Xia^{1, 2, 3, 4, #}, Man Zhou⁸, Jie Lv^{1, 2, 3, 4}, Dongyu Zhao^{1, 2, 3, 4},
5 Yanqiang Li^{1, 2, 3, 4}, Yiwen Bu^{1, 2, 3, 4}, Xin Wang^{1, 2, 3, 4}, John P. Cooke^{2, 3, 4}, Qi Cao^{5, 6}, Min
6 Gyu Lee⁷, Lili Zhang^{2, 3, 4}, Kaifu Chen^{1, 2, 3, 4, *}

7

8 ¹Center for Bioinformatics and Computational Biology, Department of Cardiovascular
9 Sciences, Houston Methodist Research Institute, Houston, TX, USA

10 ²Center for Cardiovascular Regeneration, Department of Cardiovascular Sciences,
11 Houston Methodist Research Institute, Houston, TX, USA

12 ³Department of Cardiothoracic Surgeries, Weill Cornell Medical College, Cornell
13 University, New York, NY, USA

14 ⁴Houston Methodist Institute for Academic Medicine, Houston Methodist Research
15 Institute, Houston, TX, USA

16 ⁵Department of Urology, Feinberg School of Medicine, Northwestern University,
17 Chicago, IL, USA

18 ⁶Robert H. Lurie Comprehensive Cancer Center, Feinberg School of Medicine,
19 Northwestern University, Chicago, IL, USA

20 ⁷Department of Molecular and Cellular Oncology, The University of Texas MD Anderson
21 Cancer Center, Houston, TX, USA

22 ⁸College of Natural, Applied and Health Sciences, Wenzhou Kean University, Wenzhou,
23 Zhejiang, China

24

25 * Corresponding: Kaifu Chen, kchen2@houstonmethodist.org

26 # These authors contributed equally

27

28 **ABSTRACT**

29 Numerous studies of relationship between epigenomic features have focused on their
30 strong correlation across the genome, likely because such relationship can be easily
31 identified by many established methods for correlation analysis. However, two features
32 with little correlation may still colocalize at many genomic sites to implement important
33 functions. There is no bioinformatic tool for researchers to specifically identify such
34 feature pair. Here, we develop a method to identify feature pair in which two features
35 have maximal colocalization but minimal correlation (MACMIC) across the genome. By
36 MACMIC analysis of 3,385 feature pairs in 15 cell types, we reveal a dual role of CTCF
37 in epigenetic regulation of cell identity genes. Although super-enhancers are associated
38 with activation of target genes, only a subset of super-enhancers colocalized with CTCF
39 regulate cell identity genes. At super-enhancers colocalized with CTCF, the CTCF is
40 required for the active marker H3K27ac in cell type requiring the activation, and also
41 required for the repressive marker H3K27me3 in other cell types requiring the repression.
42 Our work demonstrates the biological utility of the MACMIC analysis and reveals a key
43 role for CTCF in epigenetic regulation of cell identity.

44

45 **Keywords:** mutual information, correlation, CTCF, super-enhancer, H3K27ac,
46 H3K27me3

47

48

49 INTRODUCTION

50 As DNA sequencing data expands at an unprecedented speed, genomic (including
51 epigenomic) data such as RNA-seq, ChIP-seq and genome sequencing data can be
52 conveniently collected from public databases. Each set of sequencing data is typically
53 collected to investigate a genomic (including epigenomic) feature across the genome,
54 e.g., RNA-Seq dataset to investigate the expression profile of all genes in a genome,
55 ChIP-Seq dataset to investigate a histone modification or the binding of a transcription
56 factor at individual sites across the genome. It is commonly recognized that the function
57 of a genome cannot be fully understood by studying a single genomic feature. Many
58 studies showed that analysis of correlation between two genomic features had a strong
59 potential to identify their regulatory relationship in an important biological process¹. For
60 instance, a strong positive correlation between the binding intensity of a protein near
61 individual genes and the expression level of these genes might help define the protein to
62 be an activator of transcription². By focusing on the correlation between the RNA
63 expression and a histone modification, the roles of individual histone modifications in the
64 activation or repression of transcription have also been recognized^{3,4}.

65

66 However, in many aspects of informatics, the representation of knowledge can be more
67 efficient by using a combination of uncorrelated features⁵. In other words, highly
68 correlated features often contain redundant information⁶. For example, whereas the
69 dozens of pluripotent factors such as Oct4, Sox2, Klf-4, and c-Myc, are all useful to
70 predict genes expressed in stem cells⁷⁻⁹, combining some pluripotent factors with
71 endothelial lineage factors such as Lmo2 and Erg would add power to also predict genes
72 expressed in endothelial cells; therefore, it can be more powerful using combined
73 information from transcription factors with distinct functions, as opposed to an analysis
74 using the transcription factors with similar effects on a shared set of target genes. More

75 importantly, colocalization of low-correlation chromatin features may still happen in a
76 biologically considerable manner to implement important functions. For instance, the
77 histone modifications H3K27me3 and H3K4me3 are known to be associated with
78 repression and activation of transcription in differentiated cells, respectively¹⁰. As a result,
79 they show negative correlation and often occur at different genes in somatic cell types¹¹.
80 However, these two markers lose the correlation and colocalize at a large set of genes in
81 embryonic stem cells¹². It is well known now that the colocalizations of H3K27me3 and
82 H3K4me3 in embryonic stem cells define bivalent chromatin domains, which are
83 functionally distinct from both the repressive domains associated with H3K27me3 and
84 the active domains associated with H3K4me3. These bivalent chromatin domains play
85 a unique role in embryonic stem cells to maintain a bivalent status of the lineage factors
86 for individual somatic cell types¹³⁻¹⁵. Therefore, analyzing colocalization of two chromatin
87 features with globally low correlation in a cell has the potential to reveal novel biological
88 mechanisms. However, little is known yet about the biological implications of such
89 colocalization for the other chromatin features beyond H3K4me3 and H3K27me3.
90 Therefore, the community is in need of a robust method to identify and understand the
91 biologically important colocalization of uncorrelated chromatin features in a cell.

92

93 In this study, we utilized mutual information¹⁶ as an indication for general correlation
94 (relevance) between a pair of genomic features, and mathematically integrated it with
95 the number of colocalizations between the features to define a score for maximal
96 colocalization minimal correlation (MACMIC). The MACMIC score allows us to
97 quantitatively prioritize the feature combinations that have large number of
98 colocalizations but low correlation. We next performed a systematic analysis of MACMIC
99 score between chromatin features using 1,522 datasets for histone modifications or the
100 binding of chromatin proteins from embryonic stem cells as well as somatic cell types.

101 Our analysis successfully recaptured the previously discovered bivalent domain in
102 embryonic stem cells, and further revealed a key role for CCCTC-Binding Factor (CTCF)
103 in the epigenetic regulation of cell identity genes.

104

105 **MATERIAL AND METHODS**

106 **Data collection**

107 The RNA-seq, transcript factors and histone modifications ChIP-seq data for human
108 primary cells, human embryonic stem cells and mouse embryonic stem cells were
109 downloaded from GEO database and ENCODE project website
110 (<https://www.encodeproject.org/>)¹⁷. Processed annotated topologically associating
111 domains and loops from HUVEC⁴⁰ were downloaded from GEO. Detailed information of
112 datasets reanalyzed in this study was listed in Table S1 and Table S2.

113

114 **Data processing and analysis**

115 Human reference genome sequence version hg19, mouse reference genome sequence
116 version mm9 and UCSC Known Genes were downloaded from the UCSC Genome
117 Browser website¹⁸. TPMs of RNA-seq from ENCODE were directly downloaded from
118 ENCODE project. For GEO datasets, RNA-seq raw reads were mapped to the human
119 genome version hg19 using TopHat version 2.1.1 with default parameter values. The
120 expression value for each gene was determined by the function Cuffdiff in Cufflinks
121 version 2.2.1 with default parameter values.

122

123 For ChIP-seq data, reads were first mapped to reference genome by Bowtie version
124 1.1.0. Peak calling and generation of .wig file were performed by DANPOS 2.2.3. Bigwig
125 was generated using the tool WigToBigWig. The tool WigToBigWig was downloaded
126 from the ENCODE project website (<https://www.encodeproject.org/software/wigtobigwig/>)

127 ¹⁷. Then bigwig file was submitted to the UCSC Genome Browser
128 (<https://genome.ucsc.edu>) to visualize the ChIP-Seq signal at each base pair ^{18,19}. The
129 average density plots of epigenetic marks in promoter region around TSS were plotted
130 using the Profile function in DANPOS version 2.2.3. Heatmap was plotted using
131 Morpheus (<https://software.broadinstitute.org/morpheus>). P values of boxplots were
132 calculated with a two-sided Wilcoxon test. For the regulation network, we used CellNet
133 method²⁰ to define the network and downloaded the network nodes (genes), edges and
134 value of closeness between nodes from CellNet website (<http://cellnet.hms.harvard.edu/>).
135 As the gene number will affect the percentage and p value of overlap between gene
136 groups, we used the same number of top genes from each group to avoid this effect.
137 Because the genes associated with Broad H3K4me3 was reported to be around 500 in
138 each cell type ²¹, we used this number of genes for each gene group.

139

140 **Integrated analysis of two chromatin features**

141 For individual markers, the ranking of genes was based on the width of individual
142 markers on the gene promoter region (upstream 3kb of TSS to downstream 10kb of
143 TSS). For the ranking of genes based on the colocalization of two chromatin features,
144 the rank product of two individual markers was calculated first. We defined rank product
145 as $RP = \sqrt{\prod_{i=1}^n r_{1,i} * r_{2,i}}$, where the $r_{1,i}$ is the rank of wide for the first marker, the $r_{2,i}$ is
146 the rank of wide for the second marker. Then if no colocalization of these two chromatin
147 markers was detected in the gene promoter region, the gene was being removed from
148 the ranking. A colocalization of two chromatin markers at a specific genomic locus was
149 defined by requiring at least 1bp overlap. To measure the colocalization level of two
150 chromatin markers, we calculated the total number of genomic loci that display overlap
151 of these two chromatin markers across whole genome. Afterward, the genes associated

152 with the colocalization of these two chromatin features were ranked based on the rank
153 product of individual features. For a fair comparison, each group defined by H3K4me3,
154 H3K27ac, H3K27me3, colocalization of broad H3K4me3 and broad H3K27me3,
155 colocalization of broad H3K4me3 and broad H3K27ac contained only the top 500 genes.
156 GO term pathway analysis was performed by the web portal (<http://geneontology.org/>)²².

157

158 **Calculation of MACMIC score**

159 To calculate MACMIC score, we first calculated Mutual information (MI) that is a widely
160 used measure of the mutual dependence between two variables. More specifically, MI
161 measures how much does the knowledge of one variable reduces uncertainty of the
162 other variable. If two chromatin markers have larger MI, these two chromatin markers
163 share more information and are less independent from each other. Mathematically,
164 mutual information is calculated by following equations:

165

$$I(X; Y) = H(X) + H(Y) - H(X, Y)$$

166

167 where X and Y represent the peak width from two different chromatin features, $I(X; Y)$ is
168 the mutual information of X and Y . $H(X)$ and $H(Y)$ are the marginal entropies and $H(X, Y)$
169 is the joint entropy of X and Y . Entropies are calculated by the following equation:

170

$$H(X) = - \sum_{i=1}^n P(x_i) \log P(x_i)$$

171 where n is the total gene number, $P(x_i)$ is the probability by which the total signal of a
172 given genomic marker is x_i in the promoter region of gene i . To calculate $H(X)$, we
173 focused on the promoter region from 3kb upstream to 10kb downstream of transcription
174 start site. For a promoter that has multiple ChIP-seq peaks, we calculated the total signal

175 that is the sum of signals in these peaks. The function Selector in DANPOS was used to
176 map peaks to promoters. And we used Poisson distribution to calculate the probability of
177 the observed ChIP-seq signal in a given promoter region²³. To calculate the joint
178 entropy of two genomic features, we used the following equation:

$$H(X, Y) = - \sum_{i=1}^n P(x_i, y_i) \log P(x_i, y_i),$$

179 where n is the total gene number, $P(x_i, y_i)$ is the joint probability that the total signals of
180 the first and second markers are x_i and y_i in the promoter region of gene i .

181

182 Considering the penalty of high correlation feature pairs, MACMIC score is calculated by
183 the following equation:

184

$$MIRC = \frac{C_{observed} - C_{expected}}{C_{expected}}$$

185

186 where C represents the colocalization of two chromatin features which is counted by the
187 number of overlap events. The p-value for each term tests the null hypothesis that the
188 residual is equal to zero. A low p-value (<0.05) indicates that for a specific value of MI,
189 the feature combinations have a significant higher colocalization than the estimated
190 colocalization on the genome.

191

192 **CTCF associated super-enhancers**

193 CTCF ChIP-seq datasets were processed as previously described. Peaks with height
194 larger than upper quartile of peak height values were defined as high confidence CTCF
195 peaks. Super-enhancers were defined as previous defined²⁴, and then super-enhancers
196 were categorized into two categories based on the existence of high confidence CTCF

197 peaks within super-enhancers. Super-enhancers with high confidence CTCF peaks were
198 named as CTCF associated super-enhancers (CSEs). Super-enhancers without high
199 confidence CTCF peaks were named as CTCF associated super-enhancers (OSEs).

200

201 **Simulation of association between CTCF and enhancers**

202 For each group of typical enhancers, each typical enhancer was randomly matched to a
203 super-enhancer and then typical enhancers were enlarged towards two directions until
204 they had the same size as super-enhancers. Associations of CTCF and super-
205 enhancers, typical enhancers and enlarged typical enhancers were calculated based on
206 the overlap events between these two different epigenetic markers.

207

208 **RESULTS**

209 **Colocalization of globally low-correlation chromatin features reveals unique** 210 **functional pathways**

211 We first tested whether the colocalization of two histone modifications could identify
212 genes that were not effectively identified by each of the two modifications. We performed
213 the analysis for H3K4me3 and H3K27ac that had strong correlation across the genome
214 (**Fig. S1A**) and compared it to the analysis for H3K4me3 and H3K27me3 that had little
215 correlation across the genome (**Fig. S1B**) in human stem cell H1. We recently revealed
216 that the top 500 genes associated with broad H3K4me3 were enriched with tumor
217 suppressor genes²¹. For a fair comparison, we retrieved the top 500 genes associated
218 with broad H3K27ac and the top 500 genes associated with broad H3K27me3. There
219 were 288 (57.6%) genes associated with both broad H3K4me3 and broad H3K27ac (**Fig.**
220 **S1C**). In contrast, there was no gene associated with both broad H3K4me3 and broad
221 H3K27me3 (**Fig. S1D**). To further explore the potential colocalization between H3K4me3
222 and H3K27me3, we defined the top 500 genes by the rank product of H3K4me3 width

223 and H3K27me3 width (H3K4me3&H3K27me3 broad colocalization) (**Fig. S1E**). We also
224 defined the top 500 genes by the rank product of H3K4me3 width and H3K27ac width
225 (H3K4me3&H3K27ac broad colocalization) (**Fig. S1E**). For the genes associated with
226 H3K4me3&H3K27ac broad colocalization, only 7 genes were not captured by broad
227 H3K4me3 or broad H3K27ac (**Fig. S1C**). However, for the genes associated with
228 H3K4me3&H3K27me3 broad colocalization, 421 (84.2%) genes were not captured by
229 broad H3K4me3 or broad H3K27me3 (**Fig. S1D**). Further, for the 2168 pathways
230 significantly enriched in genes associated with H3K4me3&H3K27me3 broad
231 colocalization, 1404 pathways showed no significant enrichment in genes associated
232 with broad H3K4me3 or broad H3K27me3 (**Fig. S1F**). These pathways were mainly
233 related to somatic cell lineage specification (**Fig. S1G**), which agreed with the reported
234 role of bivalent domains. These results suggested that colocalization of globally low-
235 correlation features in a cell could be associated with unique biological implications that
236 were not associated with the localization of each of these features.

237

238 **A new method to identify features with maximal colocalization minimal correlation** 239 **(MACMIC)**

240 Here, we used mutual information as an indication for correlation because mutual
241 information is more general than other methods such as linear correlation. A large
242 mutual information value will indicate strong correlation that can be either positive or
243 negative, and either linear or nonlinear. Theoretically, two features that have a small
244 mutual information value tend to have no or a small number of colocalization, whereas a
245 large number of colocalizations are often associated with large mutual information value.
246 To develop a simple method to prioritize feature pairs that have minimal correlation but a
247 maximal number of colocalizations, we first performed a systematic analysis of the
248 relationship between the mutual information value and the number of colocalizations for

249 225 feature pairs derived from 6 chromatin features in 15 cell types (**Table S1**). We
250 analyzed 6 features, which formed 15 pairs with each other in each cell type and thus
251 resulted in 225 feature pairs in 15 cell types (**Table S3**). Most feature pairs displayed a
252 positive correlation between the mutual information value and the number of
253 colocalizations (Spearman correlation coefficient 0.46) (**Fig. 1A**). Similar results were
254 observed by replacing mutual information with absolute value of correlation coefficient or
255 PCA value (**Fig. S2**). However, there were a few feature pairs that displayed a large
256 number of colocalizations but small mutual information value (**Fig. 1A**). We therefore
257 developed a regression model that employed the mutual information value to determine
258 an expected number of colocalizations, and next utilized the normalized discrepancy
259 between the observed and the expected numbers of colocalizations as a measurement
260 of the MACMIC (**Fig. 1B**). We calculated the MACMIC scores for the 225 individual
261 feature pairs and found that the large MACMIC scores effectively prioritized feature pairs
262 that possessed large number of colocalizations but weak correlations (**Fig. 1C**). We
263 further tested our MACMIC analysis method on 3160 feature pairs derived from 80
264 chromatin features in human ESC H1. Our result again indicated that MACMIC
265 successfully prioritized the feature pairs with minimal mutual information but substantial
266 colocalizations (**Fig. 1D**).

267

268 **MACMIC identifies a unique association of CTCF with super-enhancer**

269 To further test whether MACMIC scores could effectively recapture feature pairs with
270 biological implications, we analyzed MACMIC scores between H3K4me3 and
271 H3K27me3 in 15 human primary cell types. In agreement with the reported large number
272 of bivalent domains marked by both H3K4me3 and H3K27me3 in embryonic stem cells,
273 we observed a large MACMIC score (2.8) in the H1 cell. On the other hand, in
274 agreement with the reported resolution of bivalent domains to form either repressive

275 domains marked by H3K27me3 or active domains marked by H3K4me3, the MACMIC
276 scores between H3K4me3 and H3K27me3 were low in all the 14 somatic cell types
277 (from -0.68 to 0.67) (**Fig. 2A**). Therefore, MACMIC analysis successfully recaptured
278 bivalent domains that were known to play a key role in embryonic stem cells.

279

280 We next tested whether MACMIC analysis could successfully identify new feature pairs
281 that possess a large number of functionally important colocalizations but low correlation.
282 We ranked a set of 79 chromatin features in H1 cells by the MACMIC scores between
283 the enhancer feature H3K27ac and each of these features (**Fig. 2B**). The top features
284 with the large MACMIC scores in the rank included the suppressive histone modification
285 H3K27me3, consistent with the implication that H3K27ac and H3K27me3 might co-exist
286 in bivalent domains. Interestingly, master regulators of three-dimensional chromatin
287 interaction, the CTCF²⁵ and its binding partner RAD21²⁶, topped in the rank list (**Fig. 2B**).
288 We further performed analysis in 14 human somatic cell types that each had ChIP-seq
289 datasets for a set of 6 chromatin features from the ENCODE project¹⁷ (**Table S1**). The
290 results showed that the MACMIC score between H3K27ac and the binding of CTCF was
291 significantly larger than MACMIC scores between H3K27ac and the other 4 features
292 including H3K27me3, H3K4me3, H3K9me3 and H3K79me2 (**Fig. 2C**). Moreover,
293 colocalization analysis for CTCF and H3K27ac found that CTCF binding sites had the
294 largest number of colocalizations with the broadest H3K27ac peaks (super-enhancers)
295 (**Fig. 2D**). To test whether this higher frequency of colocalization was simply due to the
296 longer DNA sequences of super-enhancers, we performed a normalization by
297 lengthening typical enhancers at the two ends of each enhancer, so that the DNA
298 sequences assigned to typical enhancers had equivalent sizes to those of super-
299 enhancers. The result showed that the frequency of colocalization with CTCF binding

300 sites still tended to be higher for super-enhancers when compared to other enhancers
301 **(Fig. 2D)**.

302

303 **A unique enrichment of CTCF associated super-enhancer in cell identity genes**

304 Since super-enhancers were reported to regulate cell identity genes²⁴, we determined to
305 investigate the role of CTCF in this regulation. We divided super-enhancers into two
306 categories, i.e., CTCF associated super-enhancers (CSEs) and other super-enhancers
307 (OSEs). To study the function of genes marked by CSEs and OSEs, we defined the
308 genes of which the gene body overlapped with CSEs (or OSEs) for at least 1bp as the
309 CSEs (or OSEs) marked genes. Intriguingly, only the genes marked by CSEs were
310 significantly enriched in the pathways associated with cell lineage specifications, e.g.,
311 the endothelial cell differentiation pathway (GO:0045601) for CSEs in human umbilical
312 vein endothelial cells (HUVECs) **(Fig. 3A)** and the neuron differentiation pathway
313 (GO:0045664) for CSEs in neural cells **(Fig. 3B)**. Manual inspection of individual known
314 cell lineage factors in these cell types further confirmed the colocalization of ChIP-seq
315 signals of H3K27ac and CTCF, e.g., at the gene NR2F2²⁷ in endothelial cells and the
316 gene FOXP1²⁸ in neural cells **(Fig. 3C, D)**. In contrast, some other genes, although also
317 displaying broad enrichment of H3K27ac, were depleted of CTCF binding sites, e.g., at
318 the gene ARF1 in endothelial cells and the gene PON1 in neural cells **(Fig. 3C, D)**.
319 Intriguingly, there were typically multiple binding sites of CTCF located within the active
320 region of each CSE. This colocalization pattern was different from the well-known
321 function of CTCF binding sites as insulators, which often happened between active
322 domain and repressive domain. Besides, a significant portion of the genes associated
323 with CSEs encoded transcription factors, whereas we did not observe this phenomenon
324 for the genes associated with OSEs **(Fig. 3E)**. Further, the genes associated with CSEs
325 were connected to a significantly large number of edges in the gene regulatory networks,

326 whereas the numbers of connected network edges were similar for genes associated
327 with OSEs and random control genes (**Fig. 3F**). The differences between CSEs and
328 OSEs in their association with genes in cell lineage pathways were highly reproducible in
329 the other 15 primary cell types that we have analyzed (**Fig. 3G**). It was reported that the
330 establishment of cell type specific chromatin loops were important during cell
331 differentiation²⁹. Consistently, we found that CSEs were enriched near chromatin loops
332 (**Fig. S3A**) and the boundaries of topologically associating domains (TADs) (**Fig. S3B**),
333 whereas no significant differences in the sizes of the associated TADs were observed
334 between CSEs and OSEs (**Fig. S3C**).

335

336 **CSE- and OSE-associated genes have similar expression levels and cell type** 337 **specificities**

338 To understand how CTCF regulates enhancer activity and in turn regulates cell identity,
339 we first compared the expression levels of associated genes between CSEs and OSEs.
340 Intriguingly, similar expression levels were observed between CSE-marked genes and
341 OSE-marked genes, and this result was highly reproducible in HUVECs (**Fig. 4A left**
342 **panel**) and neural cells (**Fig. 4A right panel**). Further, CSEs genes and OSEs genes of
343 HUVECs were both significantly up regulated in HUVECs compared to embryonic stem
344 cells and neural cells (**Fig. 4B left panels**). Consistently, CSEs genes and OSEs genes
345 of neural cells were both significantly up regulated in neural cells compared to embryonic
346 stem cells and HUVECs (**Fig. 4B right panels**). These results suggested that CSEs and
347 OSEs genes of the same cell type had similar expression levels and cell type
348 specificities.

349

350 We next compared the H3K27ac levels between CSEs and OSEs, as H3K27ac is a
351 marker for enhancer activation. The result indicated that the H3K27ac levels were similar

352 at CSEs and OSEs within HUVECs (**Fig. 4C left**). Similarly, the H3K27ac levels were
353 similar at CSEs and OSEs within neural cells (**Fig. 4C right**). Further, the H3K27ac
354 levels at HUVEC-specific CSEs and OSEs were higher in HUVECs when compared to
355 the same regions in embryonic stem cells and neural cells, whereas the H3K27ac levels
356 at neuron-specific CSEs and OSEs were higher in neural cells compared to the same
357 regions in HUVECs and embryonic stem cells (**Fig. 4D**). Therefore, in agreement with
358 result from the expression analysis, CSEs and OSEs genes of the same cell type had
359 similar epigenetic states and specificities.

360

361 Of the top 500 HUVEC CSEs, 405 (81%) lost H3K27ac in neural cells and embryonic
362 stem cells (**Fig. 4E top left**). In contrast, the binding of CTCF in 483 (97%) HUVEC
363 CSEs were retained in both neural cells and embryonic stem cells (**Fig. 4E bottom left**).
364 Similar results were observed for the neural cell CSEs. Of the top 500 neural CSEs, 388
365 (78%) lost H3K27ac in HUVECs and embryonic stem cells (**Fig. 4E top right**), while the
366 binding of CTCF in 462 (92%) neural cell CSEs were retained in both HUVECs and
367 embryonic stem cells (**Fig. 4E bottom right**). To further understand the role of CTCF in
368 CSEs, we next analyzed an RNA-Seq dataset from HeLa cells with CTCF knocked down
369 or not. The genes associated with CSEs of HeLa cells were significantly enriched in the
370 genes down regulated but not in the genes up regulated in response to CTCF
371 knockdown (**Fig. 4F**). In contrast, the OSEs associated genes showed little enrichment
372 in the down or up regulated genes induced by knockdown of CTCF (**Fig. 4F**). Of the top
373 500 HUVEC OSEs, 331 (66%) lost H3K27ac in neural cells and embryonic stem cells
374 (**Fig. S4 top left**). In contrast, the binding of CTCF in 492 (98%) HUVEC OSEs were
375 retained in both neural cells and embryonic stem cells (**Fig. S4 bottom left**). Similar
376 results were observed for the neural cell OSEs. Of the top 500 neural OSEs, 347 (69%)
377 lost H3K27ac in HUVECs and embryonic stem cells (**Fig. S4 top right**), while the

378 binding of CTCF in 476 (96%) neural cell OSEs were retained in both HUVECs and
379 embryonic stem cells (**Fig. S4 bottom right**). These results indicated that although the
380 loss of the activation state of CSEs may not require the loss of CTCF bindings, the
381 bindings of CTCF were required for the activation of CSEs and their associated genes.

382

383 **CSEs of a given cell type display increased repressive modification H3K27me3 in** 384 **other cell types**

385 A cell identity gene has two key attributes: 1) it is associated with active chromatin
386 modifications and thus activated to play an important role in the cell type that requires its
387 activation; and 2) it is associated repressive chromatin modifications and thus silenced in
388 most other cell types. Since our results demonstrated that the CSEs of one cell type lost
389 H3K27ac but retained the binding of CTCF in other cell types, we hypothesized that the
390 binding of CTCF might be also important for the repressions of these CSEs in the other
391 cell types.

392

393 We first defined CSEs, OSEs, and a set of random control genes in HUVECs, and
394 analyzed the pattern of the repressive histone modification H3K27me3 on these 3 gene
395 sets in each of three cell types including embryonic stem cells, neural cells and also
396 HUVECs. We found that the H3K27me3 signals in HUVEC showed a similar pattern at
397 the HUVEC CSEs as at the HUVEC OSEs, but are substantially weaker than at the
398 random control genes (**Fig. 5A top**). Intriguingly, only the CSEs of HUVECs, not the
399 OSEs of HUVECs or the random control genes, were marked by strong H3K27me3
400 signals in embryonic stem cells (**Fig. 5A middle**). These trends observed for H3K27me3
401 in embryonic stem cells were the same for H3K27me3 in neural cells (**Fig. 5A bottom**).
402 Similar results were observed when we defined CSEs, OSEs, and a set of random
403 control genes in neural cells to analyze the pattern of H3K27me3 on these 3 gene sets

404 in HUVECs, embryonic stem cells, and neural cells. The H3K27me3 signals in neural
405 cells showed a similar pattern at the neural CSEs as at the neural OSEs, but are
406 substantially weaker at the random control genes (**Fig. 5B bottom**). However, only the
407 CSEs of neural cells, not the OSEs of neural cells or the random control genes,
408 possessed strong H3K27me3 signals in embryonic stem cells (**Fig. 5B middle**). These
409 trends observed for H3K27me3 in embryonic stem cells were the same for H3K27me3 in
410 HUVECs (**Fig. 5B top**).

411

412 We next further expanded our analysis to 15 sets of biosamples that each had ChIP-Seq
413 data for CTCF, H3K27ac, and H3K27me3. Consistent with the results from HUVECs and
414 neural cells, CSEs and OSEs showed similar enrichment of H3K27ac (**Fig. S5A**) and
415 similar depletion of H3K27me3 (**Fig. S5B**) in cell types that defined these CSEs and
416 OSEs. Next, we analyzed these CSEs and OSEs in H3K27ac ChIP-Seq datasets from
417 84 biosamples and H3K27me3 ChIP-Seq datasets from 125 biosamples from the
418 ENCODE database. CSEs and OSEs both showed attenuated enrichment of H3K27ac
419 when the H3K27ac was analyzed in cell types different from the cell types that defined
420 the CSEs and OSEs (**Fig. S5C**). However, the CSEs were associated with significant
421 enrichment of H3K27me3, whereas the OSEs showed little enrichment of H3K27me3,
422 when the H3K27me3 was analyzed in cell types different from the cell types that defined
423 these CSEs and OSEs (**Fig. S5D**). These analyses indicated that the CSEs, but not the
424 OSEs, were under stringent epigenetic repression by H3K27me3 in cell types different
425 from the cell types that defined the CSEs and OSEs. Interestingly, CTCF and
426 H3K27me3 is also among the top feature pairs ranked by MACMIC score in H1 hESC
427 (**Fig. S6**).

428

429 **CTCF in a given cell type is required for the repression of genes associated with**
430 **the CSEs defined in other cell types**

431 Importantly, auxin-induced degradation of CTCF in embryonic stem cells led to the loss
432 of CTCF bindings and H3K27me3 signals in embryonic stem cells at the CSEs genes
433 defined in HUVECs as well as the CSEs genes defined in neural cells. For example,
434 signals of CTCF bindings and H3K27me3 in embryonic stem cells at known identity
435 genes of somatic cell types, the NR2F2²⁷ of endothelial cells (**Fig. 5C left**) and the
436 FOXG1²⁸ of neural cells (**Fig. 5C right**), were substantially attenuated after auxin
437 induced degradation of CTCF, and recovered after auxin was washed off (**Fig. 5C**). The
438 CTCF binding sites in embryonic stem cells at these genes were located within the
439 broad H3K27me3 modifications. This colocalization of CTCF binding sites and broad
440 H3K27me3 in embryonic stem cells was similar to the colocalization observed for CTCF
441 binding sites and super-enhancers in HUVECs, neural cells, heart, fibroblast cell, and
442 bone marrow macrophage cell. Our further analysis indicated that in parallel with the
443 loss of CTCF bindings in embryonic stem cells at these genes (**Fig. 5D left and Fig. 5E**
444 **left**), the H3K27me3 signals in embryonic stem cells were reduced dramatically (**Fig. 5D,**
445 **Fig. 5E, Fig. S7 middle**) and the expressions in embryonic stem cells were significantly
446 up regulated (**Fig. 5D, Fig. 5E, Fig. S7 right**). Taken together, these results suggested
447 that the CTCF in a given cell type was required for the repression of genes associated
448 with the CSEs defined in a different cell type.

449

450 **DISCUSSION**

451 Conventional analysis of relationship between chromatin features tends to focus on
452 strongly positive or negative correlation to identify the associated components within a
453 specific biological process¹. However, genomic features with weak correlation across the
454 genome may still colocalize at many genomic sites in a biologically important manner. It

455 was hard to capture the significance of such colocalizations on the basis of conventional
456 correlation analysis. In this study, we provide a new method to identify MACMIC, which
457 effectively prioritize the feature pairs with low genome-wide correlation but substantial
458 colocalizations. Using the MACMIC, we successfully recapture the reported bivalent
459 domain in embryonic stem cells, which is composed of both activating histone
460 modifications, e.g., H3K4me3, and the repressive histone modifications, e.g., H3K27me3.
461 Activating histone modification and the repressive histone modification possess low
462 genome-wide correlation in the embryonic stem cell, but the colocalizations of them at
463 bivalent domains mark important lineage specific regulators.

464

465 As proof of principle, we present a novel relationship identified by MACMIC between the
466 bindings of CTCF and the enhancer marker H3K27ac. Our analysis demonstrated that
467 their colocalization is key to both the activation and repression of cell identity genes.
468 Numerous efforts have been made to understand cell identity regulation²⁰. Somatic cells
469 such as fibroblasts³⁰, keratinocytes³¹, peripheral blood cells³², and neural progenitor
470 cells³³, have been successfully reprogrammed to induced pluripotent stem cells. Many
471 transcription factors and epigenetic regulators have been proposed to play important
472 roles in these dynamic processes. We and several other groups recently discovered that
473 cell identity genes manifested unique chromatin epigenetic signatures associated with
474 their distinct transcriptional regulation mechanisms^{24,34-36}. CTCF is well known for its
475 function as an insulator that binds between active and repressive domains on
476 chromatin³⁷, as a mediator for promoter-enhancer interaction³⁸, and as a partner of
477 cohesin in regulating chromatin 3D structure^{39,40}. It further has been proven to be an
478 essential factor for cell differentiation and development of T cell⁴¹, Neuron⁴², Heart⁴³, and
479 Limb⁴⁴. However, how these functions of CTCF are connected to the regulation of cell
480 identity genes was not known.

481

482 In this study, we separate CSEs from OSEs based on the colocalization of CTCF binding
483 sites with H3K27ac signals in CSEs. Our results suggest that CTCF contributes to the
484 activations of CSEs in cell types that require the activations, and is involved in the
485 repression of CSEs in other cell types that require the repressions. Interestingly, only
486 CSEs showed significantly higher H3K27me3 signals in the cell types that required their
487 repressions, consistent with the notion that epigenetic repression of cell identity genes of
488 a given cell type is critical in other somatic cell types (**Fig. 5, S3**). In response to the loss
489 of CTCF function in embryonic stem cells, H3K27me3 signals in embryonic stem cells at
490 the CSEs of somatic cell types were dramatically reduced but restored after recovery of
491 CTCF function (**Fig. 5**). Intriguingly, the CTCF binding sites in embryonic stem cells at
492 somatic cell identity genes were located within the repressive domains of embryonic
493 stem cells. This colocalization was similar to the colocalization of CTCF binding sites
494 with super-enhancers observed in somatic cell types. These unique CTCF associated
495 epigenetic profiles suggested a novel function of CTCF in epigenetic regulation of
496 transcription.

497

498 Recently, many epigenetic regulators have been proven to interact with CTCF in
499 different biological processes. For instance, BRD2 has been reported to directly interact
500 with CTCF during Th17 cell differentiation⁴⁵. This report suggested that CTCF might be
501 able to regulate enhancer signals by facilitating the binding of enhancer mediator on the
502 chromatin⁴⁶. Interestingly, our result indicated that CTCF played an important role for the
503 repressive histone modification, H3K27me3. This observation is consistent with recent
504 reports that depletion of CTCF does not affect the spreading of H3K27me3⁴⁷, indicating
505 that CTCF might affect H3K27me3 modification by a process other than the spreading.
506 Considering that CTCF was reported to regulate Igf2 expression by direct interaction

507 with Suz12, an important component of Polycomb complexes PRC2⁴⁸, it is possible that
508 CTCF may serve as a landmark to facilitate the localization of epigenetic regulators. Due
509 to limited availability of public datasets for human, we defined genes associated with
510 CSEs and OSEs in human HUVEC and neural cells, and analyzed homolog genes in
511 mouse ESC (mESC) with CTCF ChIP-seq and H3K27me3 ChIP-seq data under normal
512 and CTCF-AID conditions. To further validate our results, we used ChIP-Seq data for
513 CTCF and H3K27ac in three mouse primary samples including heart, embryonic
514 fibroblast and bone marrow macrophage to define genes associated with CSEs and
515 OSEs, next analyzed CTCF and H3K27me3 at these genes in mouse ESC (mESC), and
516 still got a consistent result.

517

518 Interestingly, among the top-ranked feature pairs in H1-hESC, there are many pairs that
519 each was formed by a factor associated with chromatin structure and a factor associated
520 with histone modification for transcription activation or repression. For example, the
521 combination of RBBP5⁴⁹ versus RAD21⁵⁰ and the combination of KDM4A⁵⁰ versus
522 RAD21. RBBP5 and KDM4A are important regulators of H3K4me3, and RAD21 is a
523 component of the cohesion complex that regulates chromatin looping. In addition, we
524 further observed additional combinations that each include a factor associated with
525 activation of transcription and a factor associated with repression of transcription, such
526 as CTBP2⁵¹ and H3K27ac. This kind of combination is consistent with the concept of
527 bivalent domain in the stem cells. Last but not the least, we found high-score
528 combinations that each include a factor of cohesion complex and a factor associated
529 with repression of transcription, such as the combination of CTCF and H3K27me3,
530 which we found later is also very important for the cell identity regulation.

531

532 Taken together, through MACMIC analysis, we find that CTCF plays an important role in
533 the epigenetic regulation of cell identity. Further analysis suggests that CTCF is
534 important for the regulation of both enhancer signals and repressive signals at the CSEs
535 in a cell-type specific manner. Although our analysis focused on the colocalization of
536 enhancer signal with the other chromatin features, MACMIC analysis has great potential
537 to identify many other novel biologically significant colocalizations between chromatin
538 features that have low global correlation across the genome. With the increased usage
539 of sequencing technologies, more potential feature pairs can be identified. This will
540 provide opportunities in the future to further understand the function of chromatin in
541 transcription, replication, DNA repairs and many other biological processes.

542

543 **AVAILABILITY**

544 All public datasets used in this study are listed in Table S1 for ENCODE database and
545 Table S2 for GEO database. The code for MACMIC is available at the website GitHub,
546 <https://github.com/bxia888/MACMIC> . The CSE- and OSE- associated genes can be
547 downloaded from **Table S4**.

548

549 **SUPPLEMENTARY DATA**

550 Figure S1-5.

551 Table S1. Datasets reanalyzed from ENCODE.

552 Table S2. Datasets reanalyzed from GEO.

553 Table S3. Datasets of MACMIC score and p value.

554 Table S4. Datasets of CSE- and OSE- associated genes.

555

556 **AUTHOR CONTRIBUTIONS**

557 K.C. conceived the project, designed the analysis, and interpreted the data. B.X. and

558 G.W., performed the data analysis. K.C. B.X., and G.W. wrote the manuscript with
559 comments from M.Z., J.L., D.Z., Y.L., Y.B., X.W., and L.Z.

560

561 **ACKNOWLEDGEMENT**

562 We would like to thank members of Department of Cardiovascular Sciences at Houston
563 Methodist Research Institute for their thoughtful comments on the project. We further
564 appreciate all researchers who generated the public genomic datasets analyzed in this
565 study.

566

567 **FUNDING**

568 This work was supported in part by NIH grants (R01GM125632 to K.C. R01HL133254
569 and R01HL148338 to J.P.C and K.C., and R01CA207098 and R01CA207109 to M.L.).
570 Q.C. is supported by the U.S. Department of Defense (W81XWH-17-1-0357 and
571 W81XWH-19-1-0563), American Cancer Society (RSG-15-192-01) and NIH/NCI
572 (R01CA208257, P50CA180995 DRP) and Northwestern Univ. Polsky Urologic Cancer
573 Institute.

574

575 **CONFLICT OF INTEREST**

576 The authors have no competing interest that might influence the performance or
577 presentation of the work described in this manuscript.

578

579 REFERENCES

- 580 1 Alizadeh, A. A. *et al.* Distinct types of diffuse large B-cell lymphoma identified by gene
581 expression profiling. *Nature* **403**, 503-511, doi:10.1038/35000501 (2000).
- 582 2 Wang, K. *et al.* Genome-wide identification of post-translational modulators of
583 transcription factor activity in human B cells. *Nat Biotechnol* **27**, 829-839,
584 doi:10.1038/nbt.1563 (2009).
- 585 3 Karlic, R., Chung, H. R., Lasserre, J., Vlahovicek, K. & Vingron, M. Histone modification
586 levels are predictive for gene expression. *Proc Natl Acad Sci U S A* **107**, 2926-2931,
587 doi:10.1073/pnas.0909344107 (2010).
- 588 4 Stillman, B. Histone Modifications: Insights into Their Influence on Gene Expression. *Cell*
589 **175**, 6-9, doi:10.1016/j.cell.2018.08.032 (2018).
- 590 5 COVER, T. M. The best two independent measurements are not the two best. *Ieee T*
591 *Pattern Anal* (1974).
- 592 6 Vandewollenberg, A. L. Redundancy Analysis an Alternative for Canonical Correlation
593 Analysis. *Psychometrika* **42**, 207-219, doi:10.1007/Bf02294050 (1977).
- 594 7 Chen, X. *et al.* Integration of external signaling pathways with the core transcriptional
595 network in embryonic stem cells. *Cell* **133**, 1106-1117, doi:10.1016/j.cell.2008.04.043
596 (2008).
- 597 8 Zhang, L. Q., Li, Q. Z., Su, W. X. & Jin, W. Predicting gene expression level by the
598 transcription factor binding signals in human embryonic stem cells. *Biosystems* **150**, 92-
599 98, doi:10.1016/j.biosystems.2016.08.011 (2016).
- 600 9 Shi, W. Q., Fornes, O. & Wasserman, W. W. Gene expression models based on
601 transcription factor binding events confer insight into functional cis-regulatory variants.
602 *Bioinformatics* **35**, 2610-2617, doi:10.1093/bioinformatics/bty992 (2019).
- 603 10 Wei, G. *et al.* Global Mapping of H3K4me3 and H3K27me3 Reveals Specificity and
604 Plasticity in Lineage Fate Determination of Differentiating CD4(+) T Cells. *Immunity* **30**,
605 155-167, doi:10.1016/j.immuni.2008.12.009 (2009).
- 606 11 Sims, R. J., 3rd, Nishioka, K. & Reinberg, D. Histone lysine methylation: a signature for
607 chromatin function. *Trends Genet* **19**, 629-639, doi:10.1016/j.tig.2003.09.007 (2003).
- 608 12 Liu, X. Y. *et al.* Distinct features of H3K4me3 and H3K27me3 chromatin domains in pre-
609 implantation embryos. *Nature* **537**, 558-+, doi:10.1038/nature19362 (2016).
- 610 13 Vastenhouw, N. L. & Schier, A. F. Bivalent histone modifications in early embryogenesis.
611 *Curr Opin Cell Biol* **24**, 374-386, doi:10.1016/j.ceb.2012.03.009 (2012).
- 612 14 Voigt, P. *et al.* Asymmetrically Modified Nucleosomes. *Cell* **151**, 181-193,
613 doi:10.1016/j.cell.2012.09.002 (2012).
- 614 15 Stanton, B. Z. *et al.* Smarca4 ATPase mutations disrupt direct eviction of PRC1 from
615 chromatin. *Nat Genet* **49**, 282-288, doi:10.1038/ng.3735 (2017).
- 616 16 Peng, H. C., Long, F. H. & Ding, C. Feature selection based on mutual information:
617 Criteria of max-dependency, max-relevance, and min-redundancy. *Ieee T Pattern Anal*
618 **27**, 1226-1238 (2005).
- 619 17 Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome.
620 *Nature* **489**, 57-74, doi:10.1038/nature11247 (2012).
- 621 18 Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res* **12**, 996-1006,
622 doi:10.1101/gr.229102 (2002).
- 623 19 Raney, B. J. *et al.* Track data hubs enable visualization of user-defined genome-wide
624 annotations on the UCSC Genome Browser. *Bioinformatics* **30**, 1003-1005,
625 doi:10.1093/bioinformatics/btt637 (2014).
- 626 20 Cahan, P. *et al.* CellNet: network biology applied to stem cell engineering. *Cell* **158**, 903-
627 915, doi:10.1016/j.cell.2014.07.020 (2014).
- 628 21 Chen, K. *et al.* Broad H3K4me3 is associated with increased transcription elongation and
629 enhancer activity at tumor-suppressor genes. *Nat Genet* **47**, 1149-1157,
630 doi:10.1038/ng.3385 (2015).
- 631 22 Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology
632 Consortium. *Nat Genet* **25**, 25-29, doi:10.1038/75556 (2000).

- 633 23 Chen, K. F. *et al.* Broad H3K4me3 is associated with increased transcription elongation
634 and enhancer activity at tumor-suppressor genes. *Nat Genet* **47**, 1149+,
635 doi:10.1038/ng.3385 (2015).
- 636 24 Hnisz, D. *et al.* Super-Enhancers in the Control of Cell Identity and Disease. *Cell* **155**,
637 934-947, doi:10.1016/j.cell.2013.09.053 (2013).
- 638 25 Phillips, J. E. & Corces, V. G. CTCF: master weaver of the genome. *Cell* **137**, 1194-1211,
639 doi:10.1016/j.cell.2009.06.001 (2009).
- 640 26 Zuin, J. *et al.* Cohesin and CTCF differentially affect chromatin architecture and gene
641 expression in human cells. *Proc Natl Acad Sci U S A* **111**, 996-1001,
642 doi:10.1073/pnas.1317788111 (2014).
- 643 27 You, L. R. *et al.* Suppression of Notch signalling by the COUP-TFII transcription factor
644 regulates vein identity. *Nature* **435**, 98-104, doi:10.1038/nature03511 (2005).
- 645 28 Lujan, E., Chanda, S., Ahlenius, H., Sudhof, T. C. & Wernig, M. Direct conversion of
646 mouse fibroblasts to self-renewing, tripotent neural precursor cells. *Proc Natl Acad Sci U*
647 *S A* **109**, 2527-2532, doi:10.1073/pnas.1121003109 (2012).
- 648 29 Pekowska, A. *et al.* Gain of CTCF-Anchored Chromatin Loops Marks the Exit from Naive
649 Pluripotency. *Cell Syst* **7**, 482-495 e410, doi:10.1016/j.cels.2018.09.003 (2018).
- 650 30 Takahashi, K. & Yamanaka, S. Induction of pluripotent stem cells from mouse embryonic
651 and adult fibroblast cultures by defined factors. *Cell* **126**, 663-676,
652 doi:10.1016/j.cell.2006.07.024 (2006).
- 653 31 Aasen, T. *et al.* Efficient and rapid generation of induced pluripotent stem cells from
654 human keratinocytes. *Nat Biotechnol* **26**, 1276-1284, doi:10.1038/nbt.1503 (2008).
- 655 32 Loh, Y. H. *et al.* Reprogramming of T cells from human peripheral blood. *Cell Stem Cell*
656 **7**, 15-19, doi:10.1016/j.stem.2010.06.004 (2010).
- 657 33 Kim, J. B. *et al.* Oct4-induced pluripotency in adult neural stem cells. *Cell* **136**, 411-419,
658 doi:10.1016/j.cell.2009.01.023 (2009).
- 659 34 Benayoun, B. A. *et al.* H3K4me3 Breadth Is Linked to Cell Identity and Transcriptional
660 Consistency. *Cell* **158**, 673-688, doi:10.1016/j.cell.2014.06.027 (2014).
- 661 35 Chen, K. *et al.* Broad H3K4me3 is associated with increased transcription elongation and
662 enhancer activity at tumor-suppressor genes. *Nat Genet*, doi:10.1038/ng.3385 (2015).
- 663 36 Whyte, W. A. *et al.* Master Transcription Factors and Mediator Establish Super-
664 Enhancers at Key Cell Identity Genes. *Cell* **153**, 307-319, doi:10.1016/j.cell.2013.03.035
665 (2013).
- 666 37 Bell, A. C., West, A. G. & Felsenfeld, G. The protein CTCF is required for the enhancer
667 blocking activity of vertebrate insulators. *Cell* **98**, 387-396 (1999).
- 668 38 Guo, Y. *et al.* CRISPR Inversion of CTCF Sites Alters Genome Topology and
669 Enhancer/Promoter Function. *Cell* **162**, 900-910, doi:10.1016/j.cell.2015.07.038 (2015).
- 670 39 Haarhuis, J. H. I. *et al.* The Cohesin Release Factor WAPL Restricts Chromatin Loop
671 Extension. *Cell* **169**, 693-707, doi:10.1016/j.cell.2017.04.013 (2017).
- 672 40 Sanborn, A. L. *et al.* Chromatin extrusion explains key features of loop and domain
673 formation in wild-type and engineered genomes. *Proc Natl Acad Sci U S A* **112**, E6456-
674 E6465, doi:10.1073/pnas.1518552112 (2015).
- 675 41 Heath, H. *et al.* CTCF regulates cell cycle progression of alphabeta T cells in the thymus.
676 *EMBO J* **27**, 2839-2850, doi:10.1038/emboj.2008.214 (2008).
- 677 42 Watson, L. A. *et al.* Dual effect of CTCF loss on neuroprogenitor differentiation and
678 survival. *J Neurosci* **34**, 2860-2870, doi:10.1523/JNEUROSCI.3769-13.2014 (2014).
- 679 43 Gomez-Velazquez, M. *et al.* CTCF counter-regulates cardiomyocyte development and
680 maturation programs in the embryonic heart. *PLoS Genet* **13**, e1006985,
681 doi:10.1371/journal.pgen.1006985 (2017).
- 682 44 Soshnikova, N., Montavon, T., Leleu, M., Galjart, N. & Duboule, D. Functional analysis of
683 CTCF during mammalian limb development. *Dev Cell* **19**, 819-830,
684 doi:10.1016/j.devcel.2010.11.009 (2010).
- 685 45 Cheung, K. L. *et al.* Distinct Roles of Brd2 and Brd4 in Potentiating the Transcriptional
686 Program for Th17 Cell Differentiation. *Mol Cell* **65**, 1068-1080 e1065,
687 doi:10.1016/j.molcel.2016.12.022 (2017).

688 46 Ren, G. *et al.* CTCF-Mediated Enhancer-Promoter Interaction Is a Critical Regulator of
689 Cell-to-Cell Variation of Gene Expression. *Mol Cell* **67**, 1049-1058 e1046,
690 doi:10.1016/j.molcel.2017.08.026 (2017).
691 47 Nora, E. P. *et al.* Targeted Degradation of CTCF Decouples Local Insulation of
692 Chromosome Domains from Genomic Compartmentalization. *Cell* **169**, 930-944 e922,
693 doi:10.1016/j.cell.2017.05.004 (2017).
694 48 Li, T. *et al.* CTCF regulates allelic expression of Igf2 by orchestrating a promoter-
695 polycomb repressive complex 2 intrachromosomal loop. *Mol Cell Biol* **28**, 6473-6482,
696 doi:10.1128/MCB.00204-08 (2008).
697 49 Yokoyama, A. *et al.* Leukemia proto-oncoprotein MLL forms a SET1-like histone
698 methyltransferase complex with menin to regulate Hox gene expression. *Mol Cell Biol* **24**,
699 5639-5649, doi:10.1128/Mcb.24.13.5639-5649.2004 (2004).
700 50 Zuin, J. *et al.* Cohesin and CTCF differentially affect chromatin architecture and gene
701 expression in human cells. *P Natl Acad Sci USA* **111**, 996-1001,
702 doi:10.1073/pnas.1317788111 (2014).
703 51 Turner, J. & Crossley, M. The CtBP family: enigmatic and enzymatic transcriptional co-
704 repressors. *Bioessays* **23**, 683-690, doi:DOI 10.1002/bies.1097 (2001).
705

706

Figure 1

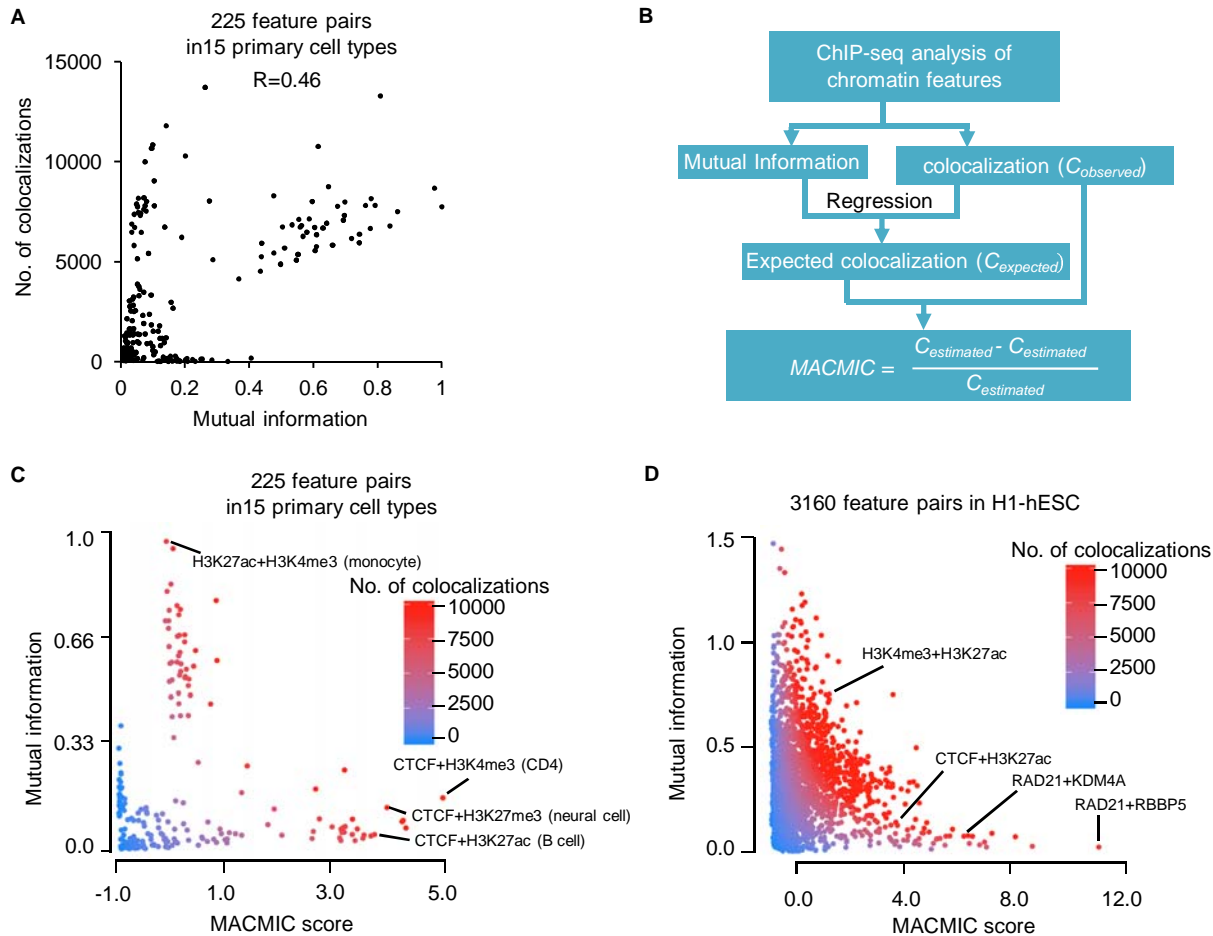


Figure 1. The MACMIC method to define mutual information redundancy of colocalizations between genomic features. (A) Scatter plot to show mutual information value and the number of colocalization for each of 225 feature pairs derived from 6 features that form 15 combinations with each other in each of 15 human primary cell types. (B) The workflow to calculate the MACMIC score. (C) Scatter plot to show MACMIC score and mutual information value for each of 225 feature pairs derived from 6 features that form 15 combinations with each other in each of 15 human primary cell types. Color scale indicates the number of colocalizations between each pair of features. (D) Scatter plot to show MACMIC score and mutual information value between each pair of features. 3160 feature pairs derived from 80 features in H1-hESC were plotted. Color scale indicates the number of colocalizations between each pair of features.

Figure 2

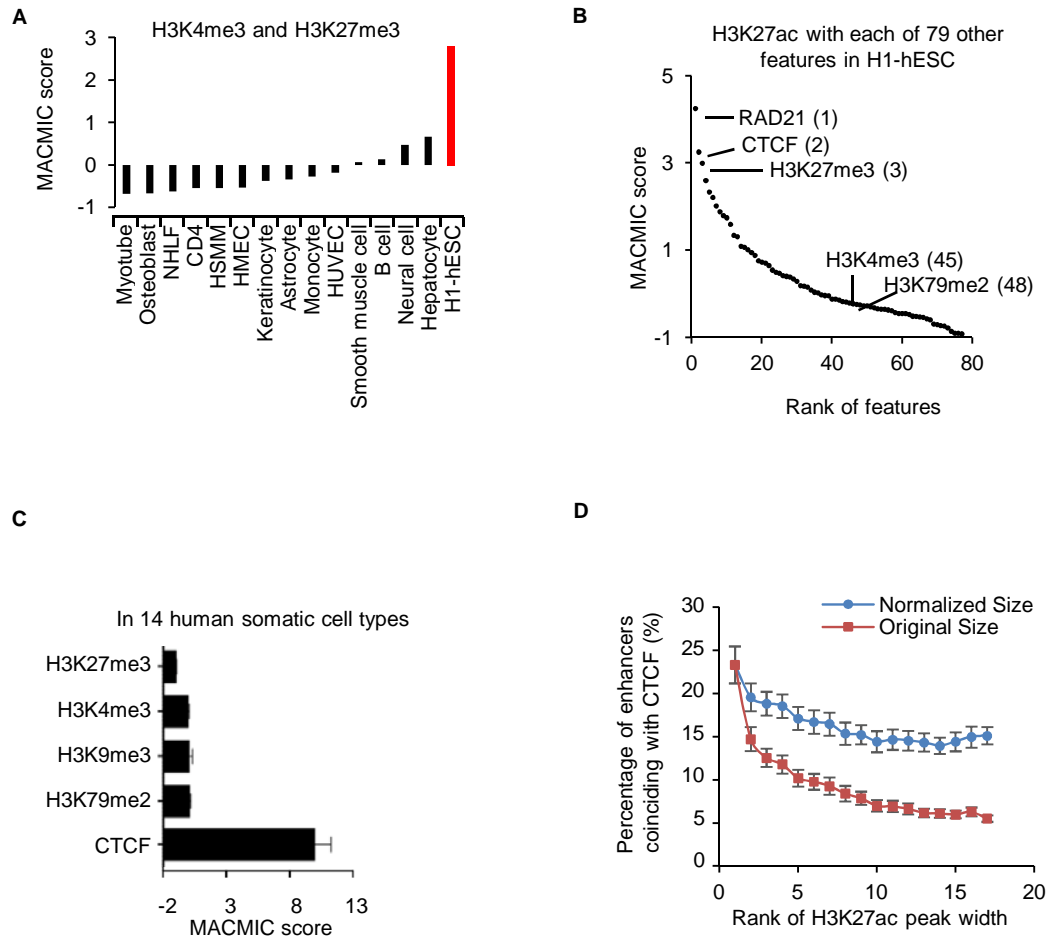


Figure 2. MACMIC reveals minimal information redundancy of frequent colocalizations between CTCF binding sites and super-enhancers. (A) Bar plot to show MACMIC scores between H3K4me3 and H3K27me3 in individual human primary cell types. (B) MACMIC scores between H3K27ac and individual other chromatin features in H1-hESC. (C) MACMIC scores between H3K27ac and individual other chromatin features in 14 human somatic primary cell types. Error bars indicate the standard deviation of MACMIC scores across cell types. (D) Percentage of enhancers that coincided with CTCF binding sites in 15 human primary cell types. Enhancers were divided into individual groups on the base of their H3K27ac width. Each group contains 500 enhancers, e.g. rank 1 contains the widest 500 enhancers; rank 2 contains the 501st to 1000th widest enhancers.

Figure 3. CTCF-associated super-enhancers mark cell identity genes. (A-B) Individual pathways enriched in CSEs or OSEs genes in HUVECs (A) or neural cells (B). (C-D) ChIP-Seq signals for H3K27ac and CTCF at CSE gene NR2F2 and OSE gene ARF1 in HUVECs (C) and CSE gene FOXG1 and OSE gene PON1 in neural cells (D). (E-F) The number of transcription factors within each gene group (E) and the number of network edges within each gene group (F) in 15 human primary cell types. Error bars indicate the standard deviation across cell types. Each gene group was defined to have the same number of genes. P values were determined by Wilcoxon test in comparison to the control group. (G) Heatmap to show $-\log_{10}$ enrichment P values of cell type related pathways (rows) in CSEs genes (top panel) or OSEs genes (bottom panel) defined in each cell type (columns).

Figure 4

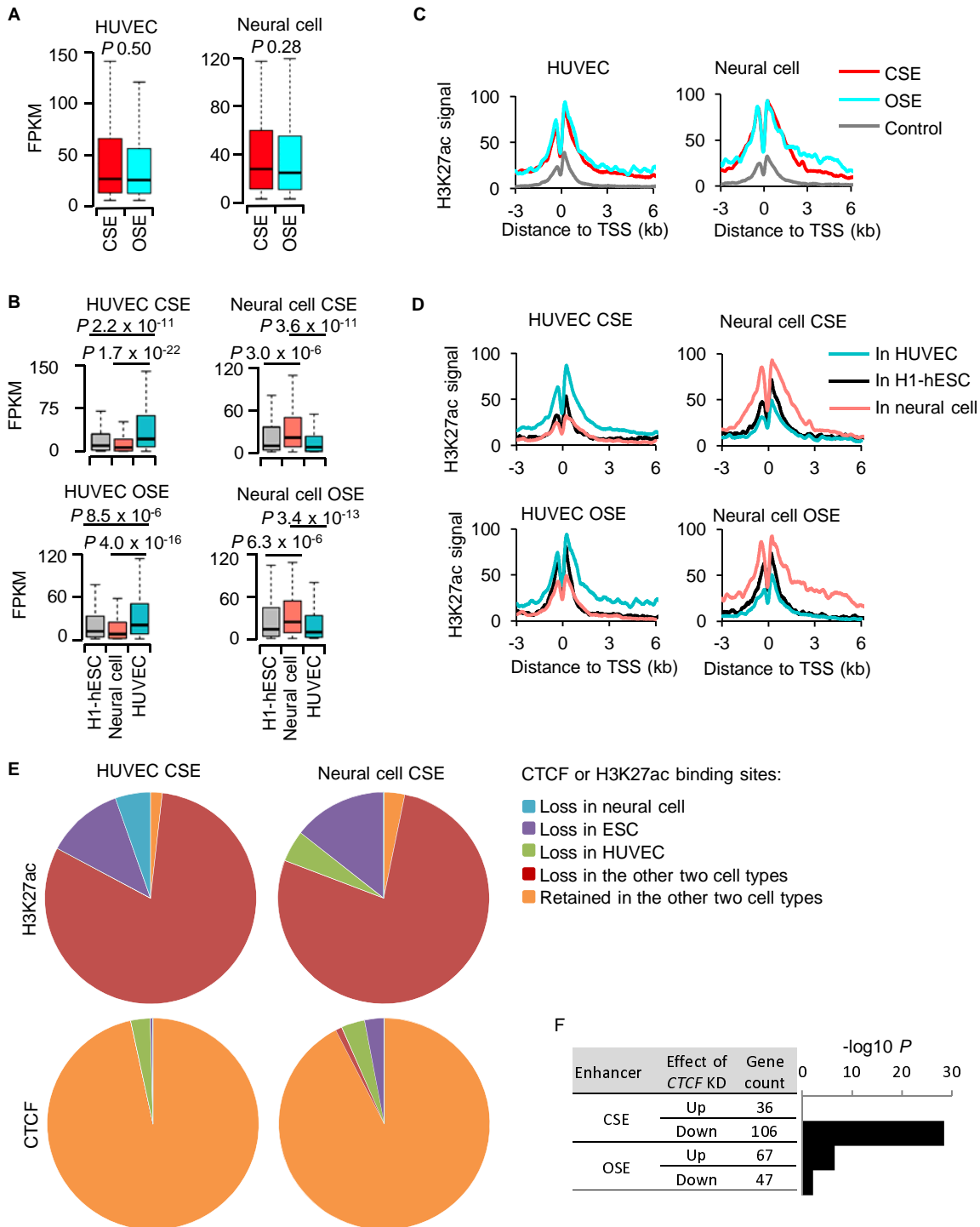


Figure 4. CTCF is linked to the activation of enhancers. (A) Box plot to show RNA expressions of HUVEC (left panel) and neural cell (right panel) CSEs genes and OSEs genes in cell types that defined them. (B) Box plot to show RNA expressions of CSE genes (top panels) and OSE genes (bottom panels) in neural cell, HUVEC, and H1-hESC. CSEs and OSEs were defined in HUVEC (left panels) or neural cells (right panels) (C) H3K27ac signals at HUVEC (left panel) and neural cell (right panel) CSEs genes, OSEs genes, and control genes in the cell type that defined these gene groups. (D) H3K27ac signals at CSE genes (top panels) and OSE genes (bottom panels) in HUVEC, H1-hESC, and neural cells. CSEs and OSEs were defined in HUVEC (left panels) or neural cells (right panels). (E) Pie charts to show binding status of CTCF at HUVEC CSEs in neural cells or embryonic stem cells (top left), binding status of CTCF at neural cell CSEs in HUVEC or embryonic stem cells (top right), H3K27ac status at HUVEC CSEs in neural cells or embryonic stem cells (bottom left), and H3K27ac status at neural cell CSEs in HUVEC or embryonic stem cells (bottom right). (F) Barplot to show $-\log_{10}$ enrichment P values of CSE genes or OSE genes in the genes up or down regulated by shCTCF in HeLa cells. P values were determined by Wilcoxon test (A, B, E, F).

Figure 5

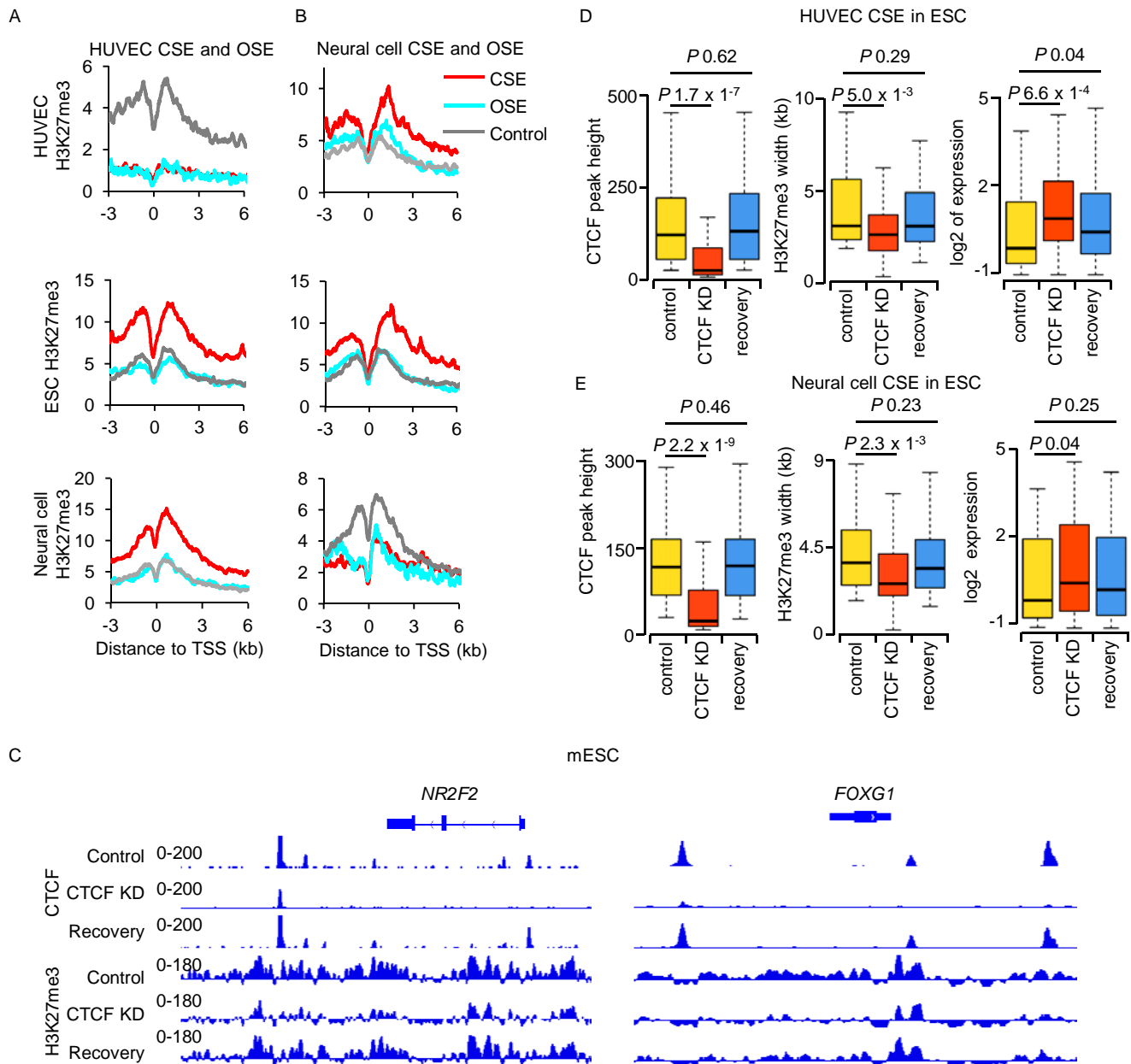


Figure 5. CTCF regulates cell identity by facilitating the suppressive marker H3K27me3.

(A-B) H3K27me3 signals in H1-hESC, neural cell and HUVECs at CSE genes, OSE genes and control genes defined in HUVECs (A) and neural cells (B). (C) ChIP-Seq signals for CTCF and H3K27me3 in mESC at the HUVEC CSE gene *NR2F2* (left) and the neural CSE gene *FOXG1* (right). (D-E) Box plot to show the heights of CTCF ChIP-Seq enrichment peaks, the widths of H3K27me3 enrichment domains, and the RNA expressions of CSE genes of HUVECs (D) and CSE genes of neural cells (E) under different conditions in embryonic stem cells. P values were determined by Wilcoxon test.