3D Chromatin Architecture Remodeling during Human Cardiomyocyte Differentiation Reveals A Novel Role Of HERV-H In Demarcating Chromatin Domains

Authors:

Yanxiao Zhang[1*], Ting Li[2*], Sebastian Preissl[1,3*], Jonathan Grinstein[2], Elie N. Farah[2], Eugin Destici[2], Ah Young Lee[1], Sora Chee[1], Yunjiang Qiu[1], Kaiyue Ma[1], Zhen Ye[1], Quan Zhu[1], Hui Huang[1], Rong Hu[1], Rongxin Fang[1], Sylvia M. Evans[2,4], Neil C. Chi[2,5,7], Bing Ren[1,3,5,6,7]

*these authors contributed equally

Affiliations:

[1]Ludwig Institute for Cancer Research, La Jolla, CA 92093, USA.

[2] Department of Medicine, Division of Cardiology, University of California, San Diego, La Jolla, California 92093, USA.

[3] Department of Cellular and Molecular Medicine, Center for Epigenomics, University of California, San Diego, School of Medicine, La Jolla, CA 92093, USA.

[4]Department of Pharmacology and Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego, La Jolla, CA 92093, USA.

[5]School of Medicine, Institute of Genomic Medicine, University of California, San Diego, La Jolla, CA 92093, USA

[6]School of Medicine, Moores Cancer Center, University of California, San Diego, La Jolla, CA 92093, USA.

[7]Correspondence: Neil Chi (nchi@ucsd.edu) and Bing Ren (biren@ucsd.edu)

Abstract/Summary

Restructuring of chromatin architecture is an essential process for establishing cell type-specific gene regulatory programs in eukaryotic cells including cardiomyocytes[1-3]. Supporting its importance, recent studies have reported that a substantial number of mutations discovered in congenital heart disease (CHD) patients reside in genes encoding chromatin remodeling factors[4-6]; yet, how chromatin structure reorganizes to assemble gene regulatory networks crucial for controlling human cardiomyocyte development remains to be elucidated. Here, through comprehensively analyzing high-resolution genomic maps that detail the dynamic changes of chromatin architecture, chromatin accessibility and modifications, and gene expression during human pluripotent stem cell (PSC) cardiomyocyte differentiation, we reveal novel molecular insights into how human PSC chromatin architecture is iteratively remodeled to build gene regulatory networks directing cardiac lineage specification. Specifically, we uncover a new class of human PSC-specific topologically associating domain (TAD) that is created by the active transcription of the primate-specific endogenous retrotransposon HERV-H. Silencing of specific HERV-Hs during the initial stages of human PSC differentiation or by genome-editing results in the elimination of corresponding TAD boundaries and reduced transcription of genes upstream of HERV-Hs. Supporting their role in maintaining pluripotency, we discovered that deletion of specific HERV-Hs leads to more efficient human PSC cardiomyocyte differentiation. Using chromatin interaction maps from these analyses, we also assigned potential target genes to distal regulatory elements involved in cardiac differentiation. Genome-editing of enhancers harboring cardiac-disease risk loci associated with congenital and adult heart diseases further confirmed that these loci regulate predicted target genes. Our results highlight a novel role for HERV-Hs in establishing human-specific PSC chromatin architecture, delineate the dynamic gene regulatory networks during cardiomyocyte development and inform how non-coding genetic variants contribute to congenital and adult heart diseases.

MAIN TEXT

The three-dimensional organization of chromosomes enables long-range communications between enhancers and promoters that are critical for building complex gene regulatory networks in multicellular species[1,3]. In somatic cells, interphase chromosomes occupy separate nuclear spaces known as chromosome territories[7]. Each chromosome is folded into a dynamic but non-random hierarchical structure characterized by stretches of transcriptionally active, megabase-long compartments that are interspersed with stretches of transcriptionally inactive compartments[8]. These compartments can be further partitioned into topologically associating domains (TADs), which exhibit high levels of intra-domain interactions and relatively low levels of inter-domain interactions[9-11]. The boundaries of TADs are generally conserved between cell types and closely related species[9,12], and are necessary to insulate enhancer/promoter interactions[13,14]. However, recent studies have shown that chromatin architecture is reconfigured in a cell-type specific manner during stem cell differentiation and somatic cell reprogramming[12,15-18].

Chromatin remodelers and transcription factors compose a large portion of genes disrupted in CHD[4-6], one of the leading causes of birth defects affecting 1-3% of newborns[19]. Thus, we investigated how chromatin architecture is dynamically remodeled to coordinate the silencing and activation of key transcriptional programs during differentiation of human PSCs to cardiomyocytes. Utilizing a transgenic human embryonic stem cell (hESC) H9 *MYL2:H2B-GFP* ventricular cardiomyocyte reporter line, which can be used to purify ventricular cardiomyocytes[20] (**Extended Data Fig. 1a**), we generated and collected samples in biological duplicates at the following six critical time points of cardiomyocyte differentiation: hESCs (Day 0), mesodermal cells (Day 2), cardiac mesodermal cells (Day 5), cardiac progenitors (Day 7), primitive cardiomyocytes (Day 15) and ventricular cardiomyocytes (Day 80) (**Fig. 1a**). Flow cytometry confirmed that differentiation efficiency for D2 – D15 time point samples was at least ~80-90%, and *MYL2:H2B-GFP*+ ventricular cardiomyocytes were sorted at day 80 (**Extended Data Fig. 1b**). *In situ* Hi-C[8] experiments were performed, and on average three billion raw read pairs and one billion unique long-range cis contacts (minimum distance 10 kb,

88 **Extended Data Table 1**) were generated for each time point. Close correlation of Hi-C
89 maps between biological replicates verified that these maps were highly reproducible at
90 multiple scales (**Extended Data Fig. 2)**. Complementing these Hi-C contact maps,
91 ATAC-seq, ChIP-seq and RNA-seq experiments were performed on these samples to
92 interrogate chromatin accessibility, locate CTCF occupancy and key histone
93 modifications, and profile transcriptomes, respectively (**Extended Data Fig. 3**).
94
95 Analyses of these genome-wide datasets revealed extensive reorganization of
96 chromatin architecture at all structural levels during hESC differentiation, which was
97 also accompanied by corresponding epigenomic changes (**Fig. 1a, b, c**). Across the
98 entire genome, the abundance of short-range chromatin interactions decreased while
99 the number of long-range chromatin interactions (above 500 kilobases) increased
100 (**Extended Data Fig. 4a**). Identification of active (A) and inactive (B) compartments at
101 all stages using principal component analysis showed that 81.6% of the genome
102 maintained the same compartment state throughout cardiomyocyte differentiation (**Fig.**
103 **1d**). Among the 18.4% of the genome that switched compartments, a similar proportion
104 of the genome switched from A to B and from B to A at each stage transition (**Extended**
105 **Data Fig. 4b**), and the majority (73%) only switched once, suggesting linear nuclear
106 architectural changes during lineage commitment (**Fig. 1d**). Hierarchical clustering of
107 genomic bins that switched compartments showed stage specific dynamics (**Fig. 1d**),
108 which positively correlated with gene expression (**Extended Data Fig. 4c**) as reported
109 previously[12]. For instance, gene loci of pluripotency factors including *SOX2* moved to
110 the inactive compartment when hESCs differentiated into mesodermal cells, whereas
111 genes involved in cardiomyocyte specification such as the transcription factor *HAND2*
112 and the cardiomyocyte-specific calcium channel *RYR2* displayed B to A transitions at
113 day 5 and day 15, respectively (**Fig. 1d; Extended Data Fig. 4d**).
114
115 Consistent with recent reports describing loss of TADs during the differentiation of
116 embryonic stem cells[15,21], we also observed that the number of TADs decreased as
117 hESCs differentiated into ventricular cardiomyocytes using multiple TAD-calling
118 algorithms[9,13,22] (**Extended Data Fig. 5a, b, c**). Clustering of stage-specific TAD

119 boundaries using Directionality Index (DI) delta scores, which estimates the strength of
120 the TAD boundaries, further showed that the majority of stage-specific TAD boundaries
121 were lost or weakened during differentiation and the effect was most pronounced at
122 early mesoderm specification (D0-D2), and ventricular cardiomyocyte
123 differentiation/maturation (D15-D80) (**Fig. 1e**). To investigate the mechanism of TAD
124 dynamics during cardiomyocyte differentiation, we examined the sequence features of
125 TAD boundaries that were lost in the course of differentiation with a particular focus on
126 D0 hESC-specific TAD boundaries [ESC(+)] and the boundaries lost during D15-D80
127 ventricular cardiomyocyte transition [vCM(-)] (**Extended Data Table 2**). While CTCF
128 has been shown to play a critical role in defining TAD boundaries[9,13,23], we observed no
129 significant difference in CTCF occupancy across all stages examined at ESC(+), vCM(-)
130 and stable TAD boundaries (**Extended Data Fig. 5d**). Thus, we explored whether other
131 potential mechanisms might be involved in the formation of these TAD boundaries.
132 Interestingly, we discovered that eight classes of repeat elements were over-
133 represented specifically at ESC(+) TAD boundaries when compared to stable TAD
134 boundaries (p-value < 0.01 and fold change > 2), whereas no enrichment of these
135 repeats was observed in vCM(-) TAD boundaries (**Extended Data Fig. 6a**).
136 Additionally, the ESC(+) TAD boundaries were associated with hESC-specific H3K27ac
137 signal and mRNA transcription (**Extended Data Fig. 5e, f**), which was particularly
138 enriched for HERV-H transcription (**Fig. 2a and Extended Data Fig. 6b**).
139
140 HERV-H is a class of primate-specific endogenous retrotransposons. It is
141 transcriptionally active in both human preimplantation embryos and pluripotent stem
142 cells[24,25], and plays a critical role in pluripotency, stem cell maintenance and somatic
143 cell reprograming[26-30]. Among the ~1200 HERV-H sequences in the human genome,
144 only a small number are transcriptionally active in hESCs (**Fig. 2b)**[26]. We found that the
145 top 50 highly transcribed HERV-H loci were predominantly located at TAD boundaries in
146 hESCs (**Fig. 2c, Extended Data. Fig 6c)**. However, transcriptional silencing of these
147 HERV-Hs occurred promptly by day 2 in mesodermal cells and coincided with the loss
148 of the corresponding TAD boundaries (**Fig. 2d, e, f and Extended Data Fig 6d, e, f**).
149 As controls, genes with similar transcriptional levels to these top 50 HERV-H loci did not

150 show any boundary activity as evidenced by the DI scores (**Extended Data Fig. 7a**).
151 This correlation was not specific to only the H9 hESC line used in this study, as we also
152 observed HERV-H associated TAD boundaries in other hESCs (H1)[12,31] and human
153 induced PSCs[32] (**Extended Data Fig. 7b**). Consistent with these findings, we notably
154 discovered that in human mesendoderm cells (an early human embryonic state that
155 gives rise to mesoderm and endoderm cells), both the expression of these HERV-Hs
156 and their corresponding TAD boundary strength were approximately half the levels
157 compared to those observed in hESCs (**Extended Data Fig. 7b**).
158
159 Because HERV-Hs might contribute to TAD boundary formation by recruiting DNA
160 binding proteins, we examined publicly available ChIP-seq data of histone modifications
161 and transcription factors (TFs) in hESCs. We identified several modifications and
162 factors enriched at sites for the top 50 transcribed HERV-Hs when compared to those
163 for the lower-ranking HERV-Hs. They included POL2 transcriptional machinery
164 (POLR2A and POLR2AphosphoS5), active histone marks (H3K4me3, H3K4me2 and
165 H3K27ac), ubiquitously active TFs (JUND and SP1), pluripotency factor NANOG[28] and
166 all three subunits of the cohesin complex (**Extended Data Fig. 7c**). To further
167 investigate the role of cohesin in HERV-H-mediated TAD boundary formation, we
168 performed ChIP-seq in hESCs for the cohesin complex subunit SMC3. Analyzing this
169 ChIP-seq data along with RNA-seq as well as CTCF and POLR2A ChIP-seq data,
170 revealed broad enrichment of cohesin, CTCF and POLR2A specifically at the 3' end of
171 these highly transcribed HERV-H sequences (**Fig. 2g**). In contrast to these findings, the
172 enrichment of these factors was not similarly observed at sites of lower ranked HERV-
173 Hs (**Fig. 2g**), suggesting that these factors may be involved in mediating the formation
174 of TAD boundaries by HERV-Hs. However, we cannot completely rule out that CTCF
175 may also participate in defining these HERV-H-associated TAD boundaries as weak but
176 reproducible enrichment of CTCF occupancy was observed in these regions despite the
177 absence of CTCF binding motifs (**Fig. 2g**). Overall, these data support that HERV-Hs
178 form TAD boundaries by accumulating cohesin complexes at its 3' end, and cohesin is
179 likely positioned by the transcribing POL2 complex[33,34].
180

181 To determine whether HERV-H sequences are indeed required for TAD boundary
182 formation, we deleted two HERV-H elements located at different hESC TAD
183 boundaries, using CRISPR-Cas9 genome editing tools (termed as HERV-H1-KO and
184 HERV-H2-KO, respectively). In both cases, deletion of these HERV-Hs resulted in
185 elimination of TAD boundaries, as evidenced by the merging of the two TADs bordering
186 each HERV-H element (**Fig. 3a**), and reduced expression of genes in the TAD domain
187 immediately 5' upstream of the HERV-H sequence (**Fig. 3a**). These findings thus reveal
188 that HERV-H sequences are indispensable for the TAD boundaries formed at these
189 regions and furthermore the presence of HERV-H increases the overall transcriptional
190 activity of the neighboring genes in the same TAD domain, supporting that HERV-Hs
191 may serve as enhancers in hESCs[35,36]. Notably, this activating effect is only seen in the
192 TAD domain on the 5' end of HERV-H and not the 3' end, suggesting that boundaries
193 present on the 3' end of HERV-Hs, may insulate the effect of enhancers. Supporting
194 this possibility, additional analyses of all genes located within 500 kb of boundary-
195 associated HERV-Hs showed that only the genes in regions 5' of HERV-H were up-
196 regulated in hESCs compared to differentiated cells (**Fig. 3b**).
197
198 Further examination of the transcriptional profiles of both HERV-H knockouts revealed
199 that expression of 353 and 144 genes were significantly changed (**Extended Data Fig.
200 8a, b, Extended Data Table 3**). The fold changes of HERV-H1-KO and HERV-H2-KO
201 were highly concordant (**Extended Data Fig. 8c**, Pearson r=0.5, p-value< 1e-15). Out
202 of the 74 genes that were significantly altered in these knockouts, 69 genes (92%) were
203 altered in the same direction with similar fold changes (**Extended Data Fig. 8c**). Of the
204 43 genes down-regulated in both, 10 (23%) were located within 20 kb of other HERV-H
205 sequences (**Extended Data Fig. 8d**), including two well studied chimeric transcripts
206 *SCGB3A2*[37] and *LINC00458*[38], which are known to be regulated by HERV-Hs
207 (**Extended Data Fig. 8e, f**). Thus, these data showed that deletion of individual HERV-
208 H sequences could lead to down-regulation of HERV-H related transcription at other
209 loci. In addition to these findings, we also discovered that the human-specific long-
210 coding RNA (lncRNA), *Heart Brake* (*HBL1*), which suppresses cardiomyocyte
211 differentiation[39], was also significantly reduced in the HERV-H1-KO (**Extended Data**

212 **Fig. 8f)**. To further explore whether this decrease in *HBL1* lncRNA in the HERV-H1-KO
213 promotes cardiomyocyte differentiation, we examined the cardiomyocyte differentiation
214 efficiency of HERV-H1-KO hESCs compared to control hESCs under suboptimal
215 cardiomyocyte differentiation conditions. While these conditions led to ~40%
216 cardiomyocyte differentiation efficiency at day 15 in control hESCs, we discovered that
217 HERV-H1-KO hESCs exhibited >80% cardiomyocyte differentiation efficiency (**Fig. 3c**),
218 which is consistent with the increased cardiomyocyte differentiation efficiency observed
219 in *HBL1* knockout hPSCs[39]. Overall, these findings suggest that HERV-Hs may globally
220 maintain pluripotency in hPSCs through a common regulatory network; however,
221 individual HERV-Hs may also regulate specific genes involved in the differentiation of
222 distinct lineages, such as *HBL1* and the cardiac lineage.
223
224 Because HERV-H sequences are primate specific, we predicted that the TAD
225 boundaries established by HERV-Hs would be absent in ESCs from non-primate
226 mammals. To test this prediction, we examined Hi-C profiles from mouse ESC and
227 observed no TAD boundaries within regions syntenic with the HERV-H-associated
228 human TAD boundaries (**Fig. 3d, e**). Next, we calculated the fraction of the human TAD
229 boundaries that overlapped or were within 25 kb of TAD boundaries in mouse ESCs
230 and discovered that HERV-H-associated TAD boundaries were less likely to be present
231 in mouse ESC than other types of TAD boundaries (**Extended Data Fig. 9a**). This
232 result is consistent with the model that HERV-H transcription is involved in forming ESC
233 specific TAD boundaries in the primate lineages. It is estimated that HERV-H integrated
234 into the genome 30–40 million years ago (MYA) at the time of divergence of Old and
235 New World monkeys[40,41], with the largest expansion (LTR7 and LTR7B) occurring at
236 about 25 MYA[40], and a second expansion (LTR7Y) occurring between 10 and 17
237 MYA[40]. We found that the top 50 most highly expressed HERV-Hs, which have the
238 potential to form TAD boundaries, were exclusively flanked by LTR7 at both ends
239 (**Extended Data Fig. 9b**). In addition, these LTRs belonged predominantly to a 450 bp
240 LTR7 subtype (**Extended Data Fig. 9c**). Finally, these TAD-boundary HERV-Hs
241 showed less sequence divergence between their 5' LTRs and 3' LTRs (**Extended Data**
242 **Fig. 9d**), suggesting that they likely were inserted more recently than other HERV-Hs

243 that do not form TAD boundaries, or were impacted by gene conversion or selection[42].
244
245 From our extensive interrogation of the genome during hESC cardiomyocyte
246 differentiation, we also identified 14,216 chromatin loops spanning a median genomic
247 distance of 180 kb at 10 kb resolution (**Extended Data Table 4**). More than 70% of the
248 loops corresponded to either promoter-to-promoter or promoter-to-enhancer interactions
249 (**Extended Data Fig. 10a**). Consistent with chromatin loops detected in different cell
250 types or model systems[13,14], the loop anchors were enriched for accessible chromatin,
251 histone marks, transcriptional start sites and CTCF binding sites (**Extended Data Fig.**
252 **10b**). More importantly, 23% (3,251) of these chromatin loops exhibited significant
253 changes in interaction strength during cardiomyocyte differentiation, which illustrates a
254 dramatic reconfiguration of the gene regulatory network (**Fig. 4a, Extended Data Fig.**
255 **10c**). Gene transcription and chromatin marks at anchors of stage-specific chromatin
256 loops were correlated to a significantly higher degree than that of anchors of static
257 chromatin loops and distance-matched control (**Fig. 4e**). Using K-means clustering, we
258 defined five distinct clusters of stage-specific chromatin loops (**Fig. 4b**), each
259 accompanied by corresponding epigenomic changes (**Extended Data Figure 10d**).
260 Remarkably, the gene sets enriched for each chromatin loop cluster recapitulated the
261 developmental processes from early differentiation, to heart development, and
262 culminated in cardiac muscle development (**Fig. 4c**). Chromatin loops gained during
263 cardiomyocyte differentiation included promoters of many well studied heart
264 transcription factors, such as *GATA6*, *TBX5*, *NKX2-5*, *TBX3* and *MEF2A*[43]. Analysis of
265 ATAC-seq peaks at loop anchors also revealed distinct patterns of transcription factor
266 binding throughout differentiation (**Fig. 4d**). For instance, enrichment for GATA motifs
267 emerged at day 2, and T-BOX motifs at day 5, whereas MEF2 motifs were only highly
268 enriched at day 80. These findings suggest a process involving pioneering factor
269 activity of GATA family members such as GATA4 for the initial rewiring of chromatin
270 loops to activate relevant early cardiac developmental gene programs, whereas MEF
271 family members such as MEF2C may play a role in the maturation of cardiomyocytes.
272 Finally, loops formed at later stages tended to span longer genomic distances, to have a
273 higher percentage of convergent CTCF-CTCF motifs, and were more likely to occur at

274 TAD boundaries (**Extended Data Fig. 10e, f, g**). These observations further support
275 the global trend of increased long-distance interactions during cardiomyocyte
276 differentiation (**Extended Data Fig. 4a**).
277
278 Integrative analysis of the chromatin architecture and ChIP-seq data revealed two major
279 types of loops: 1) loops enriched for the polycomb mark H3K27me3 (**Extended Data**
280 **Fig 10d, 11a**); and 2) loops enriched for H3K27ac (**Extended Data Fig 10d, 11b**).
281 H3K27me3-associated, polycomb dependent loop interactions, which have been
282 described in *Drosophila* and mouse ESC[15,44,45], were prominent in hESCs, spanned
283 very long genomic distances and were abruptly lost at the mesoderm stage (**Extended**
284 **Data Fig 10d, 11a**). Notably, such loops frequently involved promoters of transcription
285 factors, which were not expressed at the hESC stage (Fold enrichment=9.7, adjusted p-
286 value=1e-70, hypergeometric test), including the early cardiac transcription factor
287 *HAND2*. Further analyses of the *HAND2* locus revealed a multistep assembly of loops
288 formed between the *HAND2* promoter and several distal elements (**Fig. 1b**). In hESCs
289 and mesodermal cells, which do not express *HAND2*, the *HAND2* promoter formed an
290 H3K27me3 associated loop with the promoter region of the lncRNA *LINC02268*.
291 However, by day 5, this loop was eliminated and two new loop interactions were
292 established between the *HAND2* promoter and two distal regions marked by H3K27ac
293 and H3K4me1. This chromatin remodeling coincided with the onset of *HAND2*
294 expression in cardiac mesodermal cells (D5) and a subsequent 4-fold increase in
295 *HAND2* expression in primitive cardiomyocytes (D15) and ventricular cardiomyocytes
296 (D80); however, lncRNA *LINC02268* remained transcriptionally repressed (**Fig. 1b, c**).
297 Based on this multi-loop chromatin remodeling at the *HAND2* locus, we further
298 examined whether multi-loop formation was a general feature to tightly control
299 expression levels of transcription factors whose function may impact cell fate more than
300 that of other genes. To this end, we calculated the degree of interaction for each loop
301 anchor (**Fig. 4f**), and observed that 45% of the loop anchors interacted with at least two
302 other anchors (**Fig. 4g**), suggesting that chromatin loops frequently form network hubs.
303 In addition to *HAND2*, many other well-known cardiac TFs, including *TBX5*, *GATA4*,
304 *TBX20*, *MEIS1/2*, also interacted with four or more *cis*-elements and formed similar

network hubs (**Extended Data Fig 11c, d, e,** and **Extended Data Table. 5**). Supporting these findings, the network hubs with more chromatin interactions were more likely to include transcription start sites (TSSs) of TFs (**Fig. 4h**), and less likely to involve TSSs of house-keeping genes (**Fig. 4i**). Finally, those network hubs with more interactions frequently contained more CTCF peaks (**Extended Data. Fig. 11f**), suggesting that these interactions were likely facilitated by CTCF.

We next used the chromatin interaction maps to link non-coding genetic variants from genome-wide association studies (GWAS) to potential target genes[46-49], and found a strong enrichment of single nucleotide polymorphisms (SNPs) associated with cardiac traits in the chromatin loop anchors (**Fig. 5a**). The list of predicted candidate genes for these SNPs predominantly included genes involved in cardiac function (**Table 1** and **Extended Data Table 6**) and cardiac ion channel genes (**Fig. 5b**). One notable cardiac GWAS example was rs6781009, which was found in strong linkage disequilibrium (LD) with lead SNPs of four cardiac electrophysiologic GWAS traits (QRS complex, QRS duration, QT interval and electrocardiographic traits). Using our chromatin interaction maps, we found that this SNP was located in a stage-specific D80 enhancer that interacted with the promoter of the cardiac sodium channel gene *SCN5A* (**Extended Data Fig. 12a**), supporting similar findings recently observed by circular chromosome conformation capture (4C) studies in human cardiac tissue and functional studies in mouse hearts[50]. Another cardiac GWAS example was rs17608766, the lead SNP for six cardiac traits including aortic root size, blood pressure, pulse pressure, QRS complex, QRS duration and systolic blood pressure. Our chromatin interaction maps revealed that this SNP overlapped a D2 mesoderm specific enhancer that interacted ~110 kb away with the promoter of *WNT3* (**Fig. 5c**), which is known to be involved in cardiac mesoderm induction[51,52] and thus may explain the broad cardiac traits potentially affected by this SNP. Because this SNP is within a highly conserved genomic region (**Fig. 5d**) and predicted to alter a binding motif for the KLF4 transcription factor (**Extended Data Fig. 12b),** which has been shown to regulate cardiac function in mice[53], we deleted the putative enhancer containing this SNP in hESCs using CRISPR-mediated genome editing. This deletion consequently resulted in a ~20-50% reduction

336 in *WNT3* expression in mesodermal cells at day 2 (**Fig. 5e)**, but had no effect on
337 expression of the nearest gene *GOSR2* (**Extended Data Fig. 12c**), which has
338 previously been annotated to rs17608766. Finally, in addition to investigating SNPs
339 associated with cardiac functional traits, we also examined long-range acting SNPs for
340 cardiac developmental traits (i.e. congenital heart disease – CHD). Interrogating
341 rs870142, a SNP associated with atrial septal defect[54], revealed that it was located in an
342 active D2 mesodermal enhancer within the intron of *STX18-AS1*, which formed dynamic
343 long-range interactions (~200 kb) with the promoter of *MSX1* (**Fig. 5f, g**), a
344 developmental TF reported to interact with *TBX5* during atrial septal development[55,56].
345 Deletion of this enhancer containing rs870142 using CRISPR/Cas9-mediated genome
346 editing resulted in a 50% reduction in *MSX1* expression (**Fig. 5h**), whereas the
347 expression of the nearest gene *STX18* remained unchanged (**Extended Data Fig. 12d)**.
348 Thus, these findings support that our chromatin interaction maps can be used to
349 discover candidate genes affected by genetic disease or trait variants located in distal
350 regulatory elements controlling gene expression.
351
352 Evolution of multi-cellular organisms is driven, in large part, by the invention of new cell-
353 type specific gene regulatory circuits responsible for the fitness traits of each species.
354 Our interrogation of the dynamic re-organization of 3D chromatin architecture during
355 cardiomyocyte differentiation reveals novel insights into not only how these gene
356 regulatory networks are rewired to specifically control cardiac development but also
357 more broadly how the primate specific retrotransposon, HERV-H, creates new human
358 PSC-specific TADs during primate evolution. Applying these findings to identifying
359 chromatin loops during cardiomyocyte differentiation moreover aided in predicting
360 potential target genes for cardiac GWAS variants, which were further validated by
361 CRISPR-mediated deletion studies in both congenital and adult heart disease/trait
362 examples. Although previous research suggests that HERV-Hs integrated into the
363 human genome during primate evolution to regulate human-specific pluripotency
364 through creating novel chimeric transcripts (*ESRG*[26], *linc-ROR*[30,57] and *LINC00458*[38])
365 and providing potential binding sites to recruit pluripotency factors (*NANOG*, *SOX2* and
366 *POU5F1*)[27], our findings show that HERV-H sequences may also create human PSC

367 specific TAD boundaries to build human-specific pluripotency gene regulatory networks,
368 thus supporting a long-standing theory that retroviruses may play an important role in
369 modifying gene regulatory logic during evolution[58,59]. Additionally, we provide evidence
370 that HERV-H transcription may achieve TAD boundary formation through accumulating
371 cohesin complexes at the 3' terminal of these actively transcribed retrotransposons.
372 Consistent with this possibility, we observed that silencing of HERV-H transcription
373 during differentiation or CRISPR-mediated removal of HERV-H sequences results in
374 loss of corresponding TAD boundaries. In addition to affecting the transcription of
375 genes within the TADs that are upstream of the 5' end of HERV-H*s*, this transcriptional
376 silencing is closely linked to hPSC differentiation, and in one particular case, deletion of
377 a specific HERV-H that appears to control the cardiac regulatory lncRNA *HBL1*, leads to
378 increased differentiation of hPSCs into cardiomyocytes. Overall, these findings not only
379 highlight a novel role for endogenous retroviruses in constructing species-specific PSC
380 chromatin architecture but also provide a genomic blueprint for understanding how gene
381 regulatory networks are rewired during cardiomyocyte development and disrupted by
382 non-coding genetic variants to lead to congenital and adult heart diseases.

383 **Acknowledgements**
384 We thank Samantha Kuan and Dr. Bin Li for sequencing and bioinformatics support.
385 We thank Eric van Nostrand for RNA extraction. This project is supported by funding
386 from the Ludwig Institute for Cancer Research (to B.R.), the NIH (1UM1HL128773 to
387 S.M.E., N.C.C., E.D. and B.R.). S.P. was supported by a postdoctoral fellowship from
388 the Deutsche Forschungsgemeinschaft (DFG, PR 1668/1-1). E.N.F. was supported by
389 NIH pre-doctoral training grant (5T32HL007444-35).
390

391 **Author contributions**
392 N.C.C. and B.R. designed and supervised the experiments, analysis, and data
393 interpretation. Y.Z. implemented the analysis pipeline and analyzed all sequencing
394 datasets, interpreted the results, designed the experiments for HERV-H and chromatin
395 loop functional studies. T.L. generated the CRISPR-cas9 edited cell lines for HERV-H
396 and chromatin loop functional studies, performed differentiation and qPCR of the
397 corresponding cell lines. S.P. initiated the project, performed the Hi-C and ATAC-seq
398 experiments, and helped with interpretation of the results. J.G. and E.N.F performed cell
399 culture, differentiation and collected cells for Hi-C, ChIP-seq, ATAC-seq and RNA-seq
400 assays. E.D. contributed to analysis and interpretation of the ChIP-seq data. ChIP-seq
401 experiments were performed by A.Y.L. (histone modifications), S.C. (CTCF), Q.Z. and
402 H.H. (SMC3). Y.Q. and R.F. helped with the analysis of Hi-C datasets. K.M. helped with
403 the genome-editing experiments. Z.Y. performed the RNA-seq experiments. R.H.
404 performed the Hi-C experiments for HERV-H KO cell lines. S.M.E. helped with
405 interpretation of the results. Y.Z., T.L., S.P., N.C.C., and B.R. wrote the manuscript with
406 input from all authors.
407

408 **Author Information**
409 Sequencing data from this study have been deposited in the Gene Expression Omnibus
410 under the accession number GSE116862. Reprints and permissions information is
411 available at www.nature.com/reprints. The authors declare no competing financial
412 interests. Readers are welcome to comment on the online version of the paper.

Correspondence and requests for materials should be addressed to N.C.C. (nchi@ucsd.edu) and B.R. (biren@ucsd.edu).

416

417 # Figure Legends

418 **Figure 1. Global chromatin architecture reorganization during human ventricular**
419 **cardiomyocyte differentiation.** (**a & b**) Hi-C contact matrices for each stage of
420 cardiomyocyte differentiation at mega-base resolution and at higher resolution at the
421 *HAND2* locus with corresponding epigenomic profiles. (**c**) Schematic illustration of the
422 multi-step decommissioning of inhibitory loop and assembly of subsequent activating
423 loops anchored at *HAND2* promoter. (**d**) A heatmap showing hierarchical clustering of
424 dynamic chromatin compartments during cardiomyocyte differentiation. The pseudo
425 color reflects the PC1 values (compartment A/B) of compartment bins. Negative PC1
426 value stands for compartment B and positive for compartment A. Representative genes
427 located in corresponding compartment bins are annotated to the right of the heatmap.
428 (**e**) A heatmap showing the DI delta scores for the stage-specific TAD boundaries,
429 ordered by the presence of TADs at six stages. vCM(-) stands for TAD boundaries lost
430 in purified ventricular cardiomyocytes at D80, ESC(+) for hESC-specific TAD
431 boundaries.

432

433 **Figure 2. Transcriptionally active HERV-H forms human ESC-specific TAD**
434 **boundaries.** (**a**) Aggregate RNA-seq expression profile (RPKM normalized) at ESC(+)
435 TAD boundaries that overlap HERV-H element. (**b**) Scatterplot shows the expression
436 levels (RPKM) across different HERV-H loci (ordered by expression levels from high to
437 low in ESC) at D0, D2 and D5 stages of differentiation. (**c**) Heatmap of aggregate Hi-C
438 contact matrix [log2(observed/expected)] within 200 kb of the top 50, 51-100 and 101-
439 150 ranked HERV-Hs, at D0. (**d**) Heatmap of the aggregated Hi-C matrix
440 [log2(observed/expected)] within 200 kb of the top 50 HERV-Hs, at D0, D2 and D5. (**e,f**)
441 Representative Hi-C interaction matrices of two HERV-H loci located at ESC(+) TAD
442 boundaries at D0, D2 and D5 (top) are shown as heatmaps along with genome browser
443 tracks of POLR2A, SMC3, CTCF, H3K27ac ChIP-seq and RNA-seq data of the
444 expanded genomic region containing the TAD boundary (arrow). (**g**) Aggregated
445 genomic profiles of RNA-seq, POLR2A, SMC3 and CTCF ChIP-seq around top 50

446 HERV-Hs located on the ESC(+) TAD boundaries (red) and lower ranked HERV-Hs
447 (grey).
448
449 **Figure 3. Deletion of two HERV-H sequences leads to merging of TADs and**
450 **dysregulation of gene expression programs in hESC**. (**a**) Hi-C interaction matrices
451 of the wild-type (WT) and mutant hESC lines (HERV-H1-KO and HERV-H2-KO*)* are
452 shown, along with DI scores, and fold changes of gene expression at the HERV-H1 and
453 HERV-H2 loci. The loss of TAD boundary in the mutant cells is accompanied with
454 decrease of RNA expression 5' terminus to the HERV-H sequences. (**b**) Boxplots show
455 expression levels (RPKMs) of genes whose TSSs are located from -500 kb to the 5'
456 LTR and from 3' LTR to +500 kb of HERV-Hs of boundary-associated HERV-Hs. (**c**)
457 Line chart (mean ± standard error, N=3 cardiomyocyte differentiations) shows
458 percentage of TNNT2 positive cells during cardiomyocyte differentiation of WT and
459 HERV-H1-KO hESC lines (two HERV-H1-KO clones analyzed). HERV-H1-KO hESCs
460 display increased cardiomyocyte differentiation efficiency compared to WT hESCs. (**d**)
461 Hi-C contact matrix and DI scores of the mouse syntenic loci to the human loci
462 examined in (**a**). The arrows indicate the boundaries observed in human ESCs. (**e**)
463 Heatmap of the aggregated Hi-C contact matrix [log2(observed/expected)] of the lifted-
464 over hESC TAD boundaries associated with HERV-H loci.
465
466
467 **Figure 4. Dynamics of chromatin loops during human cardiomyocyte**
468 **differentiation.** (**a**) Bar-chart shows the percentages of static or dynamic chromatin
469 loops during cardiomyocyte differentiation. (**b**) Heatmap shows K-means clustering of
470 the stage-specific chromatin loops. Pseudo color reflects the normalized contact
471 frequencies between the loop anchors for each stage-specific loop. (**c**) Enriched Gene
472 Ontology (GO) terms for genes located at the loop anchors of each cluster in (**b**) along
473 with their adjusted p-values. (**d**) Over-represented transcription factor binding motifs for
474 distal open chromatin regions located at the loop anchors of each cluster in (**b**) and their
475 p-values. CTCF motif enrichment was not shown but observed in all clusters. (**e**)
476 Boxplots of Pearson correlation coefficients of histone marks of distal enhancers and

477 expression levels of their target genes predicted based on the chromatin loops.
478 Distance-matched controls are “mirror” loops that connect to genomic regions from the
479 opposite direction with the same distances. P-values are from Wilcoxon Rank Sum test.
480 (**f**) Diagram illustrates how the degree of chromatin interactions for each loop anchor is
481 determined. (**g**) Histogram of degrees of network connections for all loop anchors. (**h**)
482 Fraction of chromatin interaction network hubs containing at least one transcription start
483 site (TSS) of a transcription factor (TF) plotted against the degree of connectivity. (**i**)
484 Fraction of network hubs containing at least one TSS of a house-keeping gene (HKG)
485 plotted against the degree of connectivity.
486
487 **Figure 5. Identification of target gene for heart-related GWAS SNPs using**
488 **chromatin loops detected during cardiomyocyte differentiation.** (**a**) Bar graph
489 shows the enrichment of GWAS SNPs associated with cardiac traits at the chromatin
490 loop anchors detected during the cardiomyocyte differentiation. (**b**) Enriched GO terms
491 for the genes predicted to be linked to SNPs from cardiac GWAS studies along with P
492 values. Red dashed line indicates P value cutoff of 0.05. (**c**) Hi-C contact matrices, and
493 genome browser tracks of various histone modification ChIP-seq and RNA-seq data at
494 the *WNT3* locus. A cardiac GWAS SNP rs17608766 is marked by a red line. (**d**)
495 Expanded view of the genomic region containing rs17608766. The putative enhancer
496 containing rs17608766, which located in an open chromatin region flanked by H3K27ac
497 and H3K4me1 peaks, is targeted for CRISPR/Cas9-mediated deletion. (**e**) Bar charts
498 show the normalized expression levels of *WNT3* in three CRISPR genome edited
499 clones at D2 (mean ± standard error, N =5). Deleting the putative enhancer containing
500 rs17608766 specifically reduces *WNT3* expression. P-values are from one-sided paired
501 t-test. (**f**) Hi-C contact matrices, epigenomic profiles at *MSX1* locus. A cardiac GWAS
502 SNP rs870142 is marked by red line. (**g**) Expanded view of the genomic region
503 containing rs870142. CRISPR mediated deletion of the putative enhancer is marked.
504 (**h**) Bar charts show that deleting the putative enhancer containing rs870142 specifically
505 reduces *MSX1* expression in three independent clones at D2 (mean ± standard error, N
506 = 9). P-values are from one-sided paired t-test.
507

508 **Table 1. List of target genes assigned to cardiac GWAS SNPs with chromatin**
509 **loops.** The first three columns show the ATAC-seq peaks that the SNPs overlapped.
510 Last column “tier” was based on our knowledge of the gene’s known or likely
511 importance in cardiac differentiation and function.

512

513 Extended Data Figure Legends.

514

515 **Extended Data Figure 1. Imaging and Flow cytometry data for cardiomyocyte**
516 **differentiation.** (**a**) Immunostaining for MYL2 (white) shows exclusive expression in
517 *MYL2*-H2B-GFP + hESC-derived cardiomyocytes after 80 days of cardiac
518 differentiation. *MYL2*-H2B-GFP + cells were positive for cardiac troponin T
519 (cTNT) (red), but not all cTnT+ cells were MYL2+ cardiomyocytes. DNA was
520 stained with DAPI (blue). Images are representative of a minimum of three
521 independent experiments. (**b**) Flow cytometric quantification of distinct cell-
522 surface markers for D2-D15 time points: T (D2), KDR/PDGFRα (D5), TNNT2
523 (D15). H2B-GFP + ventricular cardiomyocytes were sorted at day 80. Numbers
524 represent the respective percentage of cells.

525

526 **Extended Data Figure2. The Hi-C data are highly reproducible. (a**) Smoothed
527 scatterplot of the Hi-C contacts between biological replicates for each stage. bin size =
528 100 kb. (**b**) Hierarchical clustering of Hi-C contact matrices based on 1-PCC (Pearson
529 correlation coefficients) of the contacts between all samples, bin size = 100 kb. (**c**)
530 Hierarchical clustering of the compartment A/B scores (PC1 values) between samples.
531 (**d**) Hierarchical clustering of the insulation scores between samples.

532

533 **Extended Data Figure 3. Quality metrics of ChIP-seq, RNA-seq and ATAC-seq**
534 **data used in the present study.** (**a**) Boxplot showing the number of reads for each
535 sample. (**b**) Boxplot showing the percentage of alignments to human genome for each
536 sample. (**c**) Boxplot showing the percentage of potential duplicated reads for each
537 sample. (**d**) Boxplot showing the number of peaks called for each ChIP-seq and ATAC-
538 seq sample. (**e**) Barplots showing the relative expression levels (by RNA-seq) of
539 representative genes during cardiomyocyte differentiation.

540

541 **Extended Data Figure 4. Global changes in chromatin organization during**
542 **cardiomyocyte differentiation.** (**a**) Probability of Hi-C interaction versus interacting
543 distance (log2) during cardiomyocyte differentiation. (**b**) Barplot shows the percentages
544 of compartment switches for each stage transition. (**c**) Histogram of the Pearson
545 correlation coefficients between gene expression and compartment A/B (PC1) value. (**d**)
546 Genome browser view of *SOX2*, *HAND2* and *RYR2* loci, showing PC1 (Blue/red), RNA-
547 seq (Black) and H3K27ac (Blue) signals.
548
549 **Extended Data Figure 5. Dynamics of TADs during cardiomyocyte differentiation.**
550 (**a**) Number of TADs as defined by a domain caller algorithm first reported by Dixon et
551 al. 2012[9]. (**b**) Number of TADs (or loop domains) found by Arrowhead algorithm[13]. (**c**)
552 The number of TADs found by Insulation score[22]. (**d**) Fraction of TAD boundaries that
553 contain CTCF ChIP-seq peaks at each stage of cardiomyocyte differentiation. “stable”
554 group stands for TADs that are present at all stages. (**e**) Boxplot showing the dynamics
555 of H3K27ac signal at ESC(+) TADs. (**f**) Boxplot showing the dynamics of RNA-seq
556 signal at ESC(+) TADs.
557
558 **Extended Data Figure 6. HERV-H elements are enriched at ESC(+) TAD**
559 **boundaries. (a)** Scatter plot shows the fold enrichment and -log10(p-values) for various
560 repeat element classes at ESC(+) and vCM(-) TAD boundaries relative to the rest of the
561 TAD boundaries. P-values are from two-sided proportion test. (**b**) Aggregate RNA-seq
562 expression profile (RPKM normalized) at ESC(+) TAD boundaries that overlap each of
563 the eight repeat elements (HERVH-int and LTR7 are combined as they both belong to
564 HERV-H). (**c**) Heatmap of DI scores within 70 kb of the top 50 most highly-expressed
565 HERV-Hs. (**d,e,f**) Hi-C heatmap, DI, and epigenomic profiles for three ESC(+) TAD
566 boundaries harboring HERV-H/LTR7 sequences.
567
568 **Extended Data Figure 7. Further characterization of the HERV-H-associated TAD**
569 **boundaries.** (**a**) DI score profiles around the TSSs of genes expressed at similar levels
570 to the top 50 transcribed HERV-Hs. (**b**) Aggregated DI score profiles of the TAD-
571 boundary associated with HERV-Hs in multiple H1 ESC derived lineages and iPSCs.

572 ESC: embryonic stem cell; MES: mesendoderm. MSC: mesenchymal stem cell; NPC:
573 neural stem cell; TRO: trophoblast-like cell; iPSC: induced pluripotent stem cell. (**c**)
574 Barplot shows the ChIP-seq signal fold enrichment of the top 50 HERV-Hs comparing to
575 HERV-Hs ranking 51-300.
576
577 **Extended Data Figure 8. Changes in gene expression in HERV-H knockout**
578 **hESCs.** (**a**) MA plot showing average gene expression levels and fold changes of each
579 gene in HERV-H1-KO and wild-type (WT). (**b**) Same as (**a**) but for HERV-H2-KO. (**c**)
580 Scatterplot shows the changes in gene expression in HERV-H1-KO and HERV-H2-KO
581 cells over WT cells. The red dots are genes significantly changed in both mutant cell
582 lines. The numbers at the corners show the number of significantly changed genes in
583 each Quadrant. (**d**) Barplot showing the number of significantly changed genes located
584 within 20 kb of the HERV-H sequences. Genes down-regulated in both HERV-H
585 knockouts were more likely to be within 20 kb of HERV-H sequences. (**e**, **f**) RNA-seq
586 profile of wild-type (*WT*) and HERV-H1-KO and HERV-H2-KO lines at the *SCGB3A2*
587 and *LINC00458/HBL1* gene loci.
588
589 **Extended Data Figure 9. Evolutionary analysis of HERV-H sequences and**
590 **associated TAD boundaries.** (**a**) Comparison between hESC and mESC TAD
591 boundaries reveals that a lower portion of HERV-H associated boundaries is present in
592 mESCs. P-value of the difference between groups was computed using two-sided
593 proportion test. (**b**) Bar graph shows the percentage of HERV-Hs flanked by various
594 types of LTRs. HERV-Hs are ranked by their expression levels in the hESCs, and
595 grouped by bins of 50. (**c**) Violin plot shows the length of the flanking LTRs. HERV-Hs
596 ranked and binned as described in (**b**). (**d**) Boxplot shows the sequence divergence of
597 the 5' LTR and 3' LTR for each bin of HERV-Hs. HERV-Hs ranked and binned same as
598 (**b**).
599
600 **Extended Data Figure 10. Relationships between chromatin loops and histone**
601 **modification during cardiomyocyte differentiation.** (**a**) Barplot showing fraction of
602 chromatin loops in each category of interactions. E-E: enhancer to enhancer; E-P:

603 enhancer to promoter; P-P: promoter to promoter. (**b**) Barplot showing fraction of
604 chromatin loop anchors or genome-wide background containing each of the epigenomic
605 element. (**c**) Representative data for Hi-C contact matrices of stage-specific and static
606 chromatin loop interactions (**d**) Heatmap shows the clustering of stage-specific
607 chromatin loops and histone modification states at the corresponding loop anchors. (**e**)
608 Violin plot of the genomic distances for each class of chromatin loops in log scale.
609 Symbol S: static loop. 1-5: Loop clusters 1-5 in **Fig.4b** and **Ext. Fig. 10d**. (**f**) Barplot for
610 fraction of loops harboring different type of CTCF orientations. No-CTCF: at least one
611 anchor does not contain CTCF peak; unknown: at least one anchor does not contain a
612 matched CTCF motif; incorrect: CTCF motifs at both anchors do not face inward;
613 inward: CTCF motifs at both anchors face inward. (**g**) Barplot for fraction of loops in
614 each category related to TAD boundaries. Inter-TAD: loop is outside of all TADs; TAD-
615 corner: loop is within 20 kb of any TAD corner; intra-TAD: loop is inside of TAD.
616
617 **Extended Data Figure 11. Chromatin interaction hubs around genes encoding**
618 **cardiac transcription factors.** (**a**) Example of a H3K27me3-associated inhibitory loop
619 connecting *MSX1* and *HMX1* promoters. (**b**) Example of several stage-specific and
620 H3K27ac associated activating loops at *HAND1* locus. (**c,d,e**) Hi-C contact matrices
621 (top), loop arcs (middle), and epigenomic profiles (bottom) for *GATA4* (**c**), *MEIS1* (**d**)
622 and *TBX5* (**e**). (**f**) Fraction of network hubs containing CTCF peaks plotted against the
623 degree of connectivity as illustrated by the number of CTCF peaks at each anchor
624 connectivity.
625
626 **Extended Data Figure 12. Validation of target gene predictions for heart-related**
627 **GWAS SNPs.** (**a**) Hi-C contact matrices, epigenomic profiles (ATAC-seq, H3K4me1,
628 H3K27ac and RNA-seq) and cardiac GWAS SNP rs6781009 at *SCN5A* locus. (**b**)
629 rs17608766 is predicted to alter KLF4 motif in the putative enhancer. (**c**) Normalized
630 expression values of *GOSR2* in three CRISPR-genome edited clones shows that
631 deleting the putative enhancer containing rs17608766 does not change *GOSR2*
632 expression (mean ± standard error, N=5). (**d**) Normalized expression values of *STX18*

633 in three CRISPR-genome edited clones shows that deleting the putative enhancer
634 containing rs870142 does not change *STX18* expression (mean ± standard error, N=7).
635
636
637 **Extended Data Table 1. Summary statistics for Hi-C data**
638 **Extended Data Table 2. List of stage-specific TAD boundaries.**
639 **Extended Data Table 3. List of differentially expressed genes in HERV-H1-KO and**
640 **HERV-H2-KO.**
641 **Extended Data Table 4. List of chromatin loops.**
642 **Extended Data Table 5. List of genes located on the network hubs.**
643 **Extended Data Table 6. List of additional target genes assigned to cardiac GWAS**
644 **SNPs with chromatin loops.**
645 **Extended Data Table 7. List of antibodies used for ChIP-seq.**
646 **Extended Data Table 8. List of primers and cell lines used in this study.**
647
648
649
650 **Supplementary Materials:**
651 Materials and Methods
652 Figures S1-S12
653 Tables S1-S7
654
655

656 # References

657 1 Dekker, J. & Mirny, L. The 3D Genome as Moderator of Chromosomal Communication. *Cell* **164**, 1110-
658 1121, doi:10.1016/j.cell.2016.02.007 (2016).
659 2 Dixon, J. R., Gorkin, D. U. & Ren, B. Chromatin Domains: The Unit of Chromosome Organization. *Mol*
660 *Cell* **62**, 668-680, doi:10.1016/j.molcel.2016.05.018 (2016).
661 3 Yu, M. & Ren, B. The Three-Dimensional Organization of Mammalian Genomes. *Annu Rev Cell Dev Biol*
662 **33**, 265-289, doi:10.1146/annurev-cellbio-100616-060531 (2017).
663 4 Zaidi, S. *et al.* De novo mutations in histone-modifying genes in congenital heart disease. *Nature* **498**, 220-
664 223, doi:10.1038/nature12141 (2013).
665 5 Homsy, J. *et al.* De novo mutations in congenital heart disease with neurodevelopmental and other
666 congenital anomalies. *Science* **350**, 1262-1266, doi:10.1126/science.aac9396 (2015).

667 6 Jin, S. C. *et al.* Contribution of rare inherited and de novo variants in 2,871 congenital heart disease
668 probands. *Nat Genet* **49**, 1593-1601, doi:10.1038/ng.3970 (2017).
669 7 Meaburn, K. J. & Misteli, T. Cell biology: chromosome territories. *Nature* **445**, 379-781,
670 doi:10.1038/445379a (2007).
671 8 Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of
672 the human genome. *Science* **326**, 289-293, doi:10.1126/science.1181369 (2009).
673 9 Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin
674 interactions. *Nature* **485**, 376-380, doi:10.1038/nature11082 (2012).
675 10 Sexton, T. *et al.* Three-Dimensional Folding and Functional Organization Principles of the Drosophila
676 Genome. *Cell* **148**, 458-472, doi:10.1016/j.cell.2012.01.010 (2012).
677 11 Nora, E. P. *et al.* Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485**,
678 381-385, doi:10.1038/nature11049 (2012).
679 12 Dixon, J. R. *et al.* Chromatin architecture reorganization during stem cell differentiation. *Nature* **518**, 331-
680 336, doi:10.1038/nature14222 (2015).
681 13 Rao, S. S. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin
682 looping. *Cell* **159**, 1665-1680, doi:10.1016/j.cell.2014.11.021 (2014).
683 14 Jin, F. *et al.* A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature*
684 **503**, 290-294, doi:10.1038/nature12644 (2013).
685 15 Bonev, B. *et al.* Multiscale 3D Genome Rewiring during Mouse Neural Development. *Cell* **171**, 557-572
686 e524, doi:10.1016/j.cell.2017.09.043 (2017).
687 16 Stadhouders, R. *et al.* Transcription factors orchestrate dynamic interplay between genome topology and
688 gene regulation during cell reprogramming. *Nat Genet* **50**, 238-249, doi:10.1038/s41588-017-0030-7
689 (2018).
690 17 Krijger, P. H. *et al.* Cell-of-Origin-Specific 3D Genome Structure Acquired during Somatic Cell
691 Reprogramming. *Cell Stem Cell* **18**, 597-610, doi:10.1016/j.stem.2016.01.007 (2016).
692 18 Beagan, J. A. *et al.* Local Genome Topology Can Exhibit an Incompletely Rewired 3D-Folding State
693 during Somatic Cell Reprogramming. *Cell Stem Cell* **18**, 611-624, doi:10.1016/j.stem.2016.04.004 (2016).
694 19 Gelb, B. D. Genetic basis of congenital heart disease. *Curr Opin Cardiol* **19**, 110-115 (2004).
695 20 Veevers, J. *et al.* Cell-Surface Marker Signature for Enrichment of Ventricular Cardiomyocytes Derived
696 from Human Embryonic Stem Cells. *Stem cell reports*, doi:10.1016/j.stemcr.2018.07.007 (2018).
697 21 Fields, P. A. *et al.* Dynamic reorganization of nuclear architecture during human cardiogenesis. *bioRxiv*
698 (2017).
699 22 Crane, E. *et al.* Condensin-driven remodelling of X chromosome topology during dosage compensation.
700 *Nature* **523**, 240-244, doi:10.1038/nature14450 (2015).
701 23 Nora, E. P. *et al.* Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains
702 from Genomic Compartmentalization. *Cell* **169**, 930-944 e922, doi:10.1016/j.cell.2017.05.004 (2017).
703 24 Grow, E. J. *et al.* Intrinsic retroviral reactivation in human preimplantation embryos and pluripotent cells.
704 *Nature* **522**, 221-225, doi:10.1038/nature14308 (2015).
705 25 Goke, J. *et al.* Dynamic transcription of distinct classes of endogenous retroviral elements marks specific
706 populations of early human embryonic cells. *Cell Stem Cell* **16**, 135-141, doi:10.1016/j.stem.2015.01.005
707 (2015).
708 26 Lu, X. *et al.* The retrovirus HERVH is a long noncoding RNA required for human embryonic stem cell
709 identity. *Nat Struct Mol Biol* **21**, 423-425, doi:10.1038/nsmb.2799 (2014).
710 27 Romer, C., Singh, M., Hurst, L. D. & Izsvak, Z. How to tame an endogenous retrovirus: HERVH and the
711 evolution of human pluripotency. *Curr Opin Virol* **25**, 49-58, doi:10.1016/j.coviro.2017.07.001 (2017).
712 28 Santoni, F. A., Guerra, J. & Luban, J. HERV-H RNA is abundant in human embryonic stem cells and a
713 precise marker for pluripotency. *Retrovirology* **9**, 111, doi:10.1186/1742-4690-9-111 (2012).
714 29 Ohnuki, M. *et al.* Dynamic regulation of human endogenous retroviruses mediates factor-induced
715 reprogramming and differentiation potential. *Proc Natl Acad Sci U S A* **111**, 12426-12431,
716 doi:10.1073/pnas.1413299111 (2014).
717 30 Loewer, S. *et al.* Large intergenic non-coding RNA-RoR modulates reprogramming of human induced
718 pluripotent stem cells. *Nat Genet* **42**, 1113-1117, doi:10.1038/ng.710 (2010).
719 31 Xie, W. *et al.* Epigenomic analysis of multilineage differentiation of human embryonic stem cells. *Cell*
720 **153**, 1134-1148, doi:10.1016/j.cell.2013.04.022 (2013).
721 32 Greenwald, W. W. *et al.* Integration of phased Hi-C and molecular phenotype data to study genetic and
722 epigenetic effects on chromatin looping. *bioRxiv* (2018).

723 33 Lengronne, A. *et al.* Cohesin relocation from sites of chromosomal loading to places of convergent
724 transcription. *Nature* **430**, 573-578, doi:10.1038/nature02742 (2004).
725 34 Busslinger, G. A. *et al.* Cohesin is positioned in mammalian genomes by transcription, CTCF and Wapl.
726 *Nature* **544**, 503-507, doi:10.1038/nature22063 (2017).
727 35 Jacques, P. E., Jeyakani, J. & Bourque, G. The majority of primate-specific regulatory sequences are
728 derived from transposable elements. *PLoS Genet* **9**, e1003504, doi:10.1371/journal.pgen.1003504 (2013).
729 36 Barakat, T. S. *et al.* Functional Dissection of the Enhancer Repertoire in Human Embryonic Stem Cells.
730 *Cell Stem Cell* **23**, 276-288 e278, doi:10.1016/j.stem.2018.06.014 (2018).
731 37 Wang, J. *et al.* Primate-specific endogenous retrovirus-driven transcription defines naive-like stem cells.
732 *Nature* **516**, 405-409, doi:10.1038/nature13804 (2014).
733 38 Ng, S. Y., Johnson, R. & Stanton, L. W. Human long non-coding RNAs promote pluripotency and neuronal
734 differentiation by association with chromatin modifiers and transcription factors. *EMBO J* **31**, 522-533,
735 doi:10.1038/emboj.2011.459 (2012).
736 39 Liu, J., Li, Y., Lin, B., Sheng, Y. & Yang, L. HBL1 Is a Human Long Noncoding RNA that Modulates
737 Cardiomyocyte Development from Pluripotent Stem Cells by Counteracting MIR1. *Dev Cell* **42**, 333-348
738 e335, doi:10.1016/j.devcel.2017.07.023 (2017).
739 40 Goodchild, N. L., Wilkinson, D. A. & Mager, D. L. Recent evolutionary expansion of a subfamily of
740 RTVL-H human endogenous retrovirus-like elements. *Virology* **196**, 778-788, doi:10.1006/viro.1993.1535
741 (1993).
742 41 Mager, D. L. & Freeman, J. D. HERV-H endogenous retroviruses: presence in the New World branch but
743 amplification in the Old World primate lineage. *Virology* **213**, 395-404, doi:10.1006/viro.1995.0012
744 (1995).
745 42 Kijima, T. E. & Innan, H. On the estimation of the insertion time of LTR retrotransposable elements. *Mol*
746 *Biol Evol* **27**, 896-904, doi:10.1093/molbev/msp295 (2010).
747 43 McCulley, D. J. & Black, B. L. Transcription factor pathways and congenital heart disease. *Curr Top Dev*
748 *Biol* **100**, 253-277, doi:10.1016/B978-0-12-387786-4.00008-7 (2012).
749 44 Bantignies, F. *et al.* Polycomb-dependent regulatory contacts between distant Hox loci in Drosophila. *Cell*
750 **144**, 214-226, doi:10.1016/j.cell.2010.12.026 (2011).
751 45 Schoenfelder, S. *et al.* Polycomb repressive complex PRC1 spatially constrains the mouse embryonic stem
752 cell genome. *Nat Genet* **47**, 1179-1186, doi:10.1038/ng.3393 (2015).
753 46 Javierre, B. M. *et al.* Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease
754 Variants to Target Gene Promoters. *Cell* **167**, 1369-1384 e1319, doi:10.1016/j.cell.2016.09.037 (2016).
755 47 Won, H. *et al.* Chromosome conformation elucidates regulatory relationships in developing human brain.
756 *Nature* **538**, 523-527, doi:10.1038/nature19847 (2016).
757 48 Montefiori, L. E. *et al.* A promoter interaction map for cardiovascular disease genetics. *Elife* **7**,
758 doi:10.7554/eLife.35788 (2018).
759 49 Choy, M. K. *et al.* Promoter interactome of human embryonic stem cell-derived cardiomyocytes connects
760 GWAS regions to cardiac gene networks. *Nat Commun* **9**, 2526, doi:10.1038/s41467-018-04931-0 (2018).
761 50 van der Harst, P. *et al.* 52 Genetic Loci Influencing Myocardial Mass. *J Am Coll Cardiol* **68**, 1435-1448,
762 doi:10.1016/j.jacc.2016.07.729 (2016).
763 51 Pfeiffer, M. J. *et al.* Cardiogenic programming of human pluripotent stem cells by dose-controlled
764 activation of EOMES. *Nat Commun* **9**, 440, doi:10.1038/s41467-017-02812-6 (2018).
765 52 Cohen, E. D., Tian, Y. & Morrisey, E. E. Wnt signaling: an essential regulator of cardiovascular
766 differentiation, morphogenesis and progenitor self-renewal. *Development* **135**, 789-798,
767 doi:10.1242/dev.016865 (2008).
768 53 Liao, X. *et al.* Kruppel-like factor 4 is critical for transcriptional control of cardiac mitochondrial
769 homeostasis. *J Clin Invest* **125**, 3461-3476, doi:10.1172/JCI79964 (2015).
770 54 Cordell, H. J. *et al.* Genome-wide association study of multiple congenital heart disease phenotypes
771 identifies a susceptibility locus for atrial septal defect at chromosome 4p16. *Nat Genet* **45**, 822-824,
772 doi:10.1038/ng.2637 (2013).
773 55 Boogerd, K. J. *et al.* Msx1 and Msx2 are functional interacting partners of T-box factors in the regulation
774 of Connexin43. *Cardiovasc Res* **78**, 485-493, doi:10.1093/cvr/cvn049 (2008).
775 56 Li, Q. Y. *et al.* Holt-Oram syndrome is caused by mutations in TBX5, a member of the Brachyury (T) gene
776 family. *Nat Genet* **15**, 21-29, doi:10.1038/ng0197-21 (1997).
777 57 Wang, Y. *et al.* Endogenous miRNA sponge lincRNA-RoR regulates Oct4, Nanog, and Sox2 in human
778 embryonic stem cell self-renewal. *Dev Cell* **25**, 69-80, doi:10.1016/j.devcel.2013.03.002 (2013).

58 Britten, R. J. & Davidson, E. H. Repetitive and non-repetitive DNA sequences and a speculation on the origins of evolutionary novelty. *Q Rev Biol* **46**, 111-138 (1971).

59 King, M. C. & Wilson, A. C. Evolution at two levels in humans and chimpanzees. *Science* **188**, 107-116 (1975).

Materials and Methods

Table of Contents

1 Experimental approaches 3
1.1 Cell culture, differentiation, and collections for large-scale cardiac molecular studies. 3
1.2 In-situ Hi-C 4
1.3 ChIP-seq 5
1.4 RNA-seq 6
1.5 ATAC-seq 6
1.6 Cardiac chromatin loop functional studies 6
1.7 HERV-H functional studies 7
2 Data Analysis 8
2.1 ChIP-seq data processing: 8
2.1.1 Read alignment 8
2.1.2 ChIP-seq peak calling 8
2.1.3 ChIP-seq signal visualization 8
2.2 ATAC-seq data processing: 9
2.2.1 Read alignment 9
2.2.2 ATAC-seq peak calling 9
2.3 RNA-seq data processing: 9
2.3.1 Read alignment and quantification 9
2.3.2 Differentiation gene expression analysis 10
2.4 Hi-C data processing: 10
2.4.1 Compartments A/B 10
2.4.2 Insulation score 11
2.4.3 Directionality Index 11
2.4.4 Topological associated domains (TADs) 11
2.4.5 Chromatin Loops 13
2.5 HERV-H related analysis 16
2.5.1 Directionality index profile on HERV-Hs transcribed at different levels 16
2.5.2 Directionality index profile on HERV-Hs in iPSCs and H1-derived cells (relating to **Extended Data. Fig7. b)** 16

2.5.3 ChIP-seq enrichment analysis at TAD boundary-associating HERV-H loci (relating to **Extended Data. Fig. 7c**) 17
2.5.4 HERV-H LTR type and divergence analysis (relating to **Extended Data. Fig. 9**) 17
2.6 GWAS analysis: 18
2.6.1 GWAS lead SNP and SNPs in linkage disequilibrium 18
2.6.2 Binomial test for significance of enrichment of heart-traits GWAS SNPs within chromatin loops 18
Gene Ontology Enrichment testing of the genes. 18
2.7 General data processing and plots 19
3 Data and script availability: 19

1 Experimental approaches

1.1 Cell culture, differentiation, and collections for large-scale cardiac molecular studies.

For the large-scale genomic and transcriptomic human embryonic stem cell (hESC) cardiac studies (i.e. ATAC-seq, ChIP-seq, RNA-seq and HiC experiments), an engineered H9 hESC *MLC2v:H2B-GFP* reporter transgenic line, which specifically expresses H2B-GFP in differentiated ventricular cardiomyocytes (CMs), was used[1]. This H9 hESC *MLC2v:H2B-GFP* line was differentiated into cardiomyocytes and their intermediate developmental stages utilizing a well-established Wnt-based cardiomyocyte differentiation protocol[2] that we have previously used to efficiently create hESC-CMs[3]. Specifically, the following developmental cardiac cell populations were generated and collected: Day (D)0 (hESC), D2 (mesoderm), D5 (cardiac mesoderm), D7 (cardiac progenitors), D15 (primitive CMs) and D80 (ventricular CMs). Differentiation efficiency was assessed at specific developmental cardiac stages by flow cytometry using antibodies against POU5F1 (Cell Signaling C30A3) - D0/hESC, BRACHYURY/T (R&D Systems AF2085) - D2/mesoderm, KDR/PDGFRA (R&D Systems AF357P and AF1264A, respectively) - D5/cardiac mesoderm and D7/cardiac progenitors, and cTNNT (ThermoFisher MA5-12960) - D7/cardiac progenitors, D15/primitive CMs, D80/ventricular CMs. Differentiation success for collected D2-D7 samples was further confirmed by re-plating a portion of the cells dissociated for collection and then maintaining them until ~D15 when CM differentiation could be assessed using TNNT2 flow analysis. Collected samples that resulted in high differentiation efficiency (>90%) were then further processed and analyzed for molecular studies (see below). For D80 ventricular CMs, we additionally sorted and purified these cells based on *MLC2v*:H2B-GFP prior to their processing and analyses (**Extended Data Fig.1**). For RNA-seq studies, ~2M dissociated cells were washed in PBS, lysed in Trizol and then stored at -80°C until further processing. For ChIP-seq studies, dissociated cells were washed in PBS, fixed in 1% formaldehyde for 10 min, quenched with 0.125 M Glycine for 5 min at room temperature (RT), then 10 min on ice and finally washed twice with PBS. These fixed cell pellets were then snap-frozen and stored at -80°C until further processing. For ATAC-seq, dissociated cells were washed with PBS and cryopreserved in culture medium + 10% DMSO at -80°C overnight in Mr.

Frosty freezing containers (Thermo Scientific) and then stored in liquid N_2 until further processing as previously described[4].

1.2 In-situ Hi-C

In situ HiC was performed as described previously with minor modifications[5]. Cells were harvested and pelleted by centrifugation at 300 x g for 5 min. Cells were resuspended in fresh media at 1 million cells/mL and fixed with 2% formaldehyde for 10min at RT. The reaction was quenched by incubating with 200 mM glycine for 5 min at RT before transferring onto ice. Cells were pelleted (300 x g for 5 min at 4°C) and subsequently washed with ice cold PBS. Following supernatant removal, the pellet was flash frozen in liquid nitrogen and stored at -80°C. Nuclei were isolated using lysis buffer (10mM Tris-HCl (pH 8.0), 10mM NaCl, 0.2% Igepal CA630) and incubation on ice for 15 min. Pelleted nuclei were permeabilized using 0.5% SDS and incubation for 10 min at 62°C. SDS was quenched by adding Triton-X100. Chromatin was digested using the four base cutter restriction enzyme MboI (200 units, NEB) at 37 °C overnight. Digestion efficiency was checked by gel electrophoresis. After inactivation of MboI for 20 min at 62°C, single strand overhangs were filled with biotinylated-14-ATP (Life Technologies) using Klenow DNA polymerase (40 units, NEB) for 90 min at 37°C. Next, DNA was ligated for 4 h at 16 °C using T4 ligase (2000 units, NEB). Reverse crosslinking and protein degradation was performed using proteinase K (NEB) and incubation at 55 °C for 30 min. Finally, NaCl (500mM) was added and the reaction was completed at 68 °C overnight. DNA was purified using ethanol precipitation and sonicated using an ultra sonicator (Covaris M220; duty cycle: 10; intensity: 4; cycles/burst: 220; duration: 55s) to 300-700 bp small fragments. After size selection and purification using SPRI beads (Beckman Coulter), biotinylated DNA was isolated using Dynabeads MyOne T1 Streptavidin beads (Life Technologies). Sequencing libraries were prepared on magnetic beads and final PCR amplification was carried out for 6 cycles. Size distribution of bead purified libraries was checked using a High Sensitivity D1000 ScreenTape on a Tapestation (Agilent) and quantified using Qubit (Life Technologies). Libraries were sequenced on Hiseq 4000 (100 bp PE, Illumina).

1.3 ChIP-seq

ChIP-seq experiments for all histone marks at all time points were performed as previously described[6,7]. The experiments were conducted according to experimental guidelines and read depth standards of the ENCODE consortium. Harvested cells were washed in PBS and fixed in 1% formaldehyde for 10 min at room temperature. Reaction was quenched with 125 mM Glycine for 5 min at room temperature followed by 10 min incubation on ice. Next, the cells were washed twice with PBS. Crosslinked cells were lysed and chromatin sheared using a Branson Sonifier 450 (20 cycles, 15 sec ON, 45 sec OFF at power 3). Detailed protocols can be found on the ENCODE homepage for immunoprecipitation (https://www.encodeproject.org/documents/89795b31-e65a-42ca-9d7bd75196f6f4b3/@@download/attachment/Ren%20Lab%20ENCODE%20Chromatin%20Immunoprecipitation%20Protocol_V2.pdf) and library preparation (https://www.encodeproject.org/documents/4f73fbc3-956e-47ae-aa2d41a7df552c81/@@download/attachment/Ren_ChIP_Library_Preparation_v060614.pdf).

ChIP-seq experiments for CTCF and SMC3 were carried out on two and ten million crosslinked cells respectively, as previously described[8,9]. Nuclei were isolated using hypotonic buffer and subsequecntly lysed in RIPA buffer (140 mM NaCl, 1 mM EDTA, 1% Triton X-100, 0.1% SDS, 0.1% sodium deoxycholate, 10 mM Tris-HCl pH 8.0 supplemented with protease inhibitor (Roche)) and chromatin was sheared using a Covaris M220. Cell lysates were precleared using Protein G beads (Sigma) and rotation for 3h at 4°C. For immunoprecipitation, chromatin was incubated with antibodies against CTCF overnight at 4°C. Antibody chromatin complexes were bound by Protein G beads for 2h at 4°C. After eight rounds of washing using combinations of high and low salt buffers, crosslinks were reversed with Proteinase K (NEB) at 65°C overnight. DNA was purified by phenol /chloroform/isoamylalcohol extraction and ethanol precipitation. Sequencing libraries were generated using the Accel-NGS® 2S Plus DNA Library Kit (Swift).

All ChIP-seq libraries were sequenced on HiSeq2500/4000 (50 bp SE, Illumina).

A list of antibodies used in this study can be found in **Extended Data Table 7**.

1.4 RNA-seq

Total RNA from cells was extracted using Trizol (Thermo) and purified with RNA Clean & Concentrator™-5 (R1013, Zymo Research). RNA sequencing libraries were prepared from 1-2 µg total RNA using the Illumina TruSeq Stranded mRNA Library Prep Kit Set A (Illumina, RS-122-2101) or Set B (Illumina, RS-122-2102). Sequencing was carried out on a HiSeq4000 (100 bp PE, Illumina).

1.5 ATAC-seq

ATAC-seq was performed on 50,000 cells as described previously with minor modifications[10]. Cells were thawed at 37°C and pelleted for 5 min at 500 x g at 4 °C. Cell pellets were resuspended in 250 µl permeabilization buffer (5 % BSA, 0.2 % IGEPAL-CA630, 1mM DTT and cOmplete™, EDTA-free protease inhibitor cocktail (Roche) in PBS) and incubated on an overhead shaker for 5 min at 4°C, followed by centrifugation for 5 min at 500 x g. Pellet was resupended in 25 µl tagmentation buffer (36.3 mM Tris-acetate (pH = 7.8), 72.6 mM K-acetate, 11 mM Mg-acetate, 17.6 % DMF) and counted on a hemocytometer. 40,000 nuclei were used for tagmentation and the reaction volume was adjusted to 19 µl using tagmentation buffer. After addition of 1 µl Tn5 transposomes (Illumina) tagmentation was performed at 37°C for 60 min with shaking (500 rpm). Next, samples were purified using MinElute columns (Qiagen), PCR-amplified for 9 cycles with NEBNext® High-Fidelity 2X PCR Master Mix (NEB, 72°C 5 min, 98°C 30 s,[98°C 10 s, 63°C 30 s, 72°C 60 s] x cycles, 72°C forever). Amplified libraries were purified using MinElute columns (Qiagen) and Ampure XP Bead (Beckmann Coulter). Sequencing was carried out on a HiSeq2500 or 4000 (50 bp PE, Illumina).

1.6 Cardiac chromatin loop functional studies

To further validate identified chromatin loops regulating predicted enhancer-promoter interactions, selected enhancers were genetically deleted from hESC lines using CRISPR-Cas9 genome-editing strategies. To this end, specific gRNA pairs with binding sites surrounding the target region were designed[11] using a web-based software tool CRISPOR[12], which generates gRNA sequences for optimal targeting of region of interest while minimizing potential off-target effects. These identified gRNAs were then

synthesized *in vitro* using the GeneArt Precision gRNA Synthesis kit (Invitrogen). To better monitor the differentiation efficiency, we generated both eWNT3 and eMSX1 knockout line in H9 *TNNT2*:GFP hESCs, which enables us to dynamically track the cardiomyocyte generation by GFP[13]. One day before transfection, 1×10^5 H9 *TNNT2*:GFP hESCs were seeded in 12-well plates. When cells were 30-60% confluent, a pair of RNP complexes containing 2 µg of Cas9 protein (NEB) and 400 ng of *in vitro* transcribed gRNA were transfected using Lipofectamine Stem Transfection Reagent (Invitrogen)[14,15]. Three days after transfection, cells were diluted and clonally expanded for genotyping and functional assays[16]. Clonal mutant lines were genotyped using PCR and Sanger sequencing (**Extended Data Table 8**). For heterozygous knock-out alleles, a second round of targeting was performed to create homozygous knock-out lines. After confirmation of genome-edited clones, at least three successfully targeted clones per deletion were then selected for further functional analyses. Using the Wnt-based CM differentiation protocol[2] described above, enhancer deletion clones (i.e. eWNT3$^{-/-}$ and eMSX1$^{-/-}$) were differentiated to respective stages as to when enhancer-promoter loops appear. Differentiation efficiency was assessed as described above by flow cytometry and replating studies if applicable. Samples that resulted in high efficiency differentiation (>90%) were then further processed and analyzed by qPCR for the expression of cognate genes that are predicted to be regulated by studied enhancer. As a negative control, expression of the nearest genes of the studied enhancer was also examined. Gene expression levels were normalized to the expression of TBP of each sample. For WNT3 locus, four differentiations were performed; the first three differentiations had one replicate, the 4th differentiation had two replicates. For MSX1 locus, three differentiations were performed, each with three replicates.

1.7 HERV-H functional studies

To interrogate the functional significance of HERV-Hs, selected HERV-H sequences were genetically deleted in H9 *MLC2v*:H2B-GFP hESC lines using CRISPR-Cas9 genome-editing strategies as described above for cardiac chromatin loop functional studies. After confirmation of genome-edited clones by Sanger sequencing, at least two successfully targeted clones per deletion were selected for functional analysis. To

examine the effects of deleting HERV-Hs on chromatin boundary formation and gene expression, *in-situ* Hi-C and RNA-seq studies were respectively performed as described above. Because of the potential role of HERV-H1 in cardiac differentiation, HERV-H1 deleted H9 *MLC2v*:H2B-GFP hESC clones were additionally examined for their ability to differentiate into cardiomyocytes. To this end, hESC cell seeding density and canonical Wnt signaling levels were titrated to initially find suboptimal CM differentiation conditions that result in 50% CM differentiation by Day 15 in control H9 *MLC2v*:H2B-GFP hESCs. These suboptimal CM differentiation conditions were then applied to both control and HERV-H1 deleted H9 *MLC2v*:H2B-GFP hESC clones to further assess their ability to differentiate into CMs from Day 7-15 as detected by TNNT2 FACS analyses. The differentiation experiments were repeated three times.

2 Data Analysis

2.1 ChIP-seq data processing:

2.1.1 Read alignment

Reads were aligned to hg19 (GRC37) following standards of ENCODE ChIP-seq pipeline[17]. The full pipeline is available at https://github.com/ren-lab/chip-seq-pipeline. Briefly, the reads were mapped with bwa[18]. PCR duplicates were removed using Picard[19]. Only reads with mapping quality > 30 were kept.

2.1.2 ChIP-seq peak calling

For each stage, ChIP-seq peaks were called using MACS2[20] for each biological replicate and pooled sample, using the pooled input sample as control. MACS2 was run with the default setting with "--nomodel –extsize 180" parameter. Peaks called in the pooled sample that also intersected with peaks in both replicates by at least 50% bases were defined as replicated peaks and were used as the final peak list for that stage. Peaks from all stages were merged into a union peak list using bedtools mergeBed[21].

2.1.3 ChIP-seq signal visualization

ChIP-seq bigWig files were generated using deeptools[22] bamCoverage function with the RPKM (Reads Per Kilobase of transcript per Million mapped reads) normalization

option. Aggregated signal profiles were generated with the deeptools computeMatrix and plotProfile functions and further modified in R.

2.2 ATAC-seq data processing:

2.2.1 Read alignment

Reads were aligned to hg19 (GRC37) following standards of ENCODE ATAC-dnase-seq pipeline[23]. The full pipeline (including peak-calling) is available at https://github.com/ren-lab/atac-seq-pipeline. Briefly, the reads were mapped with bowtie2[24]. PCR duplicates were removed using Picard. Only reads with mapping quality > 30 were kept.

2.2.2 ATAC-seq peak calling

For each stage, ATAC-seq peaks were called using MACS2 for each biological replicate and pooled sample, using the pooled input sample as control. MACS2 was run with the default setting with "--nomodel –extsize 73 –p 0.01" parameter. Peaks called in the pooled sample that also intersected with peaks in both replicates by at least 50% were defined as replicated peaks and were used as the final peak list for that stage. Peaks from all stages were merged into a union peak list using bedtools mergeBed. The peaks were further separated into promoter-distal and promoter-proximal by whether the peaks intersected with GENCODE v19 transcription start sites (TSSs).

2.3 RNA-seq data processing:

2.3.1 Read alignment and quantification

RNA-seq alignment and quantification pipeline is available at https://github.com/ren-lab/rnaseq-pipeline. Briefly, reads were aligned to hg19 (GRC37) and GENCODE GTF v19 with rnaSTAR[25] following "ENCODE" options outlined in the rnaSTAR manual (http://labshare.cshl.edu/shares/gingeraslab/www-data/dobin/STAR/STAR.posix/doc/STARmanual.pdf). PCR duplicates were removed using Picard. Raw reads for each gene were quantified using featureCounts[26]. Read per million reads and kilobases (RPKM) were calculated for each gene.

2.3.2 Differentiation gene expression analysis

Differentially gene expression was performed using edgeR[27] with the default normalization method, tagwise dispersion estimation function and glmLRT (negative binomial generalized log-linear model).

2.4 Hi-C data processing:

The HiC data processing pipeline is available at https://github.com/ren-lab/hic-pipeline. Hi-C reads were aligned on hg19(GRC37) using BWA-MEM[18] for each read separately, and then paired. For chimeric reads, only 5' end mapped locations were kept. Duplicated read pairs mapped to the same location were removed to leave only one unique read pair. The output bam files were transformed into juicer file format and tag directories for further analysis in Juicebox [28]and HOMER[29], respectively. Unless specified otherwise, we used the contact matrix in 10kb resolution and normalized using KR matrix balancing method.

2.4.1 Compartments A/B

Principal component analysis was performed using HOMER[29] with resolution of 50kb. DNA segments with positive PC1 value in both biological replicates were defined as compartment A, whereas negative value for compartment B. Inconsistent PC1 signs between replicates were defined as inconsistent. For dynamic and static compartment analysis (**Fig 1.d**), only DNA segments with consistent compartment state at each of the six stages (which was 85% of all genomic bins) were included. A DNA segment was defined as static if it stayed compartment A or B at all six stages; Otherwise it was named dynamic. Compartment switch types were defined as "BA" if it switched from B to A, and *vice versa*. If the compartment bin switched from A to B, and then switched back to A, it will be defined as "ABA", and the same rule applies to "BAB". Bins with more than 2 switches were very rare and left out in the analysis/plot. Hierarchical clustering was performed on the PC1 values of all dynamic bins, by both rows (segments) and columns (sample replicates), using (1- Pearson correlation coefficient) as distance and average linkage method to merge branches.

2.4.2 Insulation score

To quantify insulation based on KR normalized contact frequencies we used the insulation score defined in Crane et al. 2015[30], using 500kb as the square size, and 200kb as the delta size. Insulation TADs were called using the perl script at https://github.com/dekkerlab/crane-nature-2015.

2.4.3 Directionality Index

We calculated the Directionality Index (DI) as described in Dixon et al[31]. The KR normalized contact matrices were used as input to this transformation. We generated the DI score at two different resolutions: 10kb and 40kb. In both cases, we limited the scope of interaction to 2MB upstream or downstream.

2.4.4 Topological associated domains (TADs)

2.4.4.1 TAD identification.

We called TADs using three different algorithms: domain caller[31], Arrowhead[5] and insulation score[30]. All methods were run with the default parameters recommended in the user manual and with the resolution at 10kb. They produced different number of TADs because of inherent difference in their algorithms and choice of default parameters. Nevertheless, the same trend of TAD loss was observed (**Extended Data Fig. 5a,b,c**).

2.4.4.2 Identification of non-redundant TAD boundaries

TAD boundaries were previously found to be conserved between cell types[32]. Based on this assumption, we should be able to find a set of non-redundant TAD boundaries across different cell types. To identify non-redundant boundaries across all samples, we collapsed spatially identical or close TADs across all time points and replicates. To do this, we first quantile normalized the Directionality Index (DI) scores genome-wide between all twelve samples. Then we calculated the DI delta scores for each TAD boundary. DI delta score is the average difference in DI between the 4 bins (10kb per bin) downstream and 4 bins upstream of the boundary. We concatenated TAD boundaries from all samples (twelve) and sorted them by their DI delta scores in descending order. Then we picked one TAD boundary from the top of the list and remove any remaining boundary within 50Kb of the top TAD boundary. The number as

well as the stages of removed boundaries are recorded in the kept TAD boundaries as two additional columns. Then we picked the next TAD boundary on the list and repeat the same process, until the entire list was traversed. To remove unwanted spurious calls, any TAD boundary without a redundant (duplicate) call is discarded. For each non-redundant TAD boundary, it is denoted as a boundary (TRUE) in a stage if the DI delta score is greater than 200 in both replicates in that stage, otherwise is FALSE.

2.4.4.3 Dynamic TAD boundaries across stages

To determine dynamic TAD boundaries across stages, we first obtained the DI delta scores for each non-redundant TAD boundary. Then we log transformed the score and used the values as input to LIMMA[33]. LIMMA package was used to call differential boundary strength between each adjacent stage comparisons (e.g. D00 vs D02, D02 vs D05, etc). P-values were calculated with the eBayes function within LIMMA with trend parameter disabled, and were adjusted using Benjamini-Hochberg method. A TAD boundary was called dynamic if its adjusted p-value was less than 0.01 in any adjacent stage comparisons. Otherwise, it was called a stable TAD boundary. However, if the delta scores of that boundary were consistently greater than 200 in all samples, regardless of its state called in LIMMA, it was denoted as stable TAD boundary. Within dynamic TAD boundaries, we denoted the ones that were only present (TRUE) at D00 as ESC(+) TAD boundaries, those that were only FALSE at D80 as vCM(-) TAD boundaries (**Fig. 1e**).

2.4.4.4 Overlap of TAD and TAD boundaries with CTCF and repeat elements

Overlap of CTCF ChIP-seq peaks and repeat elements to TAD and/or TAD boundaries were performed using BEDtools intersect[21]. When performing intersections, the TAD boundaries were increased to ±25kb from the original TAD boundaries to account for algorithmic inaccuracies in calling boundaries. Repeat elements were RepeatMasker annotation downloaded directly from UCSC[34]. To avoid spurious results introduced by repeats with few copies in the genome, only repeats with more than 1000 copies genome-wide were tested here. The intersections were performed against each repeat name category separately, and for ESC(+) TADs, vCM(-) TADs and all stable TADs. For each repeat element, we obtained the fraction of TAD boundaries that contained that repeat. Then we calculated the fold enrichment of each repeat element at the TAD

boundaries of ESC(+) (or vCM(-)) TADs over stable TAD boundaries. The significance of the enrichment was computed using two-sided proportion test function in R. The repeat elements with fold enrichment greater than two and p-value less than 0.01 were reported as significantly enriched.

*2.4.4.5 Gene expression profile at the TAD boundaries containing repeat elements (relating to **Fig. 2a** and **Extended Fig. 6b**)*

For each of the eight repeat elements found in **Extended Fig.6a**, we kept the DNA sequences that intersected with ESC(+) TAD boundaries. Using these sequences as reference point, we generated the average RPKM normalized signal over ±50kb of the reference point with deeptools computeMatrix function[22]. Only LTR7 and HERVH-int sequences displayed high expression in the hESC.

*Mouse ES cell TAD boundaries and overlap with human TAD boundaries (**Fig.3d,e, Extended Fig.9a**)*

Mouse ESC Hi-C data from *Bonev et al*[35] was downloaded from SRA and aligned to mm10 with the same procedure as described above. The TAD boundary coordinates were converted to hg19 using UCSC liftover tool with minimum ratio of base remap of 0.1. Human TAD boundaries overlap with lifted-over mouse TAD boundaries were performed using bedtools intersect. The significance of the difference in fraction of overlap was computed using two-sided proportion test function in R.

2.4.5 Chromatin Loops

2.4.5.1 Identification of chromatin loops

Chromatin loops were called using an adapted algorithm from HICCUPS[5] at 10kb resolution, with the major difference being that only the “donut region” around the pixel were used to model the expected count.

The input for the method were raw and KR-normalized contact matrices of 10kb resolution. In preparation, distance-based mean interaction frequency was calculated from KR-normalized matrix, as well as the raw-to-KR normalization factor matrix. For each bin (B_i) in the matrix, we isolated the surrounding bins (B_s) within 40kb (both horizontal and vertical) of that bin. For each surrounding bin, we corrected its contact frequency with respect to B_i by distance-based curve, i.e.

$$B'_s = B_s \times F_i / F_s$$

F_i and F_s are the genome-wide average interaction frequencies for the distance of the test bin and surrounding bin, respectively. B'_s is the corrected frequency. Then we used the mean of the corrected frequencies from all surrounding bins as the expected contact (E_i) for the test bin. Because the expected contact was calculated using the balanced matrix, we transformed it to the raw expected contact by multiplying it with the raw-to-KR normalization factor. Finally, we calculated the probability of seeing the raw observed contact based on Poisson distribution with lambda equals expected raw contact.

In the first pass of post-processing, bins with p-value less than 0.01 and distance greater than 10kb were marked as candidate bins and kept for further analysis. Candidate bins often form clusters, because if a genomic locus is interacting with another, their nearby bins should also be physically close. Therefore, if a candidate bin is not adjacent to any other candidate bins, it is removed because it is very likely a false positive. Finally, we collapsed candidate bins within 20kb of each other into the bin with most significant p-value. The collapsed bins with p-value less than 1e-5 was designated as significant chromatin loops.

2.4.5.2 Merge redundant loops across stages

Non-redundant loops across all stages and replicates (total of 12 samples) were identified using a similar algorithm as that for non-redundant TAD boundaries in the previous section. To begin with, loop calls from all samples were concatenated and sorted by the p-values in ascending order. Loop from the top of the list was selected and any loops within 30kb of that top loop were collapsed into the top loop. This process was repeated for the next top loop until the whole list was traversed. To remove unwanted spurious calls, any loop without a redundant (replicated) loop call was discarded. For each non-redundant loop, it is denoted as TRUE in a stage if it is called as loop in that stage, otherwise FALSE.

2.4.5.3 Identification of stage-specific loops

For each non-redundant loop, Bioconductor package edgeR[27] was utilized to evaluate the dynamicity of the interaction frequencies over the stages. Raw interaction

frequencies were provided as an input to edgeR. The contrast matrix was constructed to test the difference between every adjacent stages, e.g. D02.vs.D00, D05.vs.D07, etc. The statistical tests were performed using the glmLRT function per the pacakge's instruction. The resulting p-values were adjusted using Bonferroni correction. A loop was defined as stage-specific if its adjusted p-value was less than 0.01.

2.4.5.4 K-means clustering

In search of over-represented patterns of stage-specific loop interactions over differentiation, we performed K-means clustering. The input matrix for K-means was number of contacts per million read pairs. Each row (loop) was further divided by a common factor to become a unit vector. K-means clustering was performed using K from two to ten. When K was greater than five, there was no apparent gain of new patterns. Therefore, we chose K=5 to split the clusters.

2.4.5.5 Gene set and motif enrichment testing at the loop anchors

For each loop cluster, we extended the loop anchors to 15kb from the center of the anchor to account for potential errors in finding the exact pixel of the loop. Loop anchors were overlapped with transcription start sites of genes and merged ATAC-seq peaks. The genes whose TSSs overlapped anchors from each loop cluster were provided as input to DAVID[36] to perform gene set enrichment analysis. On the other hand, The ATAC-seq peaks were analyzed for transcription factor motif enrichment using HOMER[29]. The default parameters were chosen according to the user manual.

2.4.5.6 Multi-loop anchors and overlap with transcription factors, house-keeping gene TSSs, and CTCF peaks

We took all loop anchors (In total 19938 bins, each with 10kb resolution) from the 14215 chromatin loops we identified and extend the anchors to 15kb around their centers. We calculated the number of loops that interacted with each bin as the degree of connectivity for that bin (**Fig. 4f**). Then we overlapped the anchors with the TSSs of 1999 transcription factors[37], 3804 house-keeping genes[38], and merged CTCF ChIP-seq peaks from our data. The fraction of anchors overlapping these features were plotted for anchors with different degree of connectivity (**Fig. 4g,h,i**).

2.4.5.7 CTCF motif and orientation

We scanned the CTCF motif (MA0139.1) from JASPAR database[39] and kept track of their orientation in merged CTCF ChIP-seq peaks with FIMO[40]. We overlapped the chromatin loops with the CTCF ChIP-seq peaks, and separated the loops into four categories with respect to their CTCF motif orientation located on the loop anchors. The four categories are: (1) No-CTCF: at least one anchor does not contain any CTCF ChIP-seq peak; (2) Unkown: Both anchors overlapped with CTCF peaks. At least one anchor does not contain a matched CTCF motif; (3) Incorrect: both anchors overlapped with CTCF peaks. CTCF motifs at loop anchors do not face inward; (4) Inward: both anchors overlapped with CTCF peaks. CTCF motifs at loop anchors face inward.

2.4.5.8 loop location relative to TADs

We assigned chromatin loops into three categories with respect to their location to TAD boundaries. (1) TAD-corner: a loop is located within 20kb of any TAD corner. (2) Intra-TAD: a loop is inside of any TAD. (3) inter-TAD: a loop that is outside of all TADs. When there are conflicts, the priority of assignment is (1)>(2)>(3).

2.5 HERV-H related analysis

2.5.1 Directionality index profile on HERV-Hs transcribed at different levels

To investigate the boundary activity of HERV-Hs that transcribed at different levels, we visualized the Directionality Index (DI) scores at 10kb resolution. We sorted the HERV-H loci by their expression RPKMs in the hESC (day 0) in descending order and assign every 50 loci to a bin. For each bin, we generated the aggregated median DI score profile 70kb surrounding the HERV-H loci. The same DI scores were computed at day 2 and day 5. As comparison, we also plotted DI score profile for genes that expressed with similar RPKMs as the top-tier HERV-Hs in the hESC, and down-regulated at least 2 fold after differentiation (**Extended Data. Fig7. a**).

*2.5.2 Directionality index profile on HERV-Hs in iPSCs and H1-derived cells (relating to **Extended Data. Fig7. b**).*

The Hi-C data for human iPSCs were each from Greenwald *et al*[41]. The Hi-C data from H1 ESC and its derived cells were from Dixon *et al*[31,32]. The data were downloaded and processed the same way as described above. Because DI scores are dependent on the

sequencing depth, the H1ESC DI profile may look less pronounced comparing to the iPSC.

*2.5.3 ChIP-seq enrichment analysis at TAD boundary-associating HERV-H loci (relating to **Extended Data. Fig. 7c**)*

To investigate which protein(s) may be responsible for TAD boundary formation at top 50 transcribing HERV-H loci, we downloaded hESC ChIP-seq bigWig files from ENCODE (primary) and other sources in the public domain (prefixed with the SRA ID if downloaded from other than ENCODE)[42]. For each ChIP-seq data, we computed the average bigWig signal on the top 50 HERV-Hs with *bigWigAverageOverBed* command from UCSC, and the average bigWig signal on HERV-Hs ranking from 51 to 300. Then we calculated the fold change with these two averages.

*2.5.4 HERV-H LTR type and divergence analysis (relating to **Extended Data. Fig. 9**)*

To investigate the sequence features of the LTRs flanking TAD boundary associated HERV-Hs and non-boundary associated HERV-Hs, we examined their presence, length and sequence divergence. The LTRs for HERV-H include LTR7, LTR7A, LTR7B, LTR7C and LTR7Y. We downloaded the annotations for those LTRs from repeatMasker and assigned them to each HERV-H-int sequence by base overlap. The HERV-H loci were ordered by their expression levels and binned as described above (the top 50 were enriched at TAD boundaries).

Sequence divergence between the 5' LTR and 3'LTR sequences was used to measure the insertion time for retrotransposons. Based on the retrotransposon insertion mechanism, at the time of insertion, 5' and 3' LTR sequences should be identical; the LTR sequences will each accumulate mutations over time, and thus their sequence differences could be used as a proxy to their ages. We followed the steps outlined in https://github.com/SIWLab/Lab_Info/wiki/Ageing-LTR-insertions to align each LTR pairs and calculated the sequence divergences[43,44].

2.6 GWAS analysis:

2.6.1 GWAS lead SNP and SNPs in linkage disequilibrium

Human GWAS association data was downloaded from human GWAS catalog (https://www.ebi.ac.uk/gwas/) as of Sept 28th, 2017. Heart-related traits were selected for analysis. These traits are: "Atrial fibrillation", "Cardiac repolarization", "Coronary heart disease", "Conotruncal heart defects", "Congenital left-sided heart lesions", "Congenital heart disease", "Congenital heart malformation", "Heart failure", "Mortality in heart failure", "PR interval", "QRS duration", "QT interval", and "Sudden cardiac arrest". For all lead SNPs, we obtained the SNPs in linkage disequilibrium (LD) using SNAP[45] with default parameters (SNP data set was set to "1000 Genome Pilot1", Population panel set to "CEU", r square threshold > 0.8 and distance limit equals 500). The output from SNAP was in hg18 coordinates, and therefore was lifted over to hg19 with UCSC liftover tool.

2.6.2 Binomial test for significance of enrichment of heart-traits GWAS SNPs within chromatin loops

The LD SNPs were overlapped with merged ATAC-seq peaks within chromatin loop anchors using bedtools. The chromatin loops used here were loops present in cardiomyocyte differentiation day 2 – day 80. For each GWAS trait, the significance of enrichment was computed using a binomial test as described in Maurano *et al*[46]. The total number of tests was total number of lead SNPs for each trait. The number of successes was the number of lead SNPs that were overlapped with open chromatin region within the loop anchor. For each lead SNP, if itself or any of the LD SNPs were overlapped with an open chromatin region within chromatin loop anchors, that lead SNP was considered a positive hit. The hypothesized probability of success is the width of ATAC-seq peaks within chromatin loop anchors divided by the length of the genome.

Gene Ontology Enrichment testing of the genes.

We collected all of the SNPs overlapped with chromatin loop anchors and ATAC-seq peaks in the previous section, and identified the candidate target gene at the other end of the chromatin loop. Using this list of gene as input, we performed GO enrichment analysis using PANTHER[47].

2.7 General data processing and plots

Most of the described data processing steps (statistical tests, clustering, plotting, etc) were performed in python 3.4.5 (www.python.org) and statistical computing environment R 3.4.3 (www.r-project.org). Boxplots were made with ggplot2 (https://cran.r-project.org/web/packages/ggplot2). The elements of the boxplot are: center line, median; box limit, upper and lower quartiles; whiskers, 1.5x interquartile range; points, outliers.

3 Data and script availability:

All sequencing datasets have been deposited on GEO with the accession number **GSE116862**.

Scripts are available at https://github.com/shawnzhangyx/cvdc_scripts.

Additional References:

1 Veevers, J. *et al.* Cell-Surface Marker Signature for Enrichment of Ventricular Cardiomyocytes Derived from Human Embryonic Stem Cells. *Stem cell reports*, doi:10.1016/j.stemcr.2018.07.007 (2018).

2 Lian, X. *et al.* Directed cardiomyocyte differentiation from human pluripotent stem cells by modulating Wnt/beta-catenin signaling under fully defined conditions. *Nature protocols* **8**, 162-175, doi:10.1038/nprot.2012.150 (2013).

3 Hashem, S. I. *et al.* Brief Report: Oxidative Stress Mediates Cardiomyocyte Apoptosis in a Human Model of Danon Disease and Heart Failure. *Stem cells* **33**, 2343-2350, doi:10.1002/stem.2015 (2015).

4 Milani, P. *et al.* Cell freezing protocol suitable for ATAC-Seq on motor neurons derived from human induced pluripotent stem cells. *Scientific reports* **6**, 25474, doi:10.1038/srep25474 (2016).

5 Rao, S. S. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665-1680, doi:10.1016/j.cell.2014.11.021 (2014).

6 Shen, Y. *et al.* A map of the cis-regulatory sequences in the mouse genome. *Nature* **488**, 116-120, doi:10.1038/nature11243 (2012).

7 Gorkin, D. *et al.* Systematic mapping of chromatin state landscapes during mouse development. *bioRxiv* (2017).

8 Jolma, A. *et al.* DNA-binding specificities of human transcription factors. *Cell* **152**, 327-339, doi:10.1016/j.cell.2012.12.009 (2013).

9 Yan, J. *et al.* Transcription factor binding in human cells occurs in dense clusters formed around cohesin anchor sites. *Cell* **154**, 801-813, doi:10.1016/j.cell.2013.07.034 (2013).

10 Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature methods* **10**, 1213-1218, doi:10.1038/nmeth.2688 (2013).

11 Liu, Z. *et al.* Efficient CRISPR/Cas9-Mediated Versatile, Predictable, and Donor-Free Gene Knockout in Human Pluripotent Stem Cells. *Stem cell reports* **7**, 496-507, doi:10.1016/j.stemcr.2016.07.021 (2016).

12 Haeussler, M. *et al.* Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR. *Genome biology* **17**, 148, doi:10.1186/s13059-016-1012-2 (2016).

13 Wrighton, P. J. *et al.* Signals from the surface modulate differentiation of human pluripotent stem cells through glycosaminoglycans and integrins. *Proc Natl Acad Sci U S A* **111**, 18126-18131, doi:10.1073/pnas.1409525111 (2014).

14 Kim, S., Kim, D., Cho, S. W., Kim, J. & Kim, J. S. Highly efficient RNA-guided genome editing in human cells via delivery of purified Cas9 ribonucleoproteins. *Genome research* **24**, 1012-1019, doi:10.1101/gr.171322.113 (2014).

15 Zuris, J. A. *et al.* Cationic lipid-mediated delivery of proteins enables efficient protein-based genome editing in vitro and in vivo. *Nature biotechnology* **33**, 73-80, doi:10.1038/nbt.3081 (2015).

16 Santos, D. P., Kiskinis, E., Eggan, K. & Merkle, F. T. Comprehensive Protocols for CRISPR/Cas9-based Gene Editing in Human Pluripotent Stem Cells. *Current protocols in stem cell biology* **38**, 5B 6 1-5B 6 60, doi:10.1002/cpsc.15 (2016).

17 https://github.com/ENCODE-DCC/chip-seq-pipeline.

18 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).

19 http://broadinstitute.github.io/picard/.

20 Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome biology* **9**, R137, doi:10.1186/gb-2008-9-9-r137 (2008).

21 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842, doi:10.1093/bioinformatics/btq033 (2010).

22 Ramirez, F., Dundar, F., Diehl, S., Gruning, B. A. & Manke, T. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res* **42**, W187-191, doi:10.1093/nar/gku365 (2014).

23 https://github.com/kundajelab/atac_dnase_pipelines.

24 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature methods* **9**, 357-359, doi:10.1038/nmeth.1923 (2012).

25 Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21, doi:10.1093/bioinformatics/bts635 (2013).

26 Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923-930, doi:10.1093/bioinformatics/btt656 (2014).

27 Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139-140, doi:10.1093/bioinformatics/btp616 (2010).

28 Durand, N. C. *et al.* Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Syst* **3**, 99-101, doi:10.1016/j.cels.2015.07.012 (2016).
29 Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38**, 576-589, doi:10.1016/j.molcel.2010.05.004 (2010).
30 Crane, E. *et al.* Condensin-driven remodelling of X chromosome topology during dosage compensation. *Nature* **523**, 240-244, doi:10.1038/nature14450 (2015).
31 Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376-380, doi:10.1038/nature11082 (2012).
32 Dixon, J. R. *et al.* Chromatin architecture reorganization during stem cell differentiation. *Nature* **518**, 331-336, doi:10.1038/nature14222 (2015).
33 Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **43**, e47, doi:10.1093/nar/gkv007 (2015).
34 Casper, J. *et al.* The UCSC Genome Browser database: 2018 update. *Nucleic Acids Res* **46**, D762-D769, doi:10.1093/nar/gkx1020 (2018).
35 Bonev, B. *et al.* Multiscale 3D Genome Rewiring during Mouse Neural Development. *Cell* **171**, 557-572 e524, doi:10.1016/j.cell.2017.09.043 (2017).
36 Huang da, W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* **37**, 1-13, doi:10.1093/nar/gkn923 (2009).
37 Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A. & Luscombe, N. M. A census of human transcription factors: function, expression and evolution. *Nat Rev Genet* **10**, 252-263, doi:10.1038/nrg2538 (2009).
38 Eisenberg, E. & Levanon, E. Y. Human housekeeping genes, revisited. *Trends Genet* **29**, 569-574, doi:10.1016/j.tig.2013.05.010 (2013).
39 Sandelin, A., Alkema, W., Engstrom, P., Wasserman, W. W. & Lenhard, B. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res* **32**, D91-94, doi:10.1093/nar/gkh012 (2004).
40 Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017-1018, doi:10.1093/bioinformatics/btr064 (2011).
41 Greenwald, W. W. *et al.* Integration of phased Hi-C and molecular phenotype data to study genetic and epigenetic effects on chromatin looping. *bioRxiv* (2018).
42 Xu, C. & Corces, V. G. Nascent DNA methylome mapping reveals inheritance of hemimethylation at CTCF/cohesin sites. *Science* **359**, 1166-1170, doi:10.1126/science.aan5480 (2018).
43 Hu, T. T. *et al.* The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. *Nat Genet* **43**, 476-481, doi:10.1038/ng.807 (2011).
44 Slotte, T. *et al.* The Capsella rubella genome and the genomic consequences of rapid mating system evolution. *Nat Genet* **45**, 831-835, doi:10.1038/ng.2669 (2013).
45 Johnson, A. D. *et al.* SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics* **24**, 2938-2939, doi:10.1093/bioinformatics/btn564 (2008).
46 Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190-1195, doi:10.1126/science.1222794 (2012).

47 Mi, H. *et al.* PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic Acids Res* **45**, D183-D189, doi:10.1093/nar/gkw1138 (2017).

Figure 1

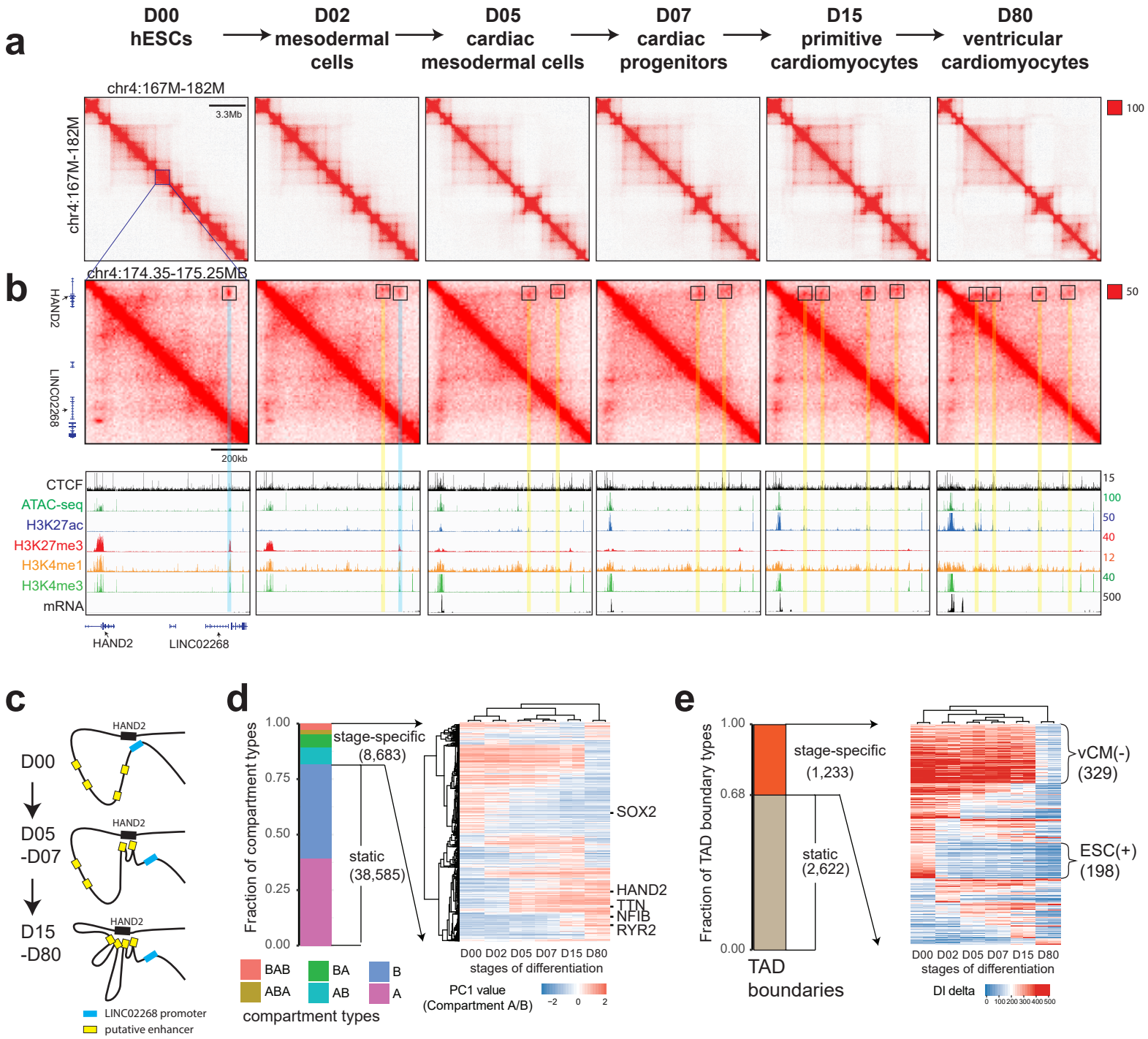

a
D00 hESCs
D02 mesodermal cells
D05 cardiac mesodermal cells
D07 cardiac progenitors
D15 primitive cardiomyocytes
D80 ventricular cardiomyocytes
chr4:167M-182M
3.3Mb
100
b
chr4:174.35-175.25MB
HAND2
LINC02268
200kb
50
CTCF
ATAC-seq
H3K27ac
H3K27me3
H3K4me1
H3K4me3
mRNA
c
D00
D05 -D07
D15 -D80
HAND2
LINC02268 promoter
putative enhancer
d
Fraction of compartment types
stage-specific (8,683)
static (38,585)
BAB
BA
B
ABA
AB
A
compartment types
SOX2
HAND2
TTN
NFIB
RYR2
D00 D02 D05 D07 D15 D80
stages of differentiation
PC1 value (Compartment A/B)
e
Fraction of TAD boundary types
stage-specific (1,233)
static (2,622)
TAD boundaries
vCM(-) (329)
ESC(+) (198)
DI delta

Figure 2

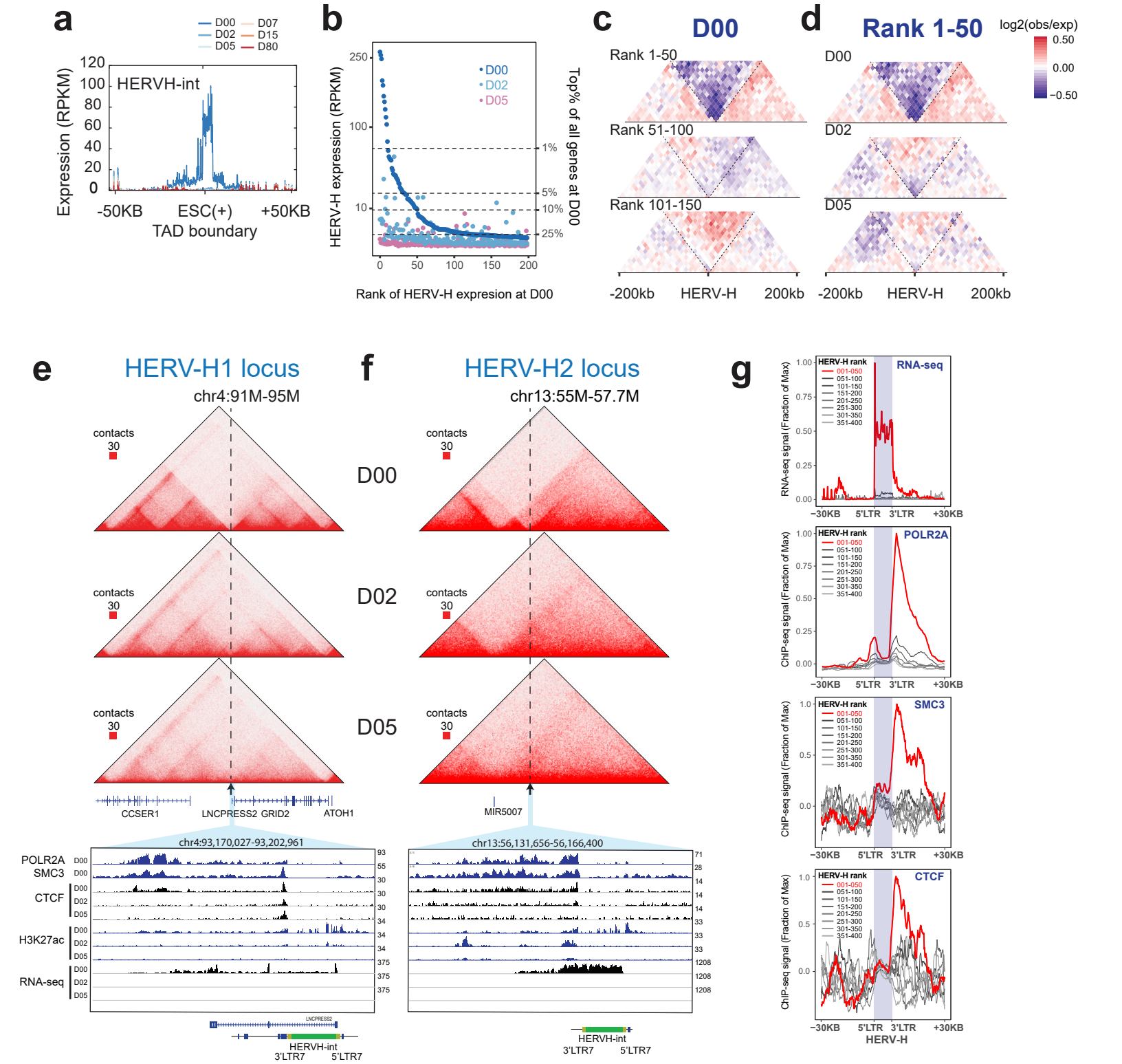

a
D00
D02
D05
D07
D15
D80
HERVH-int
Expression (RPKM)
-50KB
ESC(+)
+50KB
TAD boundary
b
D00
D02
D05
HERV-H expression (RPKM)
Top% of all genes at D00
1%
5%
10%
25%
Rank of HERV-H expresion at D00
c
D00
Rank 1-50
Rank 51-100
Rank 101-150
-200kb
HERV-H
200kb
d
Rank 1-50
log2(obs/exp)
D00
D02
D05
e
HERV-H1 locus
chr4:91M-95M
contacts
30
CCSER1
LNCPRESS2
GRID2
ATOH1
chr4:93,170,027-93,202,961
POLR2A
SMC3
CTCF
H3K27ac
RNA-seq
HERVH-int
3'LTR7
5'LTR7
f
HERV-H2 locus
chr13:55M-57.7M
MIR5007
chr13:56,131,656-56,166,400
g
RNA-seq
POLR2A
SMC3
CTCF
HERV-H rank
001-050
051-100
101-150
151-200
201-250
251-300
301-350
351-400
RNA-seq signal (Fraction of Max)
ChIP-seq signal (Fraction of Max)
-30KB
5'LTR
3'LTR
+30KB
HERV-H

Figure 3

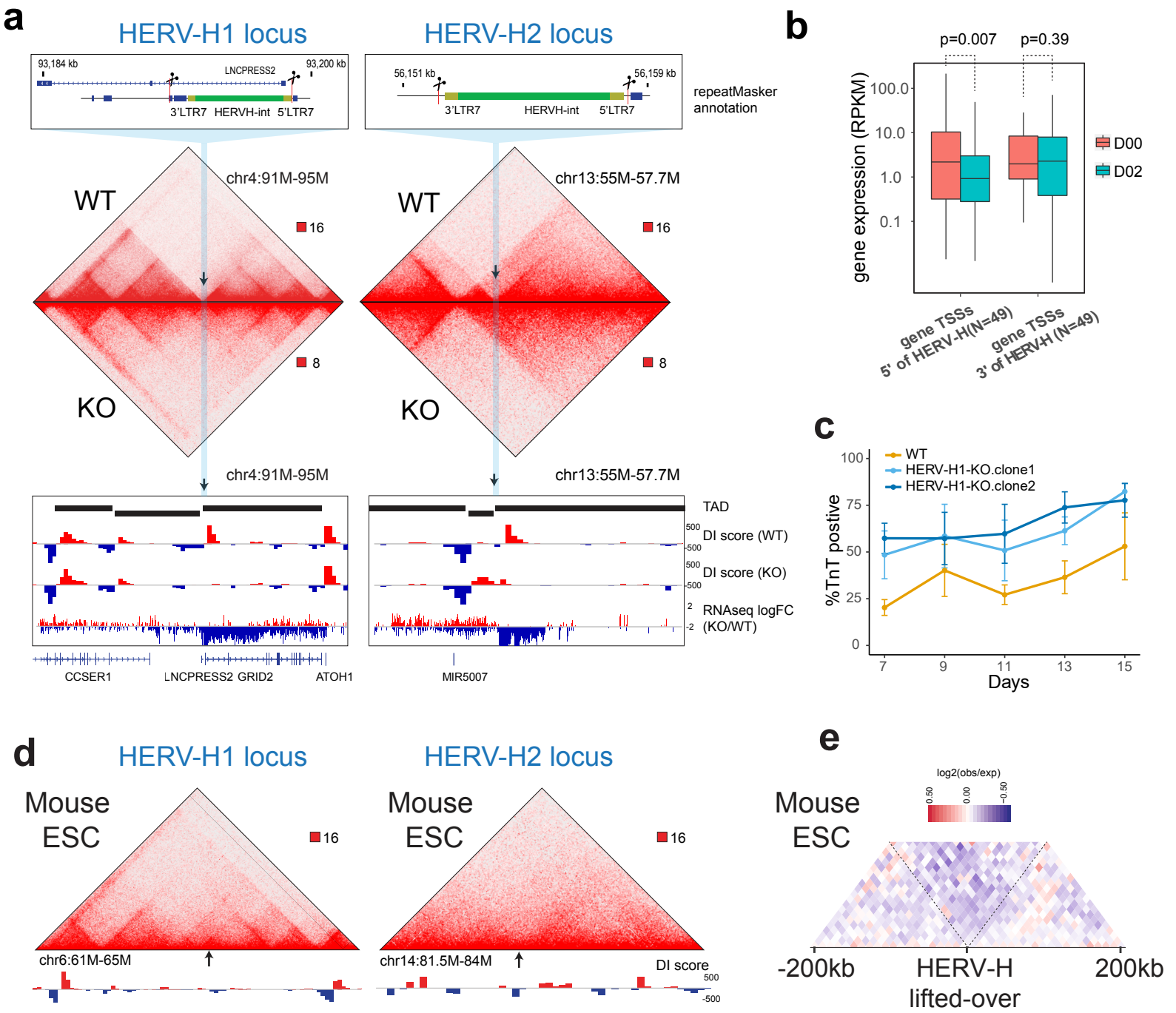

a
HERV-H1 locus
HERV-H2 locus
93,184 kb
LNCPRESS2
93,200 kb
3'LTR7 HERVH-int 5'LTR7
56,151 kb
56,159 kb
3'LTR7 HERVH-int 5'LTR7
repeatMasker annotation
WT
KO
chr4:91M-95M
chr13:55M-57.7M
16
8
TAD
DI score (WT)
DI score (KO)
RNAseq logFC (KO/WT)
CCSER1
LNCPRESS2 GRID2
ATOH1
MIR5007
b
p=0.007
p=0.39
gene expression (RPKM)
100.0
10.0
1.0
0.1
D00
D02
gene TSSs 5' of HERV-H(N=49)
gene TSSs 3' of HERV-H (N=49)
c
WT
HERV-H1-KO.clone1
HERV-H1-KO.clone2
%TnT postive
Days
d
HERV-H1 locus
HERV-H2 locus
Mouse ESC
chr6:61M-65M
chr14:81.5M-84M
DI score
e
log2(obs/exp)
-200kb
HERV-H lifted-over
200kb

Figure 4

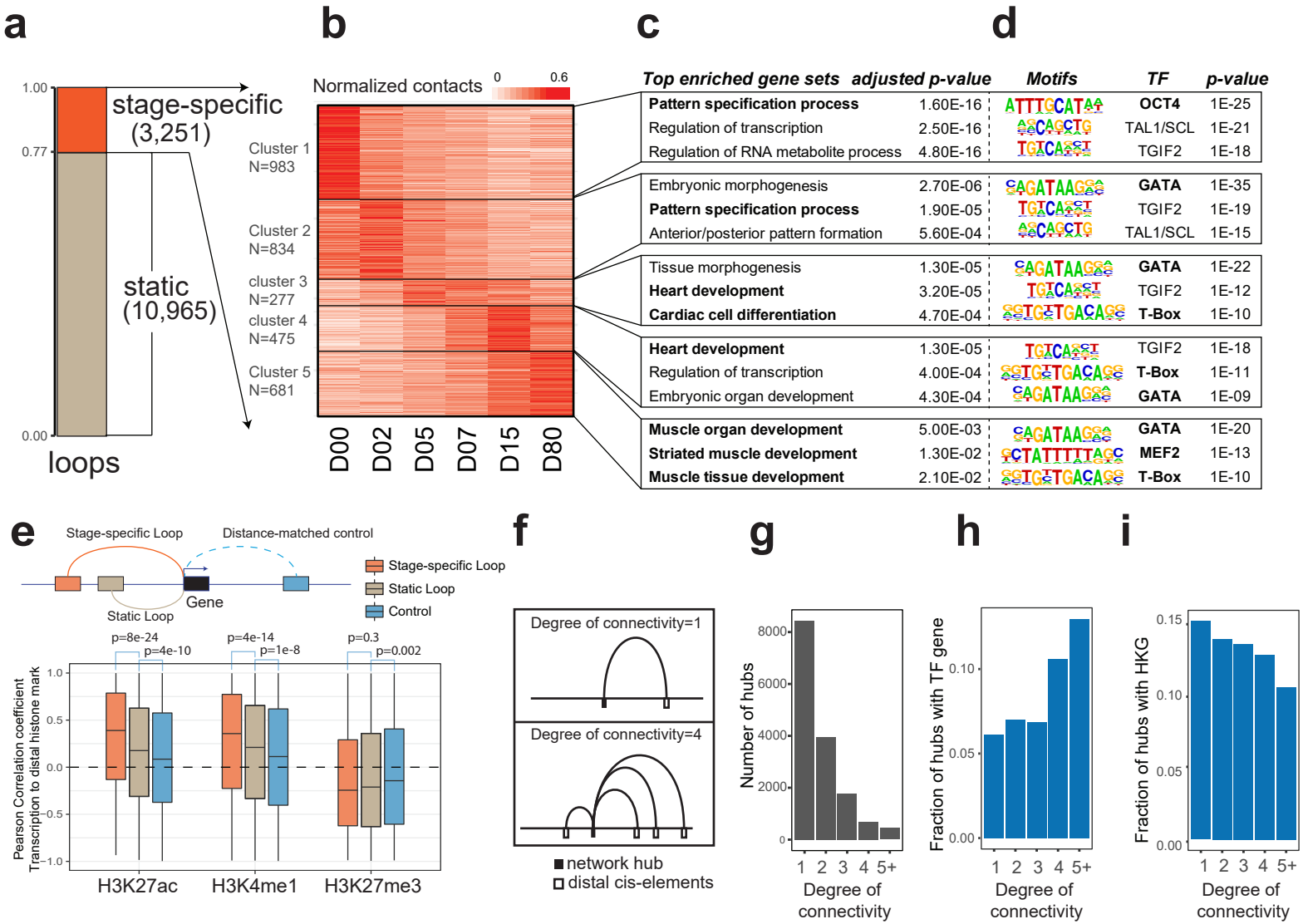
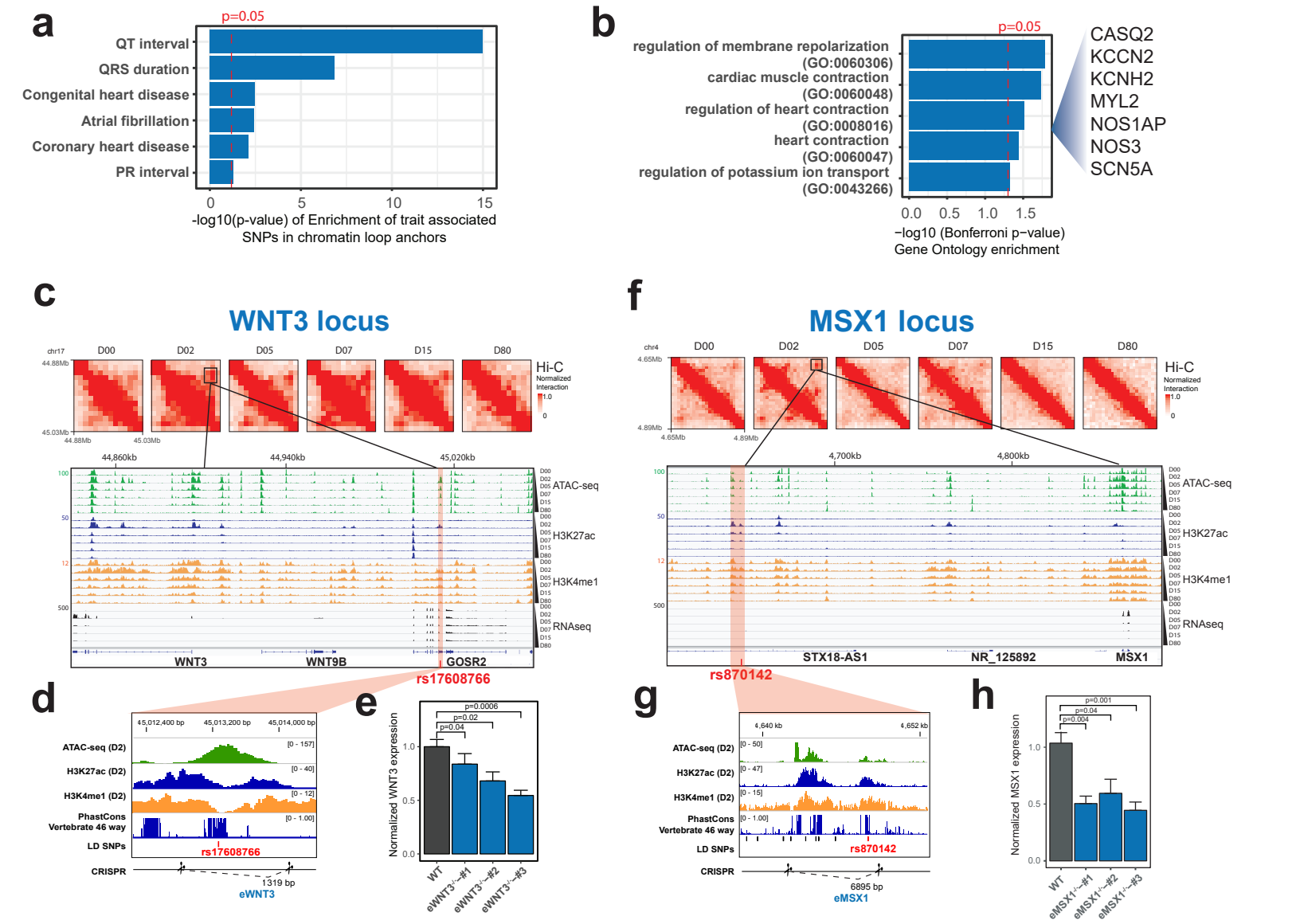
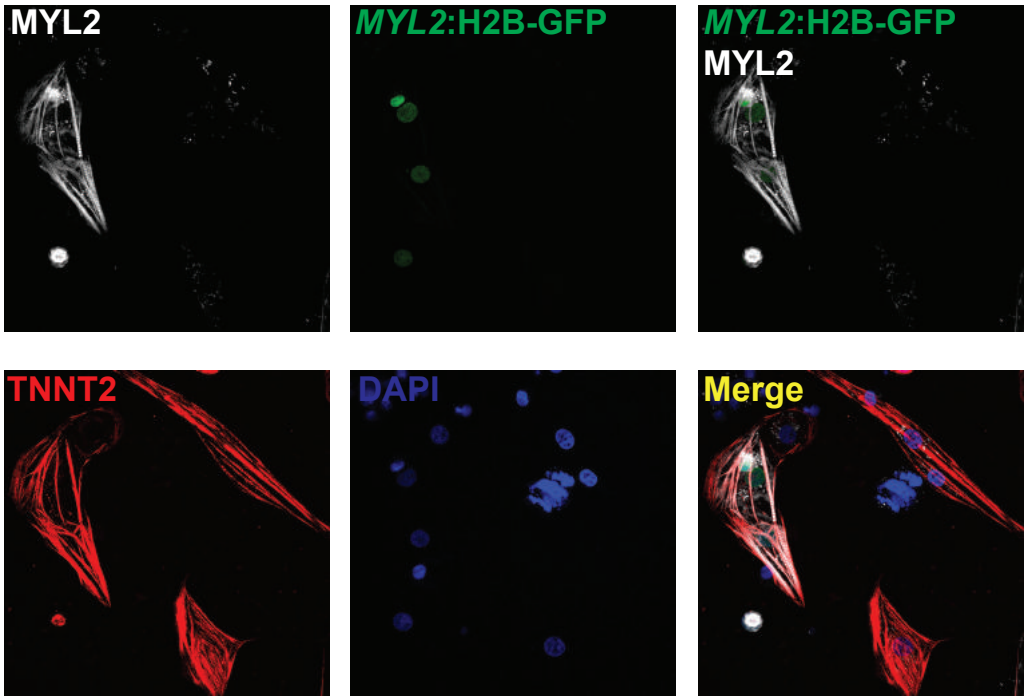

a
1.00
0.77
0.00
stage-specific
(3,251)
static
(10,965)
loops
b
Normalized contacts
0
0.6
Cluster 1
N=983
Cluster 2
N=834
cluster 3
N=277
cluster 4
N=475
Cluster 5
N=681
D00
D02
D05
D07
D15
D80
c
d
Top enriched gene sets
adjusted p-value
Motifs
TF
p-value
Pattern specification process 1.60E-16 OCT4 1E-25
Regulation of transcription 2.50E-16 TAL1/SCL 1E-21
Regulation of RNA metabolite process 4.80E-16 TGIF2 1E-18
Embryonic morphogenesis 2.70E-06 GATA 1E-35
Pattern specification process 1.90E-05 TGIF2 1E-19
Anterior/posterior pattern formation 5.60E-04 TAL1/SCL 1E-15
Tissue morphogenesis 1.30E-05 GATA 1E-22
Heart development 3.20E-05 TGIF2 1E-12
Cardiac cell differentiation 4.70E-04 T-Box 1E-10
Heart development 1.30E-05 TGIF2 1E-18
Regulation of transcription 4.00E-04 T-Box 1E-11
Embryonic organ development 4.30E-04 GATA 1E-09
Muscle organ development 5.00E-03 GATA 1E-20
Striated muscle development 1.30E-02 MEF2 1E-13
Muscle tissue development 2.10E-02 T-Box 1E-10
e
Stage-specific Loop
Distance-matched control
Gene
Static Loop
Stage-specific Loop
Static Loop
Control
p=8e-24
p=4e-10
p=4e-14
p=1e-8
p=0.3
p=0.002
Pearson Correlation coefficient
Transcription to distal histone mark
1.0
0.5
0.0
−0.5
−1.0
H3K27ac
H3K4me1
H3K27me3
f
Degree of connectivity=1
Degree of connectivity=4
network hub
distal cis-elements
g
Number of hubs
8000
6000
4000
2000
0
1 2 3 4 5+
Degree of connectivity
h
Fraction of hubs with TF gene
0.10
0.05
0.00
1 2 3 4 5+
Degree of connectivity
i
Fraction of hubs with HKG
0.15
0.10
0.05
0.00
1 2 3 4 5+
Degree of connectivity

Figure 5

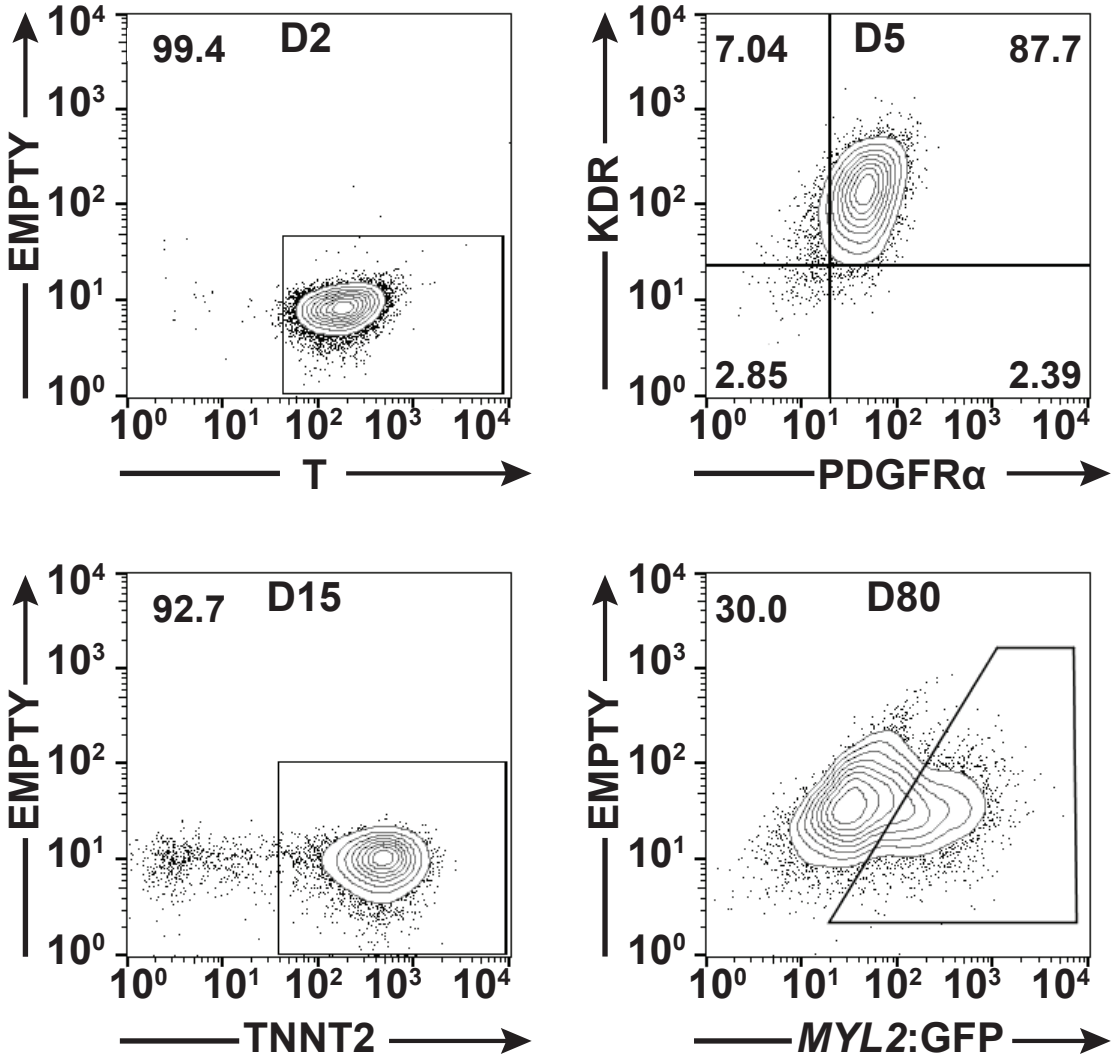

a
p=0.05
QT interval
QRS duration
Congenital heart disease
Atrial fibrillation
Coronary heart disease
PR interval
0
5
10
15
-log10(p-value) of Enrichment of trait associated SNPs in chromatin loop anchors
b
p=0.05
regulation of membrane repolarization (GO:0060306)
cardiac muscle contraction (GO:0060048)
regulation of heart contraction (GO:0008016)
heart contraction (GO:0060047)
regulation of potassium ion transport (GO:0043266)
CASQ2
KCCN2
KCNH2
MYL2
NOS1AP
NOS3
SCN5A
0.0
0.5
1.0
1.5
−log10 (Bonferroni p−value) Gene Ontology enrichment
c
WNT3 locus
chr17
D00
D02
D05
D07
D15
D80
Hi-C
Normalized Interaction
44,860kb
44,940kb
45,020kb
ATAC-seq
H3K27ac
H3K4me1
RNAseq
WNT3
WNT9B
GOSR2
rs17608766
d
45,012,400 bp
45,013,200 bp
45,014,000 bp
ATAC-seq (D2)
H3K27ac (D2)
H3K4me1 (D2)
PhastCons Vertebrate 46 way
LD SNPs
CRISPR
rs17608766
1319 bp
eWNT3
e
p=0.0006
p=0.02
p=0.04
Normalized WNT3 expression
WT
eWNT3−/−#1
eWNT3−/−#2
eWNT3−/−#3
f
MSX1 locus
chr4
4,700kb
4,800kb
STX18-AS1
NR_125892
MSX1
rs870142
g
4,640 kb
4,652 kb
rs870142
6895 bp
eMSX1
h
p=0.001
p=0.04
p=0.004
Normalized MSX1 expression
eMSX1−/−#1
eMSX1−/−#2
eMSX1−/−#3

Extended Data Figure 1

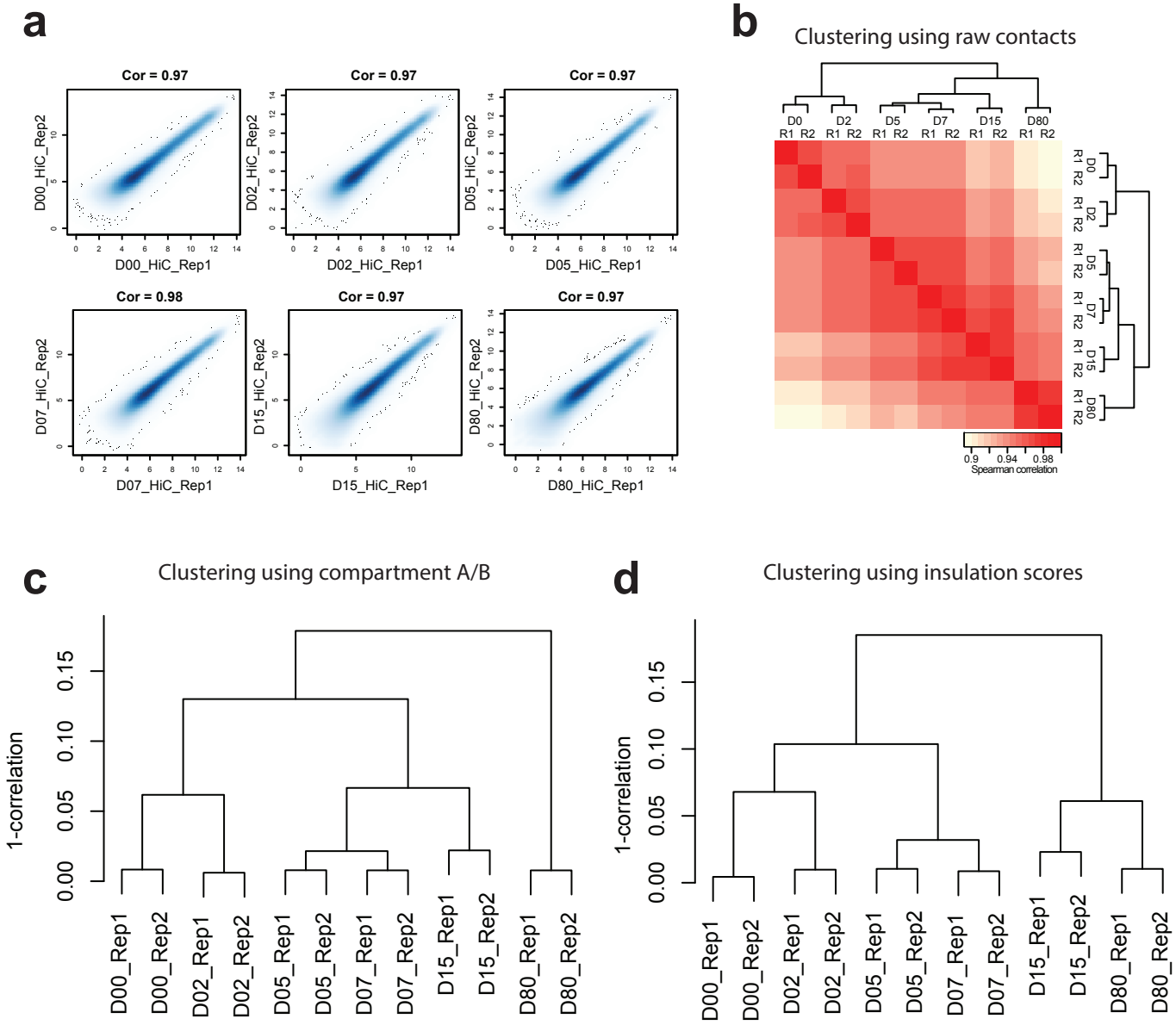
a



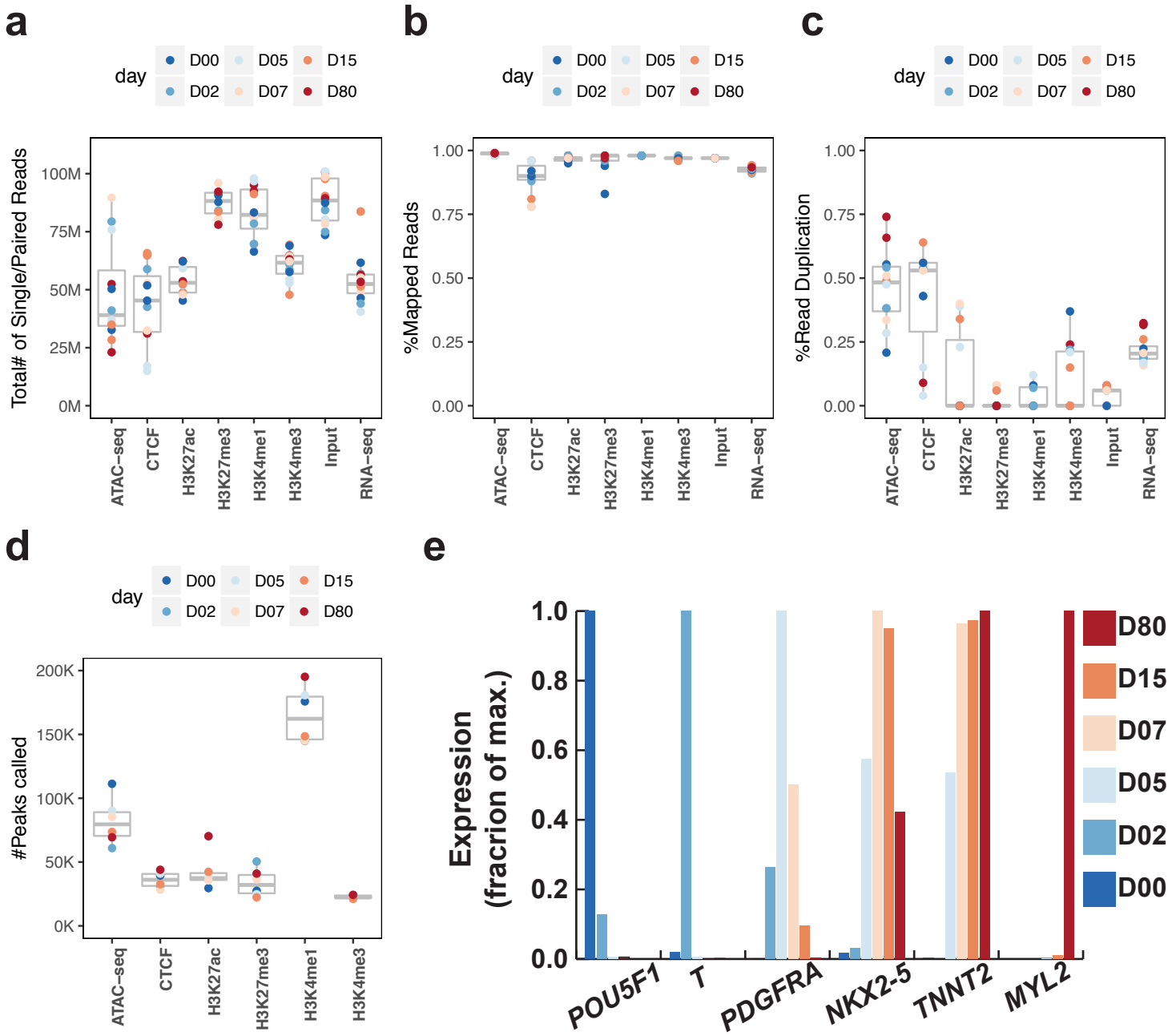
b



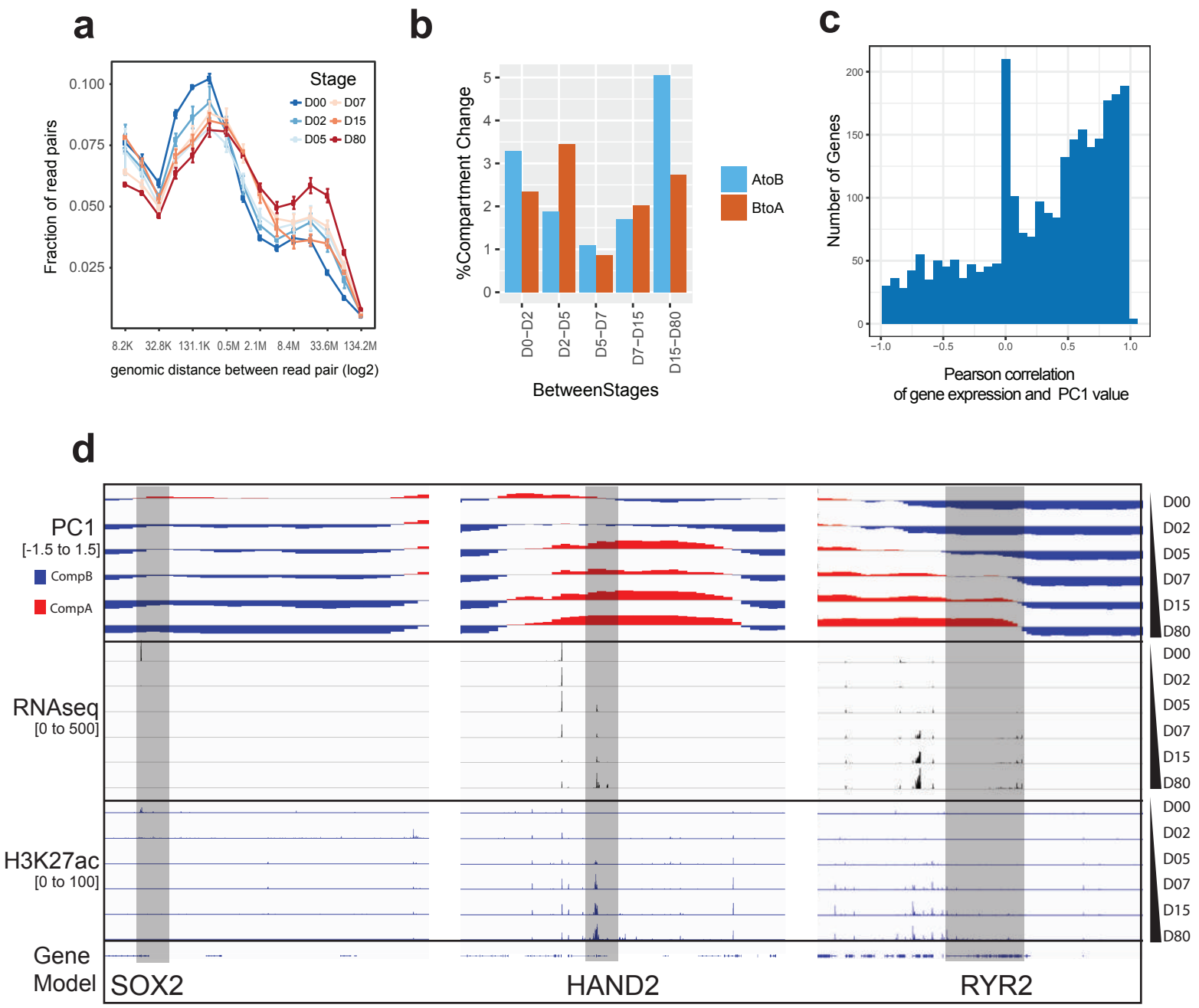
Extended Data Figure 2



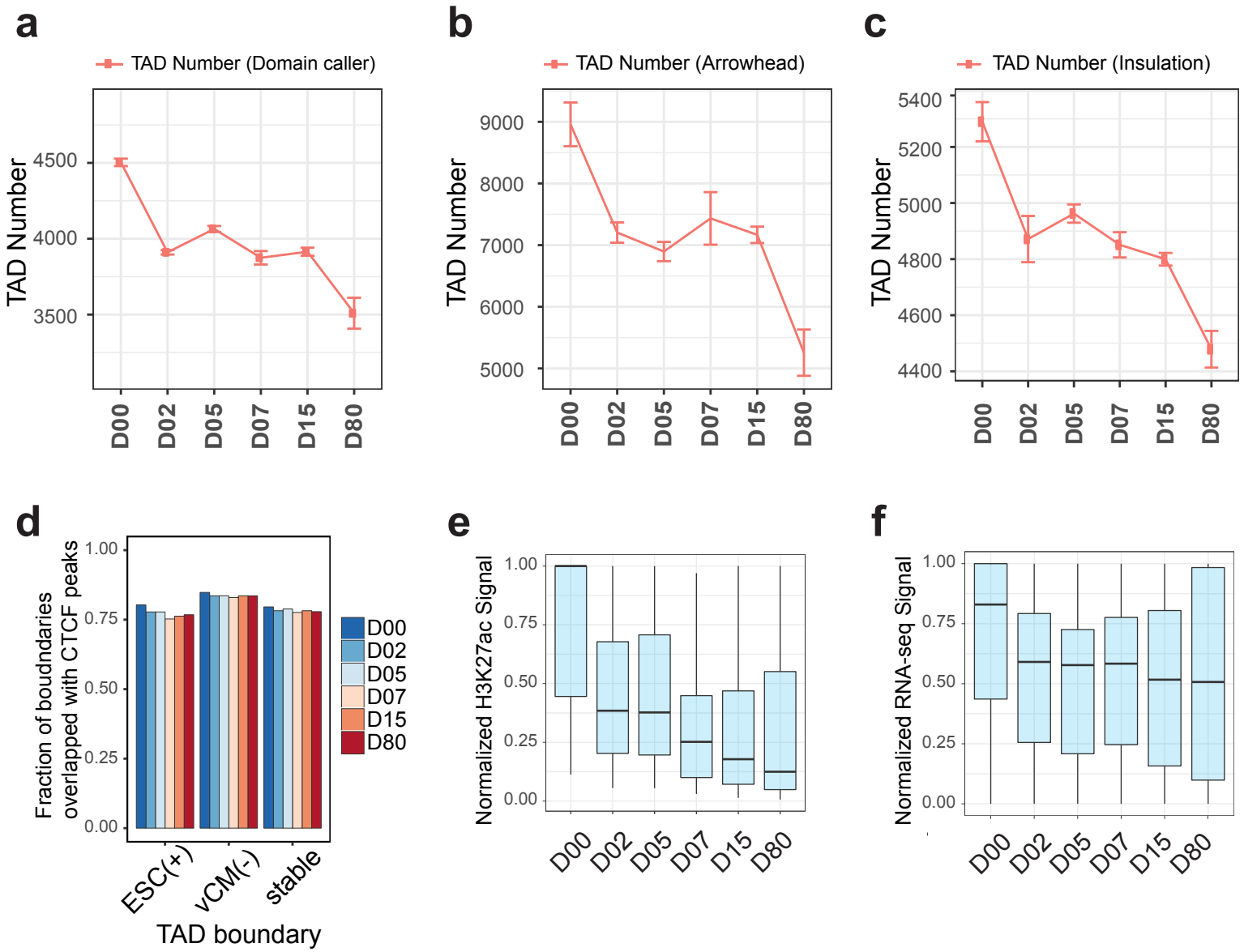
Extended Data Figure 3



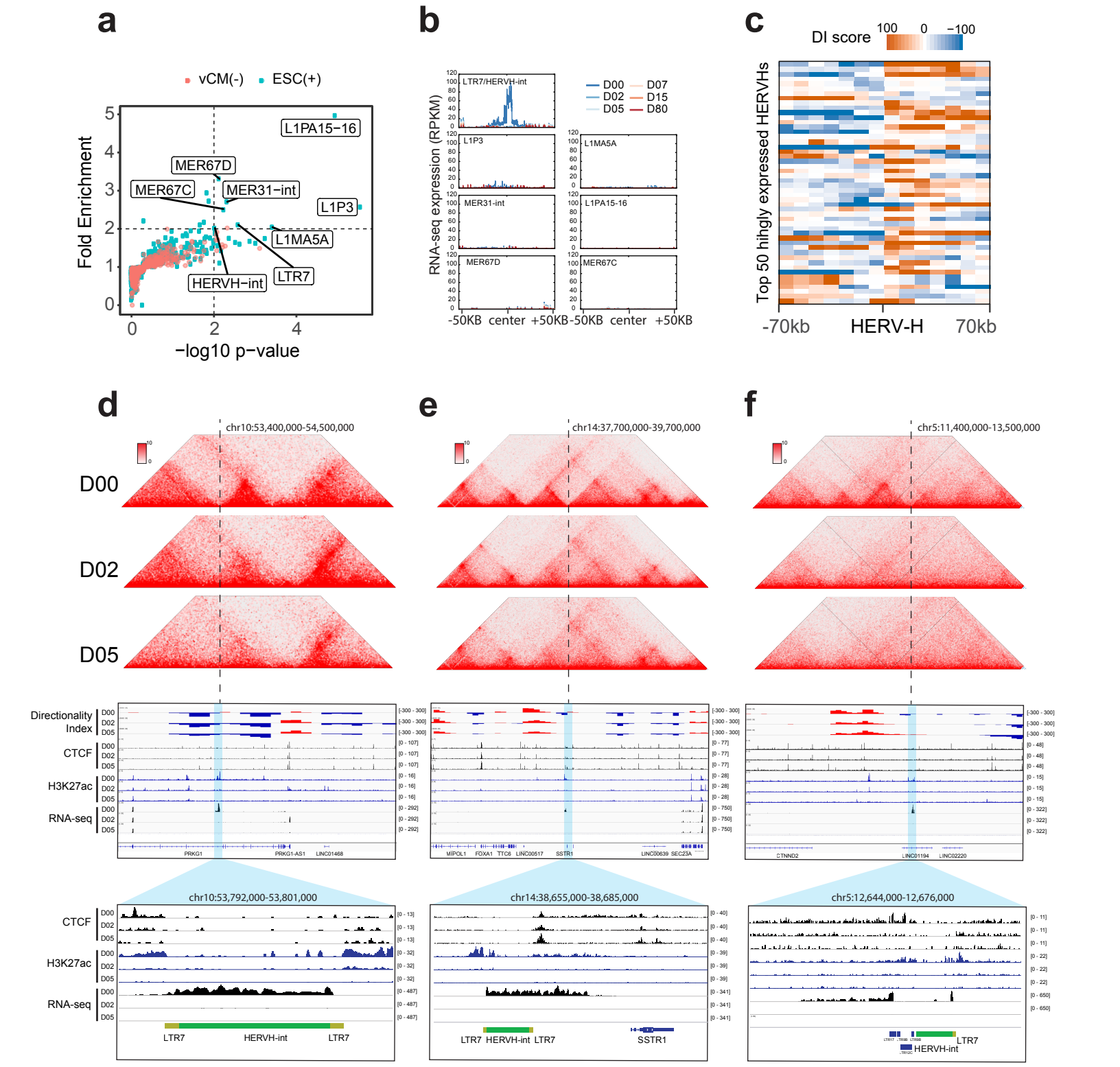
Extended Data Figure 4



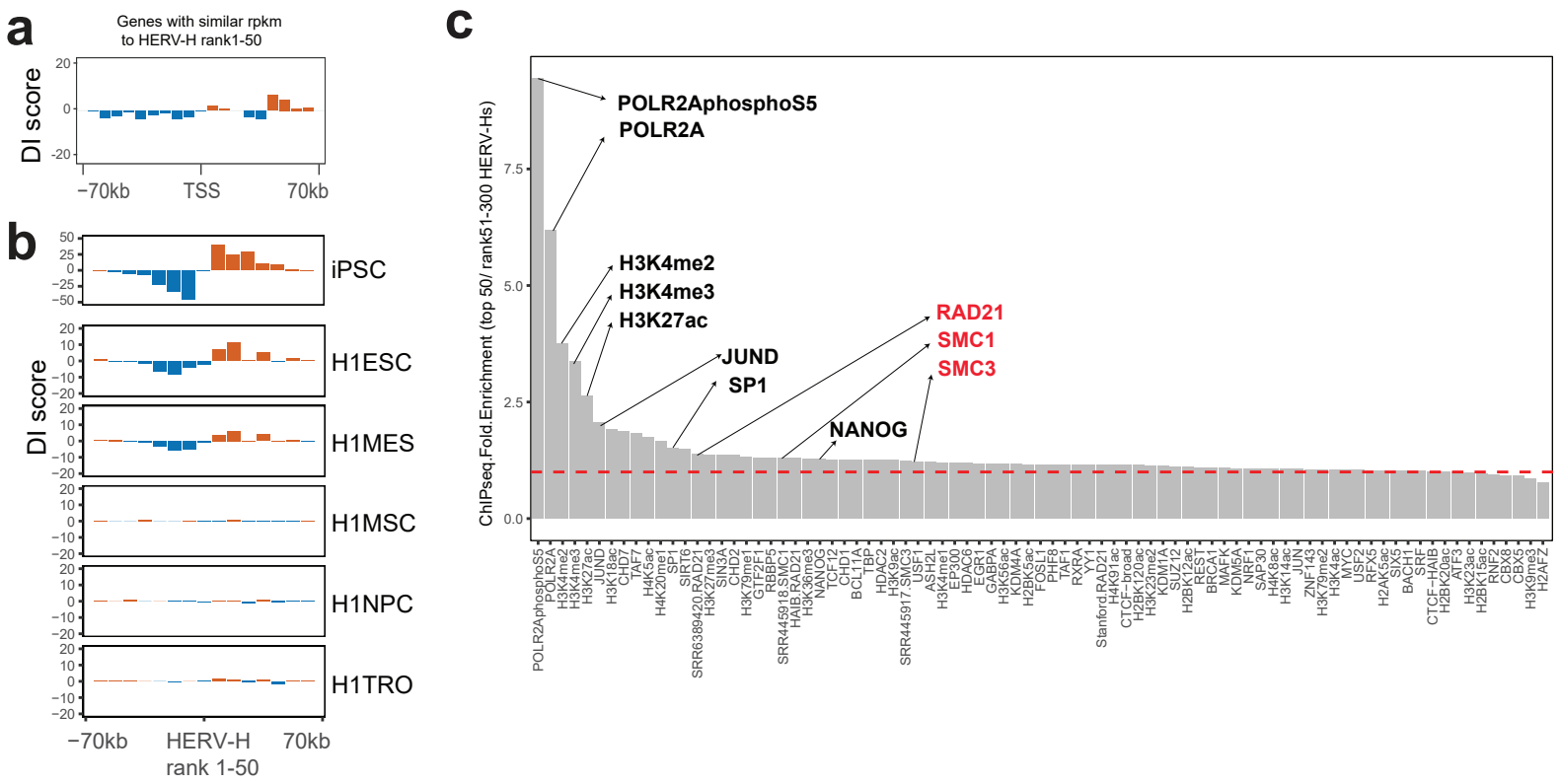
Extended Data Figure 5



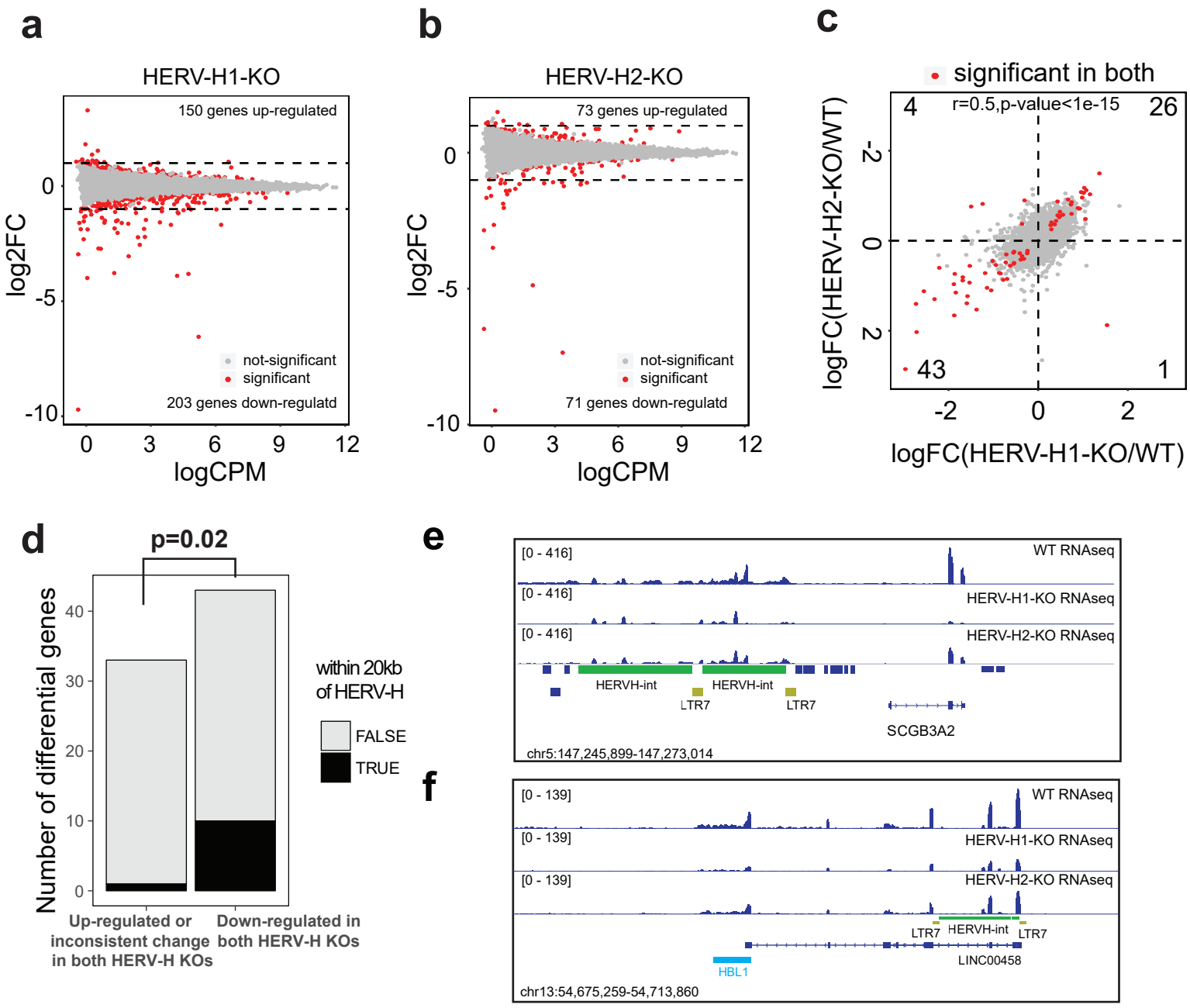
Extended Data Figure 6



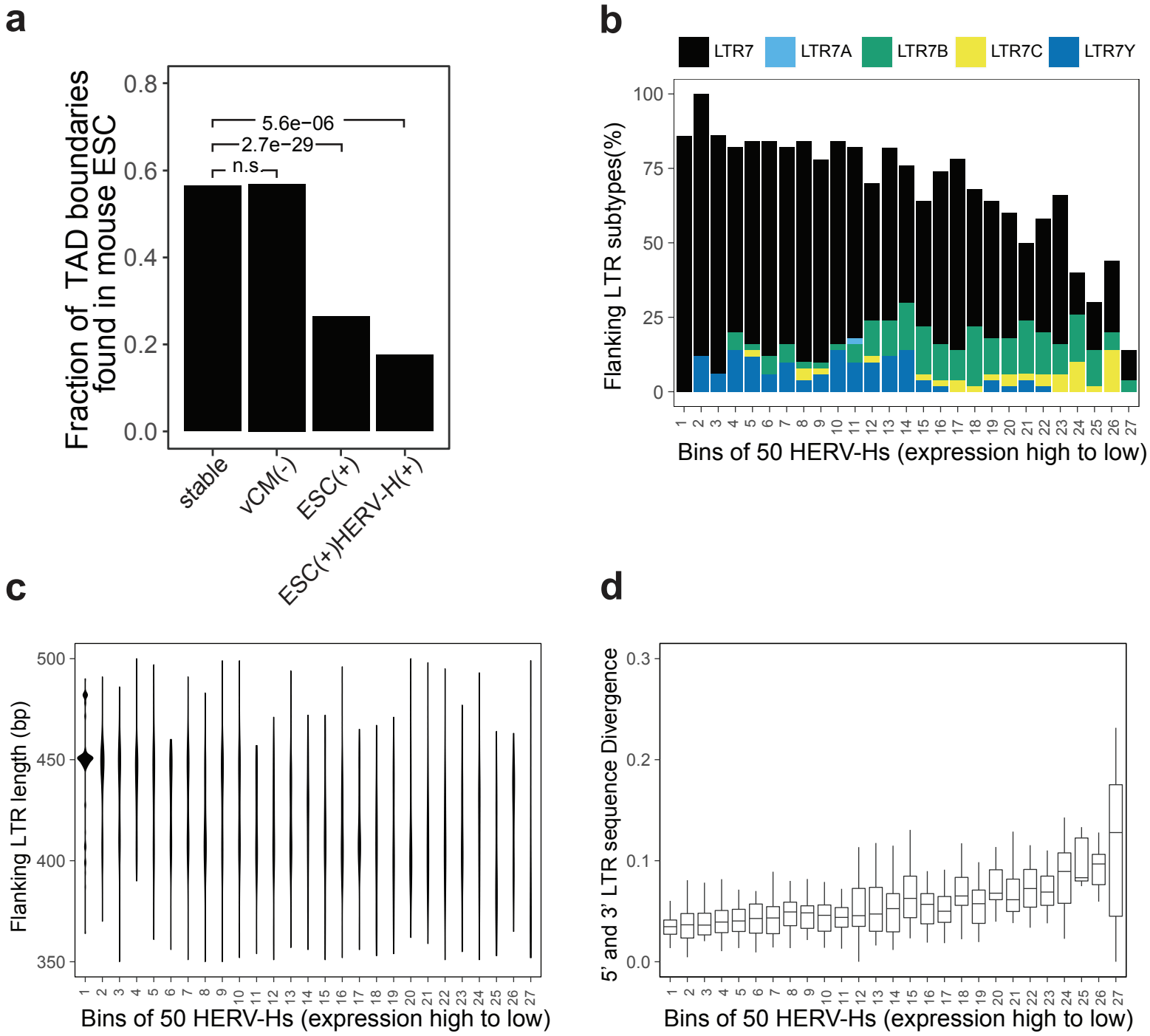
Extended Data Figure 7



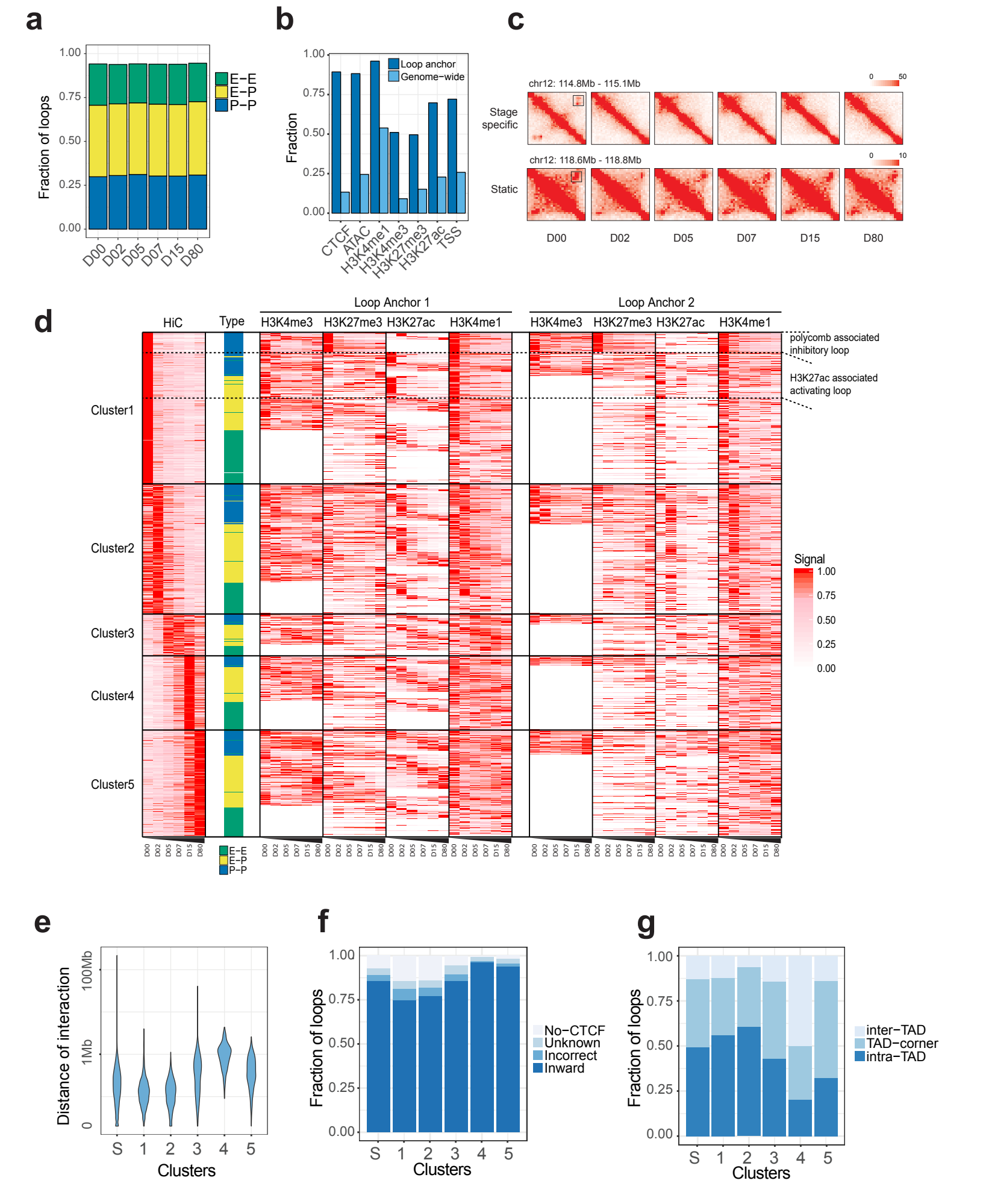
Extended Data Figure 8



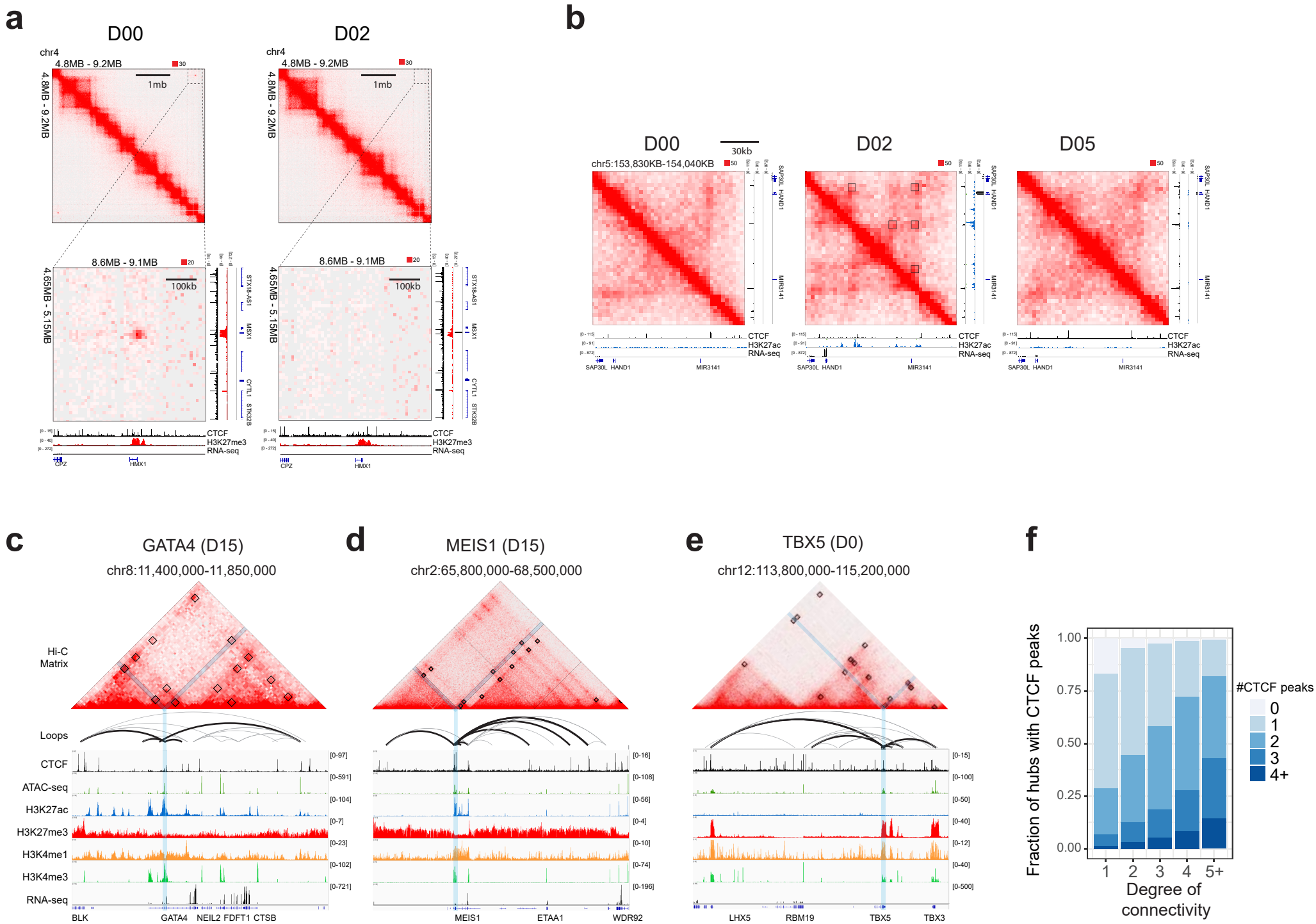
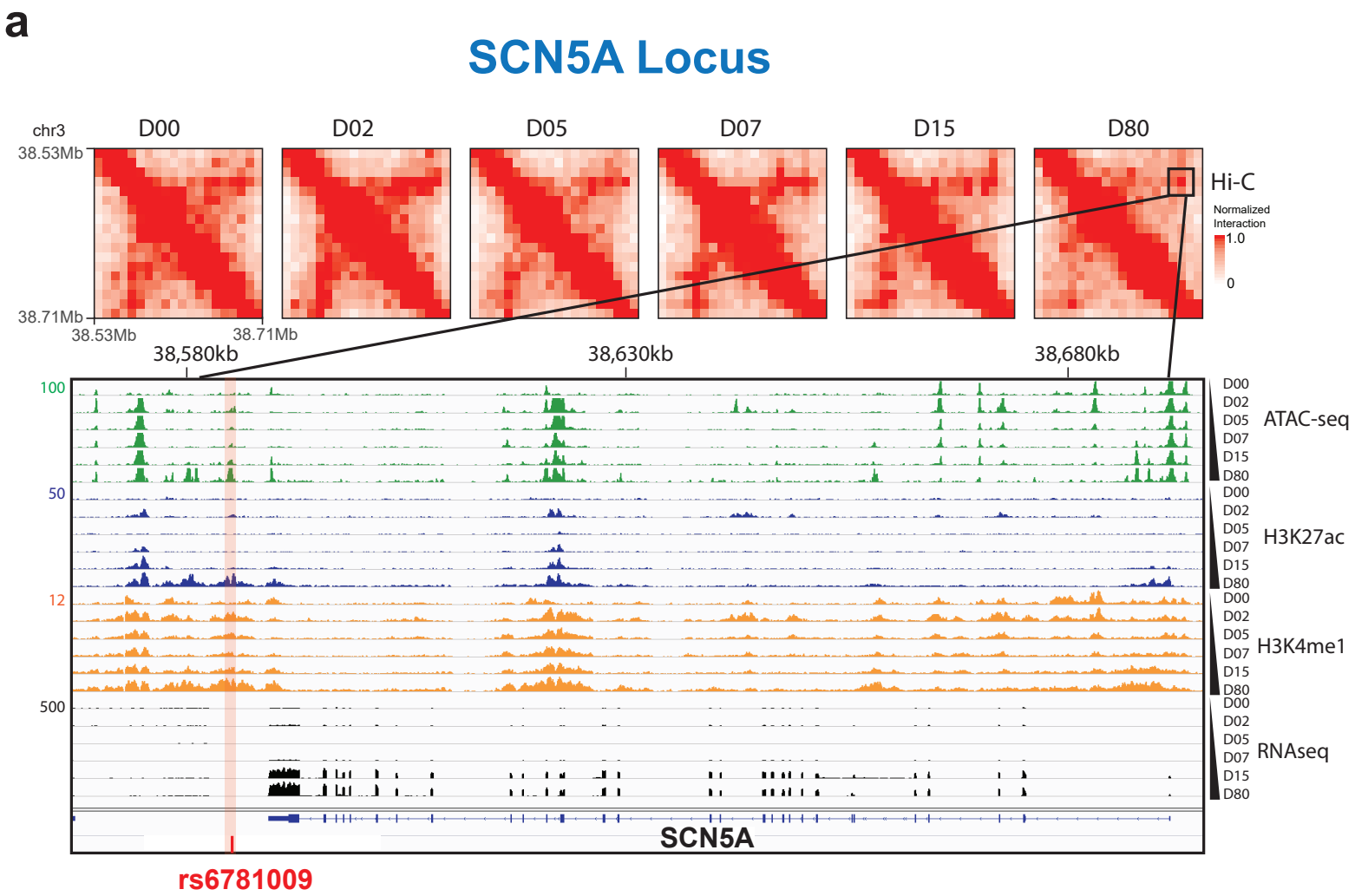
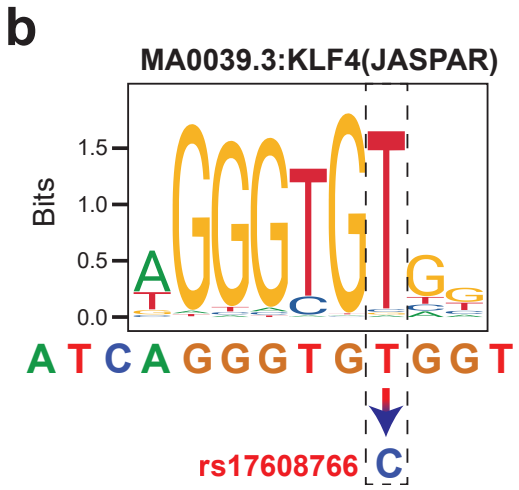
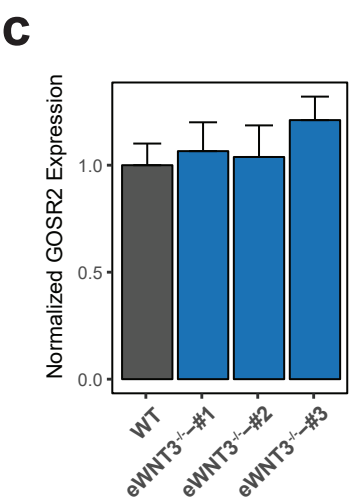
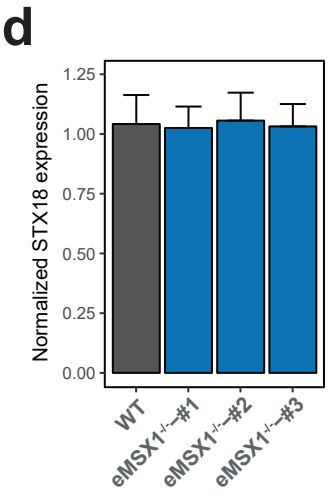
Extended Data Figure 9



Extended Data Figure 10



Extended Data Figure 11



Extended Data Figure 12