Genomic features underlie the co-option of SVA transposons as cis-regulatory elements in human pluripotent stem cells

Samantha M. Barnada[1,2,#], Andrew Isopi[3,4,#], Daniela Tejada-Martinez[1], Clément Goubert[5], Sruti Patoori[1], Luca Pagliaroli[1], Mason Tracewell[1,4], and Marco Trizzino[1,*]

[1] Department of Biochemistry and Molecular Biology, Sidney Kimmel Medical College, Thomas Jefferson University, Philadelphia, PA
[2] Genetics, Genomics and Cancer Biology PhD Program, Thomas Jefferson University, Philadelphia, PA
[3] Department of Microbiology and Immunology, Sidney Kimmel Medical College, Thomas Jefferson University, Philadelphia, PA
[4] Biochemistry and Molecular Pharmacology PhD Program, Thomas Jefferson University, Philadelphia, PA
[5] Department of Human Genetics, McGill University, Montreal, Quebec, Canada

[#] Co-first authors

* Corresponding author: Marco Trizzino, Department of Biochemistry and Molecular Biology, Sidney Kimmel Medical College, Thomas Jefferson University, 233 S 10th Street, BLSB 826, Philadelphia, PA, 19104. E-mail: marco.trizzino83@gmail.com

Abstract

Domestication of transposable elements (TEs) into functional cis-regulatory elements is a widespread phenomenon. However, the mechanisms behind why some TEs are co-opted as functional enhancers while others are not are underappreciated. SINE-VNTR-Alus (SVAs) are the youngest group of transposons in the human genome, where ~3,700 copies are annotated, nearly half of which are human-specific. Many studies indicate that SVAs are among the most frequently co-opted TEs in human gene regulation, but the mechanisms underlying such processes have not yet been thoroughly investigated. Here, we leveraged CRISPR-interference (CRISPRi), computational and functional genomics to elucidate the genomic features that underlie SVA domestication into human stem-cell gene regulation. We found that ~750 SVAs are co-opted as functional cis-regulatory elements in human induced pluripotent stem cells. These SVAs are significantly closer to genes and harbor more transcription factor binding sites than non-co-opted SVAs. We show that a long DNA motif composed of flanking YY1/2 and OCT4 binding sites is enriched in the co-opted SVAs and that these two transcription factors bind consecutively on the TE sequence. We used CRISPRi to epigenetically repress active SVAs in stem cell-like NCCIT cells. Epigenetic perturbation of active SVAs strongly attenuated YY1/OCT4 binding and influenced neighboring gene expression. Ultimately, SVA repression resulted in ~3,000 differentially expressed genes, 131 of which were the nearest gene to an annotated SVA. In summary, we demonstrated that SVAs modulate human gene expression, and uncovered that location and sequence composition contribute to SVA domestication into gene regulatory networks.

Key words: Transposable Elements, SVAs, iPSCs, NCCITs, YY1, OCT4, CRISPRi

Introduction

Transposable elements (TEs) are mobile DNA sequences that account for over 50% of the human genome, yet there is very limited knowledge on the extent of their impact on genome evolution, function, and disease.

Many elegant studies have proposed that TE sequences constantly reshape eukaryotic gene regulation [1–24] yet the underlying mechanisms are largely uncharacterized. Several TE types are active and replication competent in humans, and the genomic dispersal of these elements can affect the regulatory configurations of proximal host genes. For example, TE insertions may introduce novel cis-regulatory elements (CREs = enhancers, promoters, insulators) at the gene locus [8,13,14,16,17]. Alternatively, TE insertion can disrupt transcription factor binding sites (TFBS) within pre-existing CREs, thus attenuating or completely repressing nearby gene expression [25]. Additionally, insertion of TEs into coding sequences of genes may disrupt the open reading frames, and modify splicing sites [26–28]. Dysregulated gene expression due to TE insertions can lead to disease phenotypes as TE de-repression (i.e. de-methylation of TE sequences) is correlated with many neurological disorders and is a hallmark of multiple cancer types [29–40].

Over the last decade, the scientific community has begun to characterize the biological determinants of TE co-option in mammalian regulatory networks [8,13,14,17,20,41]. SINE-VNTR-Alus (SVAs) are the youngest human TEs. These transposons are composed of a 5' CCCTCT hexamer repeat, an Alu-like element, a variable number of tandem repeats (VNTR), a SINE element derived from an ancestral endogenous retrovirus (HERVK-10), and a poly-A tail [42]. Six main SVA subfamilies have been characterized (SVA-A through -F) and nearly half of the annotated ~3,700 copies are human-specific, including all the SVA-Es and -Fs.

Importantly, the SVAs are still actively transposing in the human genome by taking advantage of the L1-LINE machinery, and are among the most epigenetically de-repressed and transcriptionally upregulated TEs across a multitude of cancers and neurological disorders [17,24,26,33,43]. Given their young evolutionary age, SVAs provide a unique opportunity to elucidate how the human genome is evolving.

We, and others, have demonstrated that SVAs are frequently co-opted as functional enhancers and promoters in human and chimpanzee gene regulatory networks [16,17,20,24]. Yet we have an incomplete understanding of the extent of this evolutionary process and the underlying mechanisms. Why are some SVAs recruited as functional enhancers and others are not? What are the molecular properties underlying the regulatory potential of SVAs? Here, we answered these questions by exploring the highly permissive

genomic environment of induced pluripotent stem cells (iPSCs) and NCCIT cells. The latter is a human cell line derived from an embryonal carcinoma which exhibits genomic properties comparable to human embryonic stem cells and iPSCs, including widespread de-methylation. These cells are particularly favorable for the study of transposable elements [18]. Compared to iPSCs, NCCITs are easier to manipulate and more suitable to perform Cas9-mediated genome editing experiments.

We demonstrate that ~750 SVAs are depleted of repressive histone marks (i.e. H3K9me3) in iPSCs and in NCCITs. We found that the transcriptionally active SVAs are significantly closer to genes and harbor more TFBS than those harboring the repressive H3K9me3 epigenetic modification. Moreover, when comparing the sequence of the de-repressed SVAs with the repressed SVAs, we detected an enrichment for a long DNA motif composed of flanking binding sites for YY1/2 and OCT4 within the de-repressed group. The former is a ubiquitously expressed transcriptional regulator, while the latter is an essential regulator of cell pluripotency. Further the de-repressed SVAs are also enriched for individual (i.e. non-flanking) YY1 and OCT4 binding sites. We used Chromatin Immunoprecipitation followed by sequencing (ChIP-seq) to demonstrate that YY1 and OCT4 bind adjacently on many of the active SVAs. We leveraged CRISPR-interference (CRISPRi) to epigenetically repress active SVAs resulting in a loss of YY1 and OCT4 binding leading to massive alterations in gene expression.

Results

Human-specific SVAs are enriched in areas of de-repressed chromatin

We took advantage of publicly available H3K9me3 ChIP-seq data (paired-end long reads) generated in iPSCs [19] to assess to what extent SVAs are repressed in a pluripotent context. We only retained uniquely mapping high quality reads (Samtools q10 filtering). This analysis revealed that of all the SVAs annotated in the human genome (N=3,734; hg19 assembly), approximately 80% (2,983) are decorated with this repressive histone methylation in iPSCs (Fig. 1A). Conversely, 751 SVAs were depleted of H3K9me3, indicating that they are de-repressed (Fig. 1A).

Next, we wanted to compare the iPSCs findings with data generated in NCCITs in our laboratory. As mentioned, NCCITs have pluripotent characteristics but compared to human iPSCs and embryonic stem cells, the NCCITs are significantly easier to manipulate and have high transfection efficiency, which makes them suitable for CRISPR experiments. We thus performed ChIP-seq for H3K9me3 to profile repressed SVA regions in NCCITs. To ensure high-quality read mappability of repetitive regions we sequenced 100 bp long paired-end reads and upon mapping, only retained the uniquely mapping high quality reads

108 (Samtools q10 filtering). Notably, the SVA methylation pattern previously observed in iPSCs was perfectly
109 recapitulated in NCCITs (Fig. 1B & 1C). These findings indicate that NCCITs share a similar TE epigenetic
110 landscape with iPSCs making them a suitable model to study stem cell genomics, and support that 751
111 SVAs are de-repressed in pluripotent environments (hereafter "de-repressed SVAs"; Supplementary File
112 S1).

113 We wanted to ensure that the lack of H3K9me3 signal on the de-repressed SVAs was not a
114 consequence of mapping limitations, in that the youngest SVAs may be too similar to one another to allow
115 for unique read mapping. To test this hypothesis, we looked at another histone modification. Specifically,
116 we took advantage of publicly available H3K27ac ChIP-seq data (100bp paired-end) generated in NCCITs
117 by the Wysocka laboratory [18]. H3K27ac usually decorates active cis-regulatory elements and we
118 surmised that de-repressed SVAs may be transcriptionally active and therefore should exhibit ChIP-seq
119 signal for this histone mark. This analysis revealed that the large majority of the de-repressed SVAs
120 (636/751) are decorated with H3K27ac, which suggests they are in an active regulatory state, and that
121 they are "mappable" with uniquely mapped reads (hereafter "active SVAs"; Fig. 1D). Additionally, 120
122 SVAs were not marked with H3K9me3 or H3K27ac. We attribute this pattern to potential mappability
123 issues, and did not consider this SVA subset for downstream analyses. Overall, these data suggest that
124 approximately 20% of all SVAs are located within transcriptionally de-repressed and active chromatin in
125 iPSCs and in stem cell-like NCCITs (Fig. 1A-E).

126 SVAs can be further categorized into six evolutionarily conserved subfamilies (SVA-A through -F),
127 two of which (SVA-E, -F, plus the F1 subgroup) are human-specific. We investigated the subfamily
128 composition of the de-repressed and repressed SVAs and observed that human-specific SVA subfamilies
129 -E and -F (including F1) were significantly enriched in the de-repressed group (Fisher's Exact Test $p < 1.0 \times$
130 10^{-5}; Fig. 1F). In particular, ~40% of the SVA-Es were found in a de-repressed state, as compared to an
131 average of ~19% for all the other families (Fisher's Exact Test $p < 2.2 \times 10^{-16}$; Fig. 1G). In summary, these
132 findings indicate that young, human-specific SVAs are enriched not only in de-repressed chromatin, but
133 also in active regions in NCCIT cells.
134

135 **Sequence location and composition underlie SVA activation**

136 We aimed to investigate the specific genomic features underlying the selective de-repression of
137 SVAs. First, we reasoned that active SVAs preferentially located near coding genes could be co-opted as
138 active cis-regulatory elements. We used the Genecode_v33 annotations and calculated the distance from
139 the nearest transcription start site (TSS) for both repressed and active SVAs. We observed that active SVAs

140 are significantly closer to coding genes when compared to the repressed SVAs (Wilcoxon's Rank Sum Test
141 $p < 2.2 \times 10^{-16}$; Fig. 2A). Namely, the active SVAs are located approximately 10 kb closer to the nearest TSS
142 than the repressed SVAs (Fig. 2A).

143 Next, we surmised that SVA de-repression may be a consequence of gene regulatory activity.
144 Consistent with this hypothesis, we found that relative to the repressed SVAs, active SVAs were
145 significantly closer to TFBS as defined by the Encode Consortium (Wilcoxon's rank-sum test $p = 2.84 \times 10^{-}$
146 ^{7}; Fig. 2B). Further, the active SVAs directly overlap a TFBS three times more frequently than the repressed
147 SVAs (Fig. 2B). These data suggest that the active SVAs have a higher likelihood of exhibiting gene-
148 regulatory activity.

149 To further explore this hypothesis, we performed sequence-based computational motif analysis
150 on the de-repressed SVAs using HOMER with the repressed SVAs as the background control. With this
151 approach, we identified an enrichment for a long motif composed of flanking YY1/2 and OCT4 binding
152 sites in the de-repressed SVAs (Fig. 2C). YY1/2 are ubiquitously expressed transcriptional regulators, while
153 OCT4 is one of the transcription factors essential for pluripotency maintenance. The YY1-OCT motif is
154 found seven times more frequently in the de-repressed SVAs than in the repressed ones. Importantly, not
155 only is the flanking YY1-OCT4 motif enriched, but also individual (i.e. non-flanking) YY1 and OCT4 motifs
156 are enriched in the de-repressed SVAs as compared to the repressed ones. For instance, OCT4 alone is
157 found in 12% of the de-repressed SVAs, as opposed to 1% of repressed SVAs.

158 Additionally, we identified an enriched CEBPA motif (Fig. 2C). We conducted the opposite
159 analysis, by looking for motifs enriched in repressed SVAs (using de-repressed as a background) and found
160 several enriched motifs, including: SMAD3, KLF5 and several Krüppel-associated box (KRAB) domain-
161 containing zinc-finger proteins (KZFPs).

162

163 **Repression of conventionally de-repressed SVAs has global consequences on gene expression**

164 Since our results indicate that ~20% of all SVAs may act as functional CREs in NCCITs, we aimed
165 to validate our ChIP-seq data and computational predictions with a functional approach. We utilized
166 CRISPRi to repress the 751 de-repressed SVAs. We leveraged the piggyBac transposon system (Systems
167 Bioscience) to generate NCCIT cells with tetracycline-inducible expression of a functionally dead Cas9
168 (dCas9) fused to a repressive KRAB domain. We subsequently knocked-in two previously validated [20]
169 single guide RNAs (sgRNAs) that simultaneously target approximately 80% of all the SVAs annotated in the
170 human genome (Fig. 3A). The sgRNAs target the dCas9-KRAB to de-repressed SVAs leading to the
171 deposition of the repressive histone methylation H3K9me3 by the KRAB domain (Fig. 3A). This NCCIT cell

172 line exhibiting tetracycline-inducible dCas9-KRAB expression and constitutive dual sgRNA expression is
173 hereafter referred to as SCi-NCCITs (i.e. SVA-CRISPRi-NCCITs). We treated the SCi-NCCITs with doxycycline
174 for 72 hours to robustly induce dCas9-KRAB activation and function (Fig. 3B). Activation of dCas9-KRAB
175 led to the accumulation of repressive H3K9me3 modifications on 620 of the 751 SVAs originally de-
176 repressed in NCCITs (Fig. 3C & 3D). Importantly, this further demonstrates that reads can be uniquely
177 mapped on this group of SVAs.

178 To assess the impact of global SVA repression on genome-wide gene expression, we performed
179 RNA-sequencing on the SCi-NCCITs in the presence and absence of doxycycline treatment (three replicates
180 per condition). We identified 3,085 genes as differentially expressed upon induction of SVA repression
181 (FDR < 0.05; Supplementary File S2) of which 1,596 were identified as upregulated and 1,489 as
182 downregulated (Fig. 4A). The sgRNAs used for this experiment were originally designed to target a DNA
183 sequence shared by SVAs with the LTR5Hs family [20]. Given this premise, we restricted our analysis
184 exclusively to genes putatively associated with human SVAs (i.e. considering only the genes that represent
185 the closest gene to an annotated SVA). Overall, 131 of the 3,085 differentially expressed genes
186 represented the closest gene to an SVA (Fig. 4B), 101 of which were specifically near a de-repressed SVA.
187 This number is significantly higher than expected by chance (Fisher's Exact Test $p < 0.001$), suggesting that
188 the expression of most of these 131 genes is likely under the direct control of SVAs. Importantly, 109 of
189 the 131 SVA-regulated genes (83.2%) were downregulated upon doxycycline treatment (Fig. 4B),
190 indicating that SVA de-repression is necessary for gene activation.

191 We then assessed if any of the differential gene expression is due to repressing LTR5Hs since the
192 sgRNAs also target DNA sequences in this TE family. Only 120/3,085 differentially expressed genes
193 represented the closest gene to an LTR5H. Additionally, we detected very minimal overlap (8 genes in
194 total) between the differentially expressed genes near LTR5Hs and the differentially expressed genes near
195 SVAs. These analyses ensure that the alterations in gene expression are due to the repression of SVAs and
196 not LTR5Hs. In summary, these data indicate that the SVAs directly regulate the expression of at least 130
197 genes and that SVA repression may contribute to the differential expression of up to ~2,800 genes,
198 potentially as a cascade effect.

199 We performed Ingenuity Pathway Analysis (IPA) on the 131 SVA-regulated genes and found an
200 enrichment for gap junction signaling processes including Gap Junction Signaling, Sertoli Cell-Sertoli Cell
201 Junction Signaling, and Germ Cell-Sertoli Cell Junction Signaling (Fig. 4C). These processes are important
202 during gametogenesis and given that NCCIT cells retain pluripotency-like characteristics, we speculate
203 that co-opted SVAs could play a role in regulating germ line development and gametogenesis. However,

204 it is important to note that the NCCIT cell line is derived from an embryonal testicular carcinoma, and this
205 could bias our IPA. Nonetheless, recent studies have corroborated an SVA contribution in germ cells and
206 showed that in human primordial germ cells SVA sites are largely hypomethylated and genes proximal to
207 SVAs are expressed at a higher level compared to embryonic stem cells [20,44,45]. Additionally, SVA
208 transcription and retrotransposition are specifically seen in both human spermatozoa [46] and oocytes
209 [47]. INSERT IPA DATA UP vs DOWN REGULATED HERE (Fig. 4D).
210 We next used IPA to look for top upstream transcriptional regulators of the 131 SVA-associated
211 genes. This analysis identified transcription factors previously highlighted by the motif analyses (CEBPA
212 and YY1/2), along with several others including SMARCA1, CREB1, and WT1 (Fig. 4E). While YY1 is an
213 essential ubiquitous developmental regulator [48], CEBPA is a regulator of differentiation in the
214 hematopoietic [49] and adipocyte [50] lineages, and is involved in gametogenesis. SMARCA1 is a
215 chromatin remodeler [51], CREB1 plays a role in both steroidal [52] and non-steroidal [53] female
216 hormonal stimulation, and WT1 is involved in urogenital specification distinctively associated with the
217 differentiation of Sertoli cells [54]. These data suggest a possible SVA contribution to germ line
218 developmental processes and demonstrates the genome-wide impact of SVA repression on gene
219 expression.
220
221 **YY1 and OCT4 mediate SVA regulatory activity**
222 Since our motif analysis revealed that de-repressed SVAs are enriched for adjacent YY1/2-OCT4
223 binding motifs and given that many differentially expressed genes were YY1/2 targets, we performed
224 ChIP-seq for YY1 and OCT4 in SCi-NCCITs with and without doxycycline treatment. As done previously for
225 the histone modification ChIP-seq, we sequenced 100 bp long paired-end reads to ensure high
226 mappability efficiency.
227 These experiments revealed that under normal conditions (i.e. no doxycycline treatment) YY1 and
228 OCT4 bind 288 and 54 SVAs, respectively (Figs. 5A & 5B). Specifically, we identified two different clusters
229 of YY1 binding: one located in the *Alu*-like region (Fig. 5A top cluster), and a second near the start of the
230 SINE element, likely in proximity of the HERVK10-derived promoter in the SINE region (Fig. 5A second
231 cluster). OCT4 binding was limited to this second region, whereas we did not detect binding for this
232 transcription factor in the *Alu*-like region (Fig. 5B). The SVAs bound by YY1 and OCT4 were all de-repressed
233 SVAs decorated with H3K27ac. Importantly, binding of both YY1 and OCT4 was significantly attenuated
234 upon SVA repression via CRISPRi (Figs. 5A, B). We observed a significant overlap with 33 SVAs bound by
235 both YY1 and OCT4 in the SINE region. In this case, the binding of the two transcription factors was

236 sequential (i.e. one next to each other) as originally predicted by our motif analysis (Fig. 5C). This pattern
237 was found in de-repressed SVAs of all the main families (SVA-A through -F).
238 Finally, we leveraged the nearest TSS approach to determine the closest gene to each of the SVAs
239 bound by YY1 and/or OCT4. Using our RNA-seq data, we investigated whether loss of YY1 and OCT4
240 binding altered the expression of neighboring genes. Notably, 44 of 288 genes located near YY1-bound
241 SVAs and 24 of 54 genes located near OCT4-bound SVAs were differentially expressed upon CRISPRi-
242 induced SVA repression (Fig. 5D). Interestingly, of the 6 genes upregulated upon doxycycline treatment,
243 4 were located near SVAs bound exclusively by OCT4 (Fig. 5D). Overall, 85% of the genes (28/33) located
244 near SVAs that were bound by both YY1 and OCT4 were differentially expressed upon CRISPRi-induced
245 SVA repression (Fig. 5D). This suggests that synergistic binding of these two transcription factors on the
246 SVA sequence is important for the regulatory activity driven by these transposable elements (Fig. 5D &
247 Fig. 6). Pathway Analysis revealed that the genes regulated by the cooperative YY1-OCT4 binding (and by
248 OCT4-only) are enriched for processes related to pluripotency and DNA repair (Fig. 5E). Conversely, the
249 genes bound by YY1-only are enriched for processes related to germline development (Fig. 5E).
250
251 **Discussion**
252 SVAs are evolutionarily young transposable elements, which colonized the great ape genomes in
253 the last 10-15 million years. In fact, they are not found in gibbons whose lineage split from the remaining
254 apes ~17 million years ago [55]. Interestingly, the gibbon genome has been independently colonized by a
255 distinct family of transposons called LAVAs (LINE-AluSz-VNTR-Alu) [56,57]. Notably, SVA and LAVA
256 structures are similar, and both have shown high cis-regulatory potential and massive co-option into gene
257 regulatory networks [16,17,20,57]. However, the mechanisms underlying SVA and LAVA domestication
258 and recruitment into primate gene regulatory networks have not been explored in depth.
259 Here, we aimed at elucidating such mechanisms, focusing specifically on SVAs and stem cells. We
260 show that ~20% of all the human SVAs are found in a de-repressed chromatin state in iPSCs and in NCCITs,
261 with a near perfect overlap between the two cell types. Most of these de-repressed SVAs were also
262 decorated with histone modifications characteristic of active enhancers and promoters (H3K27ac),
263 suggesting that their de-repression is associated with cis-regulatory activity. This pattern may indicate
264 that either the SVAs are enriched in regions that were already transcriptionally active before their
265 insertion, or that the SVAs themselves dictated the epigenetic landscape as a consequence of their
266 sequence. A body of literature has emerged supporting the contribution of specific TE families to the
267 dispersion of TFBS and cis-regulatory elements in the eukaryotic genomes (Fueyo et al., 2022). This can

268 happen in multiple ways: TEs already harbor the TFBS before the transposition event, or they gain TFBS
269 afterwards as a consequence of new mutations. In the former case, if TEs harboring TFBSs insert near
270 genes, this will increase the likelihood for the TE to be co-opted as an enhancer/promoter and thus lose
271 repressive H3K9me3 and gain H3K27ac. Consistent with this scenario, it is estimated that in human
272 embryonic stem cells ~20% of the TFBSs for pluripotency factors are located within transposable elements
273 (Kunarso et al. 2010; Sundaram et al 2017; Fueyo et al. 2022).

274 The youngest SVAs (which are human-exclusive), and especially the SVA-Es, are the most enriched
275 among the active copies. To this end, we speculate that the older SVAs accumulated genetic mutations
276 over time, hampering their regulatory potential. Alternatively, the human genome may be adapted to
277 silence older transposons, and this may have affected copies with higher regulatory potential.

278 According to our data, SVA location and sequence composition are the best predictors of cis-
279 regulatory activity. Active SVAs are, on average, 10 kb closer to genes than the repressed ones. This is
280 likely due to the fact that the chromatin environment near gene loci may be more frequently in an
281 accessible and de-repressed state and is thus more suitable for co-option of novel cis-regulatory elements.
282 Moreover, a shorter distance from the nearest TSS may facilitate the tri-dimensional interaction between
283 the SVA-derived enhancer and the gene promoter. This is also in line with a recent study that
284 demonstrated that eQTL-rich TEs tend to be significantly closer to genes than eQTL-poor TEs [25].

285 We show that active SVAs host a significantly higher number of TFBS than the repressed ones.
286 This coincides with many studies suggesting that TE exaptation into gene regulatory networks is largely
287 driven by the evidence that they propagate TFBS across the genome [60]. We demonstrate that a long
288 motif with flanking YY1/2-OCT4 binding sites is enriched in de-repressed SVA copies relative to repressed
289 SVAs. This may mediate their function in gene regulatory networks of stem and stem-like cells. In fact,
290 OCT4 is one of the four Yamanaka factors [61] essential for pluripotency maintenance. YY1 is one of the
291 major transcriptional regulators in human cells; it is ubiquitously expressed across all cell-types and
292 performs many different functions in transcriptional regulation, including transcriptional activation,
293 repression, as well as mediation of enhancer-promoter looping [62]. To this regard, over 40% of the SVAs
294 de-repressed in iPSCs and NCCITs are bound by YY1, OCT4, or both. Importantly, when both are present,
295 they bind alongside each other as predicted by computational motif analysis. As expected, depositing
296 repressive histone methylation (H3K9me3) on the SVA locus nearly abolishes the binding of these two
297 transcription factors near SVAs, leading to an alteration of nearby gene expression. In fact, even when
298 restricting the analysis to the genes that are located near SVAs, repressing 620 of the 751 SVAs normally
299 active in NCCITs results in 131 of differentially expressed genes. The expression of most (88%) of these

300 genes is attenuated with the repression of the nearby SVA, further supporting the enhancer activity
301 provided by these transposons. These genes include important regulators of cell pluripotency and cell
302 differentiation, such as *MYC*, *MYBL2*, *FUS*, *ITGAX*, *SP4* and several others. We remark that this analysis
303 was very conservative. In fact, given the nature of the sgRNAs that we chose, which target both SVAs and
304 LTR5Hs, we exclusively focused on the nearest gene to an SVA transposon, and thus the number of genes
305 affected by distal SVA repression is likely much higher.

306 Finally, repressing SVAs did not result in any obvious alteration in cellular phenotypes, although
307 we cannot rule out that a longer experiment (i.e. with cells collected more than 72 hours post doxycycline
308 treatment) may result in alterations in cell viability/proliferation as a consequence of SVA-repression.
309 Future studies may assess longer term SVA repression, focusing on the ability of iPSCs to act as pluripotent
310 cells upon sustained repression of the SVA-derived enhancers. In summary, in this study we provide
311 further evidence that SVA transposons are an important component of the human gene regulatory
312 networks, specifically in stem and stem-like cells. We propose a potential mechanism underlying this cis-
313 regulatory activity where SVA location and sequence composition regulate this co-option. Additional
314 studies are required to determine if the YY1-OCT4 combination is a driver of SVA regulatory activity only
315 in pluripotent cells or, alternatively, in a broader, more universal context. In this study most genomics
316 experiments were conducted in NCCIT cells. Further genomic studies in human iPSCs will be necessary to
317 confirm the proposed mechanism.
318

319 **Materials and Methods**
320

321 ***Antibodies and sgRNAs***

322 YY1 ChIP-seq: Cell Signaling Technology D5D9Z/46395S (15ug per ChIP). OCT4 ChIP-seq: Abcam ab181557
323 (15ug per ChIP). H3K27me3 ChIP-seq: Abcam ab8898 (3ug per ChIP). Cas9 Western Blot: Active Motif
324 61757 (1:100). GAPDH Western Blot: Cell Signaling Technology D16H11/5174 (1:1000). Anti-Rabbit IgG,
325 HRP-linked Western Blot: Cell Signaling Technology 7074 (1:10000). Anti-Mouse IgG, HRP-linked Western
326 Blot: Cell Signaling Technology 7076 (1:10000). The two sgRNAs were designed and used in a previous
327 study [20]: sgRNA1: 5' CTCCCTAATCTCAAGTACCC 3' ; sgRNA2: 5' TGTTTCAGAGAGCACGGGGT 3'.
328

329 ***NCCIT Cell Line Culture***

330 The NCCIT cell line (ATCC) was maintained in RPMI media supplemented with 10% tet-free FBS, 1%
331 penicillin-streptomycin solution, and 1% L-glutamine and incubated at 5% CO_2, 20% O_2 at 37°C.
332

333 ***ChIP-Sequencing***

334 All samples from different conditions were processed together to prevent batch effects. Between 10-15
335 million cells were cross-linked with 1% formaldehyde for 5 minutes at room temperature, quenched with
336 125 mM glycine, harvested, and washed twice with 1x PBS. The fixed cell pellet was resuspended in ChIP
337 lysis buffer (150 mM NaCl, 1% Triton X-100, 0.7% SDS, 500 µM DTT, 10 mM Tris-HCl, 5 mM EDTA) and
338 chromatin was sheared to an average length of 200–900 base-pairs, using a Covaris S220 Ultrasonicator.
339 The chromatin lysate was diluted with SDS-free ChIP lysis buffer. 15µg of antibody was used for YY1 and
340 OCT4 and 3µg of antibody for H3K9me3. The antibody was added to 5µg of sonicated chromatin along
341 with Dynabeads Protein G magnetic beads (Invitrogen) and incubated at 4°C overnight. The beads were
342 washed twice with each of the following buffers: Mixed Micelle Buffer (150 mM NaCl, 1% Triton X-100,
343 0.2% SDS, 20 mM Tris-HCl, 5 mM EDTA, 65% sucrose), Buffer 200 (200 mM NaCl, 1% Triton X-100, 0.1%
344 sodium deoxycholate, 25 mM HEPES, 10 mM Tris-HCl, 1 mM EDTA), LiCl detergent wash (250 mM LiCl,
345 0.5% sodium deoxycholate, 0.5% NP-40, 10 mM Tris-HCl, 1 mM EDTA) and a final wash was performed
346 with 0.1X TE. Finally, beads were resuspended in 1X TE containing 1% SDS and incubated at 65°C for 10
347 min to elute immunocomplexes. The elution was repeated twice and the samples were incubated
348 overnight at 65°C to reverse cross-linking, along with the untreated input (5% of the starting material).
349 The DNA was digested with 0.5 mg/ml Proteinase K for 1 hour at 65°C and then purified using the ChIP
350 DNA Clean & Concentrator kit (Zymo) and quantified with QUBIT. Barcoded libraries were made with
351 NEBNext Ultra II DNA Library Prep Kit for Illumina (New England BioLabs) and sequenced on an Illumina
352 NextSeq 2000 producing 100 bp paired-end reads.

353

354 ***ChIP-seq Analysis***

355 After removing the adapters with TrimGalore!, the sequences were aligned to the reference hg19, using
356 Burrows-Wheeler Alignment tool, with the MEM algorithm [63]. Uniquely mapping aligned reads were
357 filtered based on mapping quality (MAPQ > 10) to restrict our analysis to higher quality and likely uniquely
358 mapped reads, and PCR duplicates were removed. Peaks were called for each SVA site using the default
359 parameters, at 5% FDR, with default parameters

360

361 ***Generation and Culturing of SCi-NCCIT Stable Cell Lines***

362 A plasmid with a tetracycline-inducible dCas9-KRAB expression cassette flanked by piggyBac
363 recombination sites was obtained from the Wysocka Lab at Stanford University. This plasmid 'p-dCas9-
364 KRAB' confers constitutive puromycin resistance, allowing for selection of stably transduced clones when

365 co-expressed with the piggyBac transposase plasmid ('p-PB-Transposase', Systems Bioscience). The p-
366 dCas9-KRAB and p-PB-Transposase plasmids were co-transfected into NCCIT cells (ATCC) at 70%
367 confluency using a 6:1 ratio of Fugene HD (Promega) for 48 hours. Two days post-transfection, cells were
368 treated with puromycin selective media at a concentration of 1 ug/mL. Stable clones were isolated and
369 dCas9 expression assessed via Western blot. Next, we obtained a piggyBac transposon plasmid containing
370 the two sgRNAs [20] targeting ~80% of all annotated SVAs in humans termed 'p-sgRNA' (Systems
371 Bioscience). This plasmid constitutively confers dual sgRNA expression and geneticin resistance. The p-
372 sgRNA and p-PB-Transposase plasmids were co-transfected into the NCCIT-dCas9KRAB cells as above. Two
373 days post-transfection cell were treated with geneticin selective media at a concentration of 400 ug/mL.
374 Following antibiotic selection, the NCCIT-dCas9KRAB-SVAsgRNA (SCi-NCCITs) cell line was maintained in
375 ATCC-formulated RPMI media supplemented with 10% tet-free FBS, 1% L-glutamine, 1µg/mL puromycin,
376 and 400 µg/mL geneticin and incubated at 5% CO2, 20% O2 at 37°C. The SCi-NCCITs were seeded to 40%
377 confluency and treated with 2 ug/mL doxycycline (Sigma Aldrich) for 72 hours. For all molecular and
378 genomic CRISPRi experiments, dCas9-KRAB expression was induced with doxycycline for 72 hours.
379
380 ***Western Blot***
381 Cells were washed three times in PBS and lysed in radioimmunoprecipitation assay buffer (RIPA buffer)
382 (50 mM Tris-HCl pH7.5, 150 mM NaCl, 1% Igepal, 0.5% sodium deoxycholate, 0.1% SDS 500 µM DTT) with
383 protease inhibitors. Approximately 40 µg of whole cell lysate were loaded in Novex WedgeWell 4–20%
384 Tris-Glycine Gel (Invitrogen) and subject to SDS-PAGE. Proteins were then transferred to a Immun-Blot
385 PVDF membrane (ThermoFisher) for antibody probing. Membranes were blocked with a 10% BSA in TBST
386 solution for 30 minutes then incubated with primary antibodies in a 5% BSA in TBST, diluted as above.
387 Next, membranes were washed with TBST and incubated with secondary antibodies, diluted as above.
388 Chemiluminescent signal was detected using the Pierce ECL Plus Western Blotting Substrate
389 (ThermoFisher) and an Amersham Imager 680.
390
391 ***RNA Extraction and Library Preparation for RNA sequencing***
392 Cells were lysed in Tri-reagent (Zymo) and total RNA was extracted using Direct-zol RNA Miniprep kit
393 (Zymo) according to the manufacturer's instructions. RNA was quantified using DeNovix DS-11
394 Spectrophotometer while the RNA integrity was checked on a Bioanalyzer 2100 (Agilent). Only samples
395 with RIN value above 8.0 were used for transcriptome analysis. RNA libraries were prepared using 1µg of
396 total RNA input using NEB- Next® Poly(A) mRNA Magnetic Isolation Module, NEBNext® UltraTM II

397 Directional RNA Library Prep Kit for Illumina® and NEBNext® UltraTM II DNA Library Prep Kit for Illumina®
398 according to the manufacturer's instructions (New England Biolabs). Paired-end 100 bp reads were
399 generated.
400
401 ***RNA-seq Analysis***
402 Reads were aligned to hg19 using STAR v2.567 [64], in 2-pass mode. Bam files were filtered based on
403 alignment quality (q = 10) using Samtools [63]. We used the latest annotations obtained from Ensembl to
404 build reference indexes for the STAR alignment. Adapters were removed with TrimGalore! and Kallisto68
405 [65] was used to count reads mapping to each gene. We analyzed differential gene expression levels with
406 DESeq2 [66].
407
408 ***Statistical and Genomic Analyses***
409 All statistical analyses were performed using BEDTools v2.27.1 [67], DeepTools, and R v4.1.2. Fasta files
410 of the regions of interest were produced using BEDTools v2.27.1 [67]. Shuffled input sequences were used
411 as background. E-values < 0.001 were used as a threshold for significance. Motif analysis of de-repressed
412 SVAs on a repressed SVA background was performed using HOMER [68]. Pathway analysis was performed
413 with Ingenuity-Pathway Analysis Suite (Qiagen Inc.,
414 https://www.qiagenbioinformatics.com/products/ingenuity-pathway-analysis).
415
416 **Competing interests**
417 The authors declare no competing interests.
418
419 **Acknowledgements**
420 The authors are grateful to Dr. Joanna Wysocka's lab for providing the dCas9-KRAB piggyBac plasmid and
421 in particular to Dr. Raquel Fueyo for providing critical support for the SCi-NCCIT CRISPRi line generation.
422 The authors thank the Genomic Facility at The Wistar Institute (Philadelphia, PA) for the Next Generation
423 Illumina Sequencing.
424
425 **Funding**
426 For this work, M. Trizzino was funded by the National Institute of Health (NIH-NIGMS R35GM138344) and
427 by the G. Harold and Leila Y. Mathers Foundation.
428
429 **Data availability**
430 The original genome-wide data generated in this study have been deposited in the GEO database under
431 accession code GSE192951.
432
433 **Author contributions**
434 MTrizzino designed the project. SMB and AI generated the stable SCi-NCCIT CRISPR line; AI performed a
435 preliminary set of genomic experiments. SMB performed all the genomic experiments. MTrizzino, SMB,

436 and DTM analyzed the data. MTrizzino and SMB wrote the manuscript. CB provided essential advice and
437 contribution for the planning of the CRISPR experiment. LP, SP and MTracewell contributed to some of
438 the experiments. All the authors read and approved the manuscript.
439
440 **References**

441 1. McClintock B. The Origin and Behavior of Mutable Loci in Maize. Proc Natl Acad Sci U S A. 1950;36:
442 344–355.

443 2. McClintock B. The significance of responses of the genome to challenge. Science. 1984;226: 792–
444 801. doi:10.1126/science.15739260

445 3. Davidson EH, Britten RJ. Regulation of Gene Expression: Possible Role of Repetitive Sequences.
446 Science. 1979;204: 1052–1059. doi:10.1126/science.451548

447 4. Jordan IK, Rogozin IB, Glazko GV, Koonin EV. Origin of a substantial fraction of human regulatory
448 sequences from transposable elements. Trends in Genetics. 2003;19: 68–72. doi:10.1016/S0168-
449 9525(02)00006-9

450 5. Bejerano G, Lowe CB, Ahituv N, King B, Siepel A, Salama SR, et al. A distal enhancer and an
451 ultraconserved exon are derived from a novel retroposon. Nature. 2006;441: 87–90.
452 doi:10.1038/nature04696

453 6. Wang T, Zeng J, Lowe CB, Sellers RG, Salama SR, Yang M, et al. Species-specific endogenous
454 retroviruses shape the transcriptional network of the human tumor suppressor protein p53. PNAS.
455 2007;104: 18613–18618. doi:10.1073/pnas.0703637104

456 7. Bourque G, Leong B, Vega VB, Chen X, Lee YL, Srinivasan KG, et al. Evolution of the mammalian
457 transcription factor binding repertoire via transposable elements. Genome Res. 2008;18: 1752–
458 1762. doi:10.1101/gr.080663.108

459 8. Kunarso G, Chia N-Y, Jeyakani J, Hwang C, Lu X, Chan Y-S, et al. Transposable elements have
460 rewired the core regulatory network of human embryonic stem cells. Nat Genet. 2010;42: 631–
461 634. doi:10.1038/ng.600

462 9. Lynch VJ, Leclerc RD, May G, Wagner GP. Transposon-mediated rewiring of gene regulatory
463 networks contributed to the evolution of pregnancy in mammals. Nat Genet. 2011;43: 1154–1159.
464 doi:10.1038/ng.917

465 10. Lynch VJ, Nnamani MC, Kapusta A, Brayer K, Plaza SL, Mazur EC, et al. Ancient transposable
466 elements transformed the uterine regulatory landscape and transcriptome during the evolution of
467 mammalian pregnancy. Cell Rep. 2015;10: 551–561. doi:10.1016/j.celrep.2014.12.052

468 11. Schmidt D, Schwalie PC, Wilson MD, Ballester B, Gonçalves Â, Kutter C, et al. Waves of
469 Retrotransposon Expansion Remodel Genome Organization and CTCF Binding in Multiple
470 Mammalian Lineages. Cell. 2012;148: 335–348. doi:10.1016/j.cell.2011.11.058

471 12. Jacques P-É, Jeyakani J, Bourque G. The Majority of Primate-Specific Regulatory Sequences Are
472 Derived from Transposable Elements. PLOS Genetics. 2013;9: e1003504.
473 doi:10.1371/journal.pgen.1003504

474 13. Chuong EB, Rumi MAK, Soares MJ, Baker JC. Endogenous retroviruses function as species-specific
475 enhancer elements in the placenta. Nat Genet. 2013;45: 325–329. doi:10.1038/ng.2553

476 14. Chuong EB, Elde NC, Feschotte C. Regulatory evolution of innate immunity through co-option of
477 endogenous retroviruses. Science. 2016;351: 1083–1087. doi:DOI: 10.1126/science.aad5497

478 15. Sundaram V, Cheng Y, Ma Z, Li D, Xing X, Edge P, et al. Widespread contribution of transposable
479 elements to the innovation of gene regulatory networks. Genome Res. 2014;24: 1963–1976.
480 doi:10.1101/gr.168872.113

481 16. Trizzino M, Park Y, Holsbach-Beltrame M, Aracena K, Mika K, Caliskan M, et al. Transposable
482 elements are the primary source of novelty in primate gene regulation. Genome Res. 2017;27:
483 1623–1633. doi:10.1101/gr.218149.116

484 17. Trizzino M, Kapusta A, Brown CD. Transposable elements generate regulatory novelty in a tissue-
485 specific fashion. BMC Genomics. 2018;19: 468. doi:10.1186/s12864-018-4850-3

486 18. Fuentes DR, Swigut T, Wysocka J. Systematic perturbation of retroviral LTRs reveals widespread
487 long-range effects on human gene regulation. Heard E, Weigel D, editors. eLife. 2018;7: e35989.
488 doi:10.7554/eLife.35989

489 19. Ward MC, Zhao S, Luo K, Pavlovic BJ, Karimi MM, Stephens M, et al. Silencing of transposable
490 elements may not be a major driver of regulatory evolution in primate iPSCs. Wittkopp PJ, editor.
491 eLife. 2018;7: e33084. doi:10.7554/eLife.33084

492 20. Pontis J, Planet E, Offner S, Turelli P, Duc J, Coudray A, et al. Hominoid-Specific Transposable
493 Elements and KZFPs Facilitate Human Embryonic Genome Activation and Control Transcription in
494 Naive Human ESCs. Cell Stem Cell. 2019;24: 724-735.e5. doi:10.1016/j.stem.2019.03.012

495 21. Miao B, Fu S, Lyu C, Gontarz P, Wang T, Zhang B. Tissue-specific usage of transposable element-
496 derived promoters in mouse development. Genome Biology. 2020;21: 255. doi:10.1186/s13059-
497 020-02164-3

498 22. Judd J, Sanderson H, Feschotte C. Evolution of mouse circadian enhancers from transposable
499 elements. Genome Biology. 2021;22: 193. doi:10.1186/s13059-021-02409-9

500 23. Mika K, Marinić M, Singh M, Muter J, Brosens JJ, Lynch VJ. Evolutionary transcriptomics implicates
501 new genes and pathways in human pregnancy and adverse pregnancy outcomes. Rokas A, Perry
502 GH, Stevens A, Wildman DE, Mesiano S, editors. eLife. 2021;10: e69584. doi:10.7554/eLife.69584

503 24. Patoori S, Barnada S, Trizzino M. Young transposable elements rewired gene regulatory networks
504 in human and chimpanzee hippocampal intermediate progenitors. 2021 Nov p.
505 2021.11.24.469877. doi:10.1101/2021.11.24.469877

506 25. Goubert C, Zevallos NA, Feschotte C. Contribution of unfixed transposable element insertions to
507 human regulatory variation. Philosophical Transactions of the Royal Society B: Biological Sciences.
508 2020;375: 14.

509 26. Ostertag EM, Goodier JL, Zhang Y, Kazazian HH. SVA Elements Are Nonautonomous
510 Retrotransposons that Cause Disease in Humans. The American Journal of Human Genetics.
511 2003;73: 1444–1451. doi:10.1086/380207

512 27. Cosby RL, Judd J, Zhang R, Zhong A, Garry N, Pritham EJ, et al. Recurrent evolution of vertebrate
513 transcription factors by transposase capture. Science. 2021;371. doi:DOI: 10.1126/science.abc6405

514 28. Payer LM, Steranka JP, Ardeljan D, Walker J, Fitzgerald KC, Calabresi PA, et al. Alu insertion variants
515 alter mRNA splicing. Nucleic Acids Research. 2019;47: 421–431. doi:10.1093/nar/gky1086

516 29. Lee E, Iskow R, Yang L, Gokcumen O, Haseley P, Luquette LJ, et al. Landscape of Somatic
517 Retrotransposition in Human Cancers. Science. 2012;337: 967–971. doi:10.1126/science.1222077

518 30. Li W, Jin Y, Prazak L, Hammell M, Dubnau J. Transposable Elements in TDP-43-Mediated
519 Neurodegenerative Disorders. PLOS ONE. 2012;7: e44099. doi:10.1371/journal.pone.0044099

520 31. Li W, Prazak L, Chatterjee N, Grüninger S, Krug L, Theodorou D, et al. Activation of transposable
521 elements during aging and neuronal decline in Drosophila. Nat Neurosci. 2013;16: 529–531.
522 doi:10.1038/nn.3368

523 32. Pugacheva EM, Teplyakov E, Wu Q, Li J, Chen C, Meng C, et al. The cancer-associated CTCFL/BORIS
524 protein targets multiple classes of genomic repeats, with a distinct binding and functional
525 preference for humanoid-specific SVA transposable elements. Epigenetics & Chromatin. 2016;9:
526 35. doi:10.1186/s13072-016-0084-2

527 33. Anwar SL, Wulaningsih W, Lehmann U. Transposable Elements in Human Cancer: Causes and
528 Consequences of Deregulation. International Journal of Molecular Sciences. 2017;18: 974.
529 doi:10.3390/ijms18050974

530 34. Krug L, Chatterjee N, Borges-Monroy R, Hearn S, Liao W-W, Morrill K, et al. Retrotransposon
531 activation contributes to neurodegeneration in a Drosophila TDP-43 model of ALS. PLOS Genetics.
532 2017;13: e1006635. doi:10.1371/journal.pgen.1006635

533 35. Guo C, Jeong H-H, Hsieh Y-C, Klein H-U, Bennett DA, De Jager PL, et al. Tau Activates Transposable
534 Elements in Alzheimer's Disease. Cell Reports. 2018;23: 2874–2880.
535 doi:10.1016/j.celrep.2018.05.004

536 36. Kong Y, Rose CM, Cass AA, Williams AG, Darwish M, Lianoglou S, et al. Transposable element
537 expression in tumors is associated with immune infiltration and increased antigenicity. Nat
538 Commun. 2019;10: 5228. doi:10.1038/s41467-019-13035-2

539 37. Jang HS, Shah NM, Du AY, Dailey ZZ, Pehrsson EC, Godoy PM, et al. Transposable elements drive
540 widespread expression of oncogenes in human cancers. Nat Genet. 2019;51: 611–617.
541 doi:10.1038/s41588-019-0373-3

542 38. Ivancevic A, Chuong EB. Transposable elements teach T cells new tricks. PNAS. 2020;117: 9145–
543 9147. doi:10.1073/pnas.2004493117

544 39. Ewing AD, Smits N, Sanchez-Luque FJ, Faivre J, Brennan PM, Richardson SR, et al. Nanopore
545 Sequencing Enables Comprehensive Transposable Element Epigenomic Profiling. Molecular Cell.
546 2020;80: 915-928.e5. doi:10.1016/j.molcel.2020.10.024

547 40. Scott EC, Gardner EJ, Masood A, Chuang NT, Vertino PM, Devine SE. A hot L1 retrotransposon
548 evades somatic repression and initiates human colorectal cancer. Genome Res. 2016;26: 745–755.
549 doi:10.1101/gr.201814.115

550 41. Haring NL, van Bree EJ, Jordaan WS, Roels JRE, Sotomayor GC, Hey TM, et al. ZNF91 deletion in
551 human embryonic stem cells leads to ectopic activation of SVA retrotransposons and up-regulation
552 of KRAB zinc finger gene clusters. Genome Res. 2021;31: 551–563. doi:10.1101/gr.265348.120

553 42. Wang H, Xing J, Grover D, Hedges DJ, Han K, Walker JA, et al. SVA Elements: A Hominid-specific
554 Retroposon Family. Journal of Molecular Biology. 2005;354: 994–1007.
555 doi:10.1016/j.jmb.2005.09.085

556 43. Raiz J, Damert A, Chira S, Held U, Klawitter S, Hamdorf M, et al. The non-autonomous
557 retrotransposon SVA is trans -mobilized by the human LINE-1 protein machinery. Nucleic Acids
558 Research. 2012;40: 1666–1683. doi:10.1093/nar/gkr863

559 44. Tang WWC, Dietmann S, Irie N, Leitch HG, Floros VI, Bradshaw CR, et al. A Unique Gene Regulatory
560 Network Resets the Human Germline Epigenome for Development. Cell. 2015;161: 1453–1467.
561 doi:10.1016/j.cell.2015.04.053

562 45. Molaro A, Hodges E, Fang F, Song Q, McCombie WR, Hannon GJ, et al. Sperm Methylation Profiles
563 Reveal Features of Epigenetic Inheritance and Evolution in Primates. Cell. 2011;146: 1029–1041.
564 doi:10.1016/j.cell.2011.08.016

565 46. Lazaros L, Kitsou C, Kostoulas C, Bellou S, Hatzi E, Ladias P, et al. Retrotransposon expression and
566 incorporation of cloned human and mouse retroelements in human spermatozoa. Fertility and
567 Sterility. 2017;107: 821–830. doi:10.1016/j.fertnstert.2016.12.027

568 47. Georgiou I, Noutsopoulos D, Dimitriadou E, Markopoulos G, Apergi A, Lazaros L, et al.
569 Retrotransposon RNA expression and evidence for retrotransposition events in human oocytes.
570 Human Molecular Genetics. 2009;18: 1221–1228. doi:10.1093/hmg/ddp022

571 48. Gordon S, Akopyan G, Garban H, Bonavida B. Transcription factor YY1: structure, function, and
572 therapeutic implications in cancer biology. Oncogene. 2006;25: 1125–1142.
573 doi:10.1038/sj.onc.1209080

574 49. Pabst T, Mueller BU, Zhang P, Radomska HS, Narravula S, Schnittger S, et al. Dominant-negative
575 mutations of CEBPA, encoding CCAAT/enhancer binding protein-α (C/EBPα), in acute myeloid
576 leukemia. Nat Genet. 2001;27: 263–270. doi:10.1038/85820

577 50. Umek RM, Friedman AD, McKnight SL. CCAAT-Enhancer Binding Protein: A Component of a
578 Differentiation Switch. Science. 1991;251: 288–292. doi:10.1126/science.1987644

579 51. Clapier CR, Cairns BR. The Biology of Chromatin Remodeling Complexes. Annual Review of
580 Biochemistry. 2009;78: 273–304. doi:10.1146/annurev.biochem.77.062706.153223

581 52. Zubenko GS, Hughes HB. Effects of the G(−656)A variant on CREB1 promoter activity in a neuronal
582 cell line: Interactions with gonadal steroids and stress. Mol Psychiatry. 2009;14: 390–397.
583 doi:10.1038/mp.2008.23

584 53. Sirotkin AV, Benčo A, Mlynček M, Harrath AH, Alwasel S, Kotwica J. The involvement of the
585 phosphorylatable and nonphosphorylatable transcription factor CREB-1 in the control of human
586 ovarian cell functions. Comptes Rendus Biologies. 2019;342: 90–96. doi:10.1016/j.crvi.2019.03.002

587 54. Chen M, Zhang L, Cui X, Lin X, Li Y, Wang Y, et al. Wt1 directs the lineage specification of sertoli and
588 granulosa cells by repressing Sf1 expression. Development. 2016; dev.144105.
589 doi:10.1242/dev.144105

590 55. Grabowski M, Jungers WL. Evidence of a chimpanzee-sized ancestor of humans but a gibbon-sized
591 ancestor of apes. Nat Commun. 2017;8: 880. doi:10.1038/s41467-017-00997-4

592 56. Meyer TJ, Held U, Nevonen KA, Klawitter S, Pirzer T, Carbone L, et al. The Flow of the Gibbon LAVA
593 Element Is Facilitated by the LINE-1 Retrotransposition Machinery. Genome Biology and Evolution.
594 2016;8: 3209–3225. doi:10.1093/gbe/evw224

595 57. Okhovat M, Nevonen KA, Davis BA, Michener P, Ward S, Milhaven M, et al. Co-option of the
596 lineage-specific LAVA retrotransposon in the gibbon genome. PNAS. 2020;117: 19328–19338.
597 doi:10.1073/pnas.2006038117

598 58. Fueyo R, Judd J, Feschotte C, Wysocka J. Roles of transposable elements in the regulation of
599 mammalian transcription. Nat Rev Mol Cell Biol. 2022 [cited 19 Mar 2022]. doi:10.1038/s41580-
600 022-00457-y

601 59. Sundaram V, Choudhary MNK, Pehrsson E, Xing X, Fiore C, Pandey M, et al. Functional cis-
602 regulatory modules encoded by mouse-specific endogenous retrovirus. Nat Commun. 2017;8:
603 14550. doi:10.1038/ncomms14550

604 60. Sundaram V, Wysocka J. Transposable elements as a potent source of diverse cis-regulatory
605 sequences in mammalian genomes. Philosophical Transactions of the Royal Society B: Biological
606 Sciences. 2020;375: 20190347. doi:10.1098/rstb.2019.0347

607 61. Takahashi K, Yamanaka S. Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult
608 Fibroblast Cultures by Defined Factors. Cell. 2006;126: 663–676. doi:10.1016/j.cell.2006.07.024

609 62. Weintraub AS, Li CH, Zamudio AV, Sigova AA, Hannett NM, Day DS, et al. YY1 Is a Structural
610 Regulator of Enhancer-Promoter Loops. Cell. 2017;171: 1573-1588.e28.
611 doi:10.1016/j.cell.2017.11.008

612 63. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map
613 format and SAMtools. Bioinformatics. 2009;25: 2078–2079. doi:10.1093/bioinformatics/btp352

614 64. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-
615 seq aligner. Bioinformatics. 2013;29: 15–21. doi:10.1093/bioinformatics/bts635

616 65. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat
617 Biotechnol. 2016;34: 525–527. doi:10.1038/nbt.3519

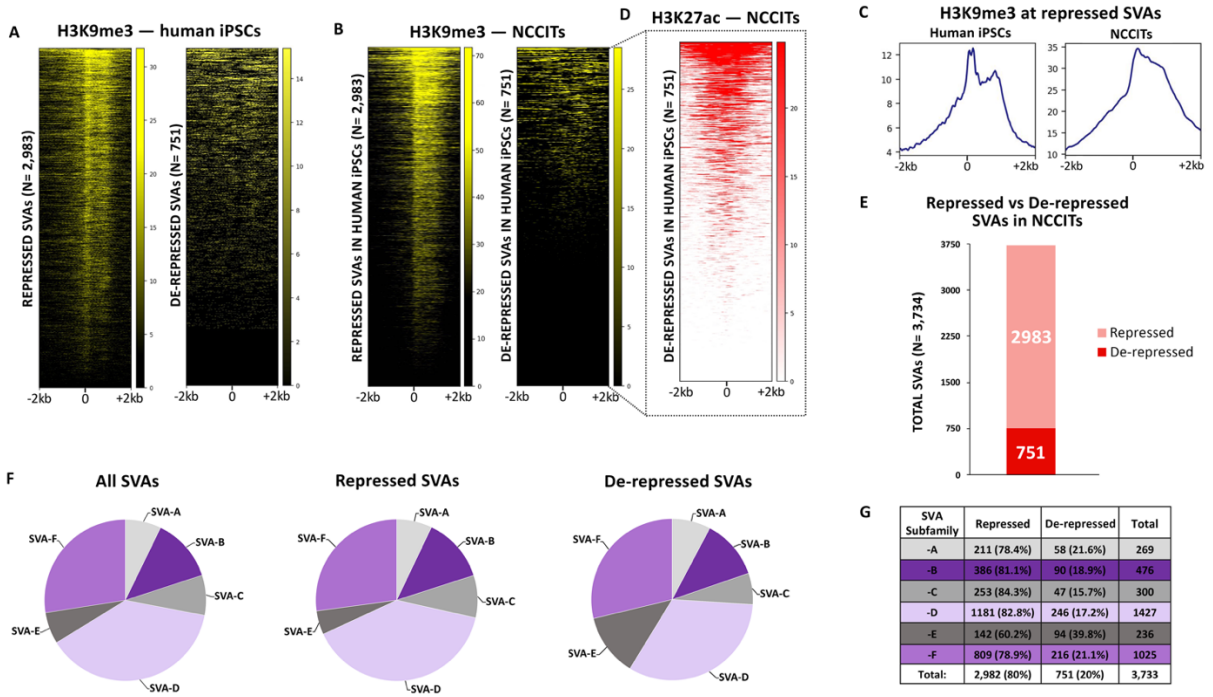
618 66. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq
619 data with DESeq2. Genome Biology. 2014;15: 550. doi:10.1186/s13059-014-0550-8

620 67. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features.
621 Bioinformatics. 2010;26: 841–842. doi:10.1093/bioinformatics/btq033

622 68. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple Combinations of Lineage-
623 Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B
624 Cell Identities. Molecular Cell. 2010;38: 576–589. doi:10.1016/j.molcel.2010.05.004

625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648

Figure 1



SVA Subfamily	Repressed	De-repressed	Total
-A	211 (78.4%)	58 (21.6%)	269
-B	386 (81.1%)	90 (18.9%)	476
-C	253 (84.3%)	47 (15.7%)	300
-D	1181 (82.8%)	246 (17.2%)	1427
-E	142 (60.2%)	94 (39.8%)	236
-F	809 (78.9%)	216 (21.1%)	1025
Total:	2,982 (80%)	751 (20%)	3,733

Figure 1 –751 human SVAs are active in human pluripotent cells. (A) Heatmaps depicting H3K9me3 ChIP-seq signal at all human SVAs in iPSCs. (B) ChIP-seq for H3K9me3 in human NCCITs at SVA regions previously classified as repressed and de-repressed in human iPSCs. (C) Average profiles of H3K9me3 enrichment at SVAs repressed in human iPSCs and NCCITs. (D) ChIP-seq for H3K27ac in NCCIT cells. The heatmap is centered on the 751 de-repressed SVAs. (E) Barplot representing the number of total repressed and de-repressed human SVAs (i.e. SVAs decorated with H3K9me3 versus lacking H3K9me3). 2982 of the ~3700 human SVAs were repressed, while 751 were de-repressed. (F) & (G) Human-specific SVAs, subfamilies SVA-E and -F are enriched within the de-repressed SVA population.

681 **Figure 2**

682

683 684 685 686 687 688 689 690 691 692 693 694 695 696 697 698 699 700 701 702 703

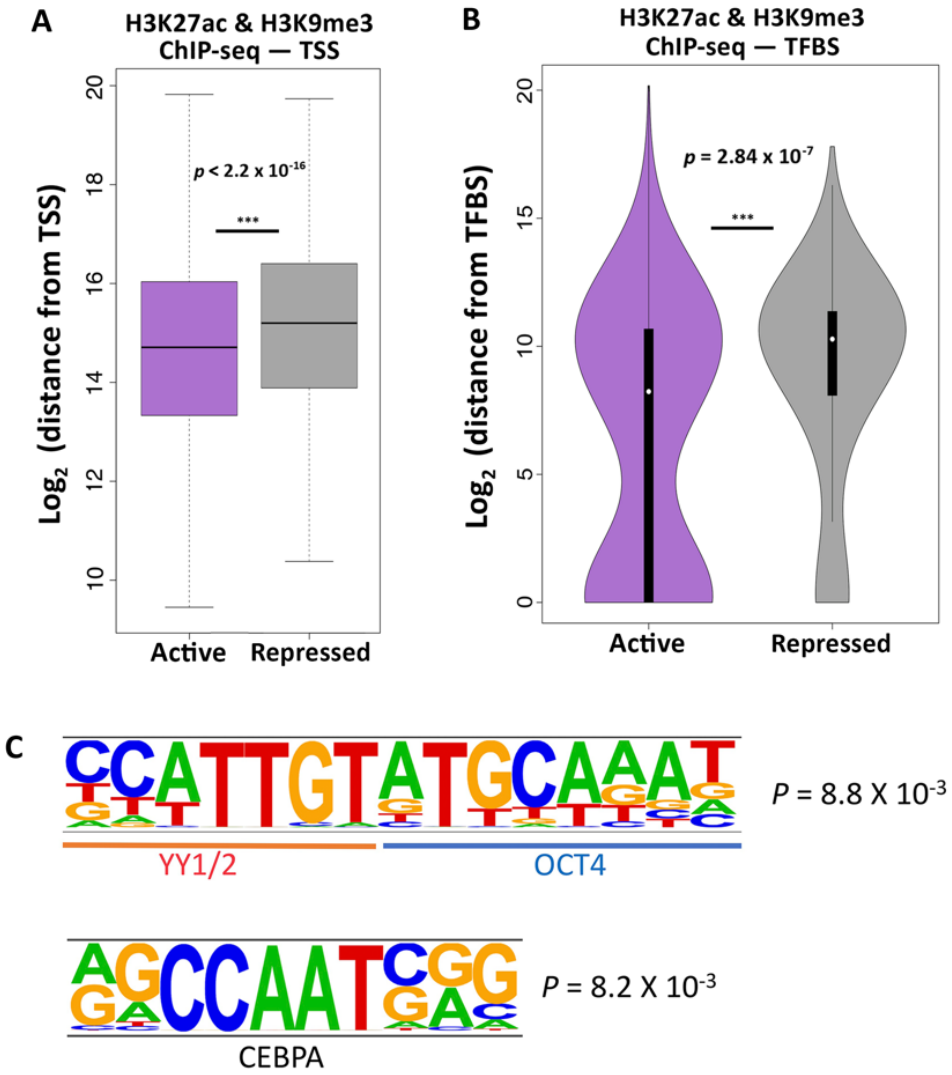


Figure 2 – Specific genomic features characterize the de-repressed SVAs. (A) SVAs in an active configuration are approximately 10 kb closer to TSS than SVAs in a repressed configuration (Wilcoxon's Rank Sum Test $p < 2.2 \times 10^{-16}$). (B) Active SVAs are significantly closer to, and directly overlap with, TFBS three times more than repressed SVAs (Wilcoxon's rank-sum test $p = 2.84 \times 10^{-7}$). (C) Motif analysis performed on the de-repressed SVAs shows enrichment for consecutive YY1/2 and OCT4 motifs. A CEBPA motif was also enriched.

710

711

712

713 **Figure 3**

714

715

716

717

718

719

720

721

722

723

724

725

726

727

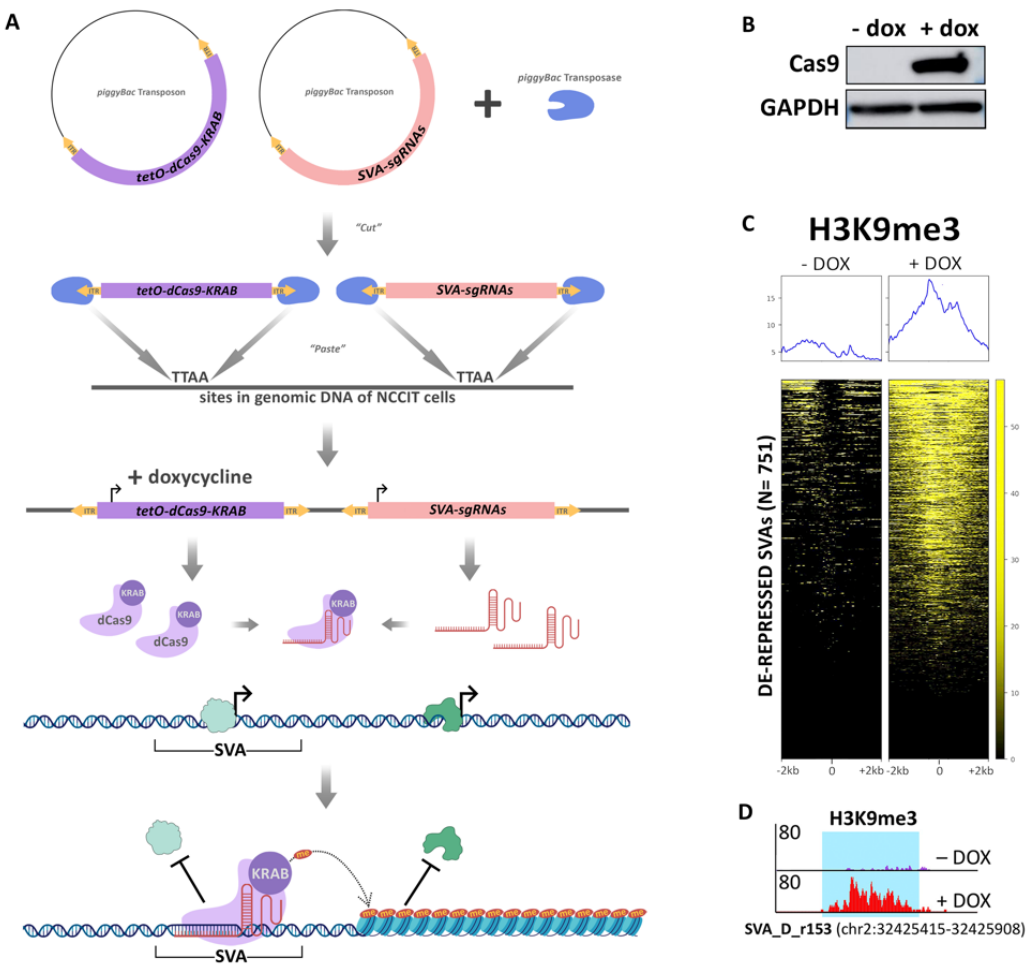
728

729

730

731

732



733

734

735 **Figure 3** – Epigenetic manipulation of NCCITs via CRISPRi results in SVA repression. (A) Schematic of the

736 generation of the SCi-NCCIT line (created with BioRender.com). (B) Cas9 immunoblot displaying activation

737 of dCas9-KRAB 72 hours post-doxycycline induction. (C) H3K9me3 ChIP-seq heatmap of the SCi-NCCITs

738 shows increased H3K9me3 signal post-doxycycline treatment. (D) Genome browser screenshot displaying

739 increased H3K9me3 before and after treatment with doxycycline at a de-repressed SVA-D.

740

741

742

743

744

Figure 4

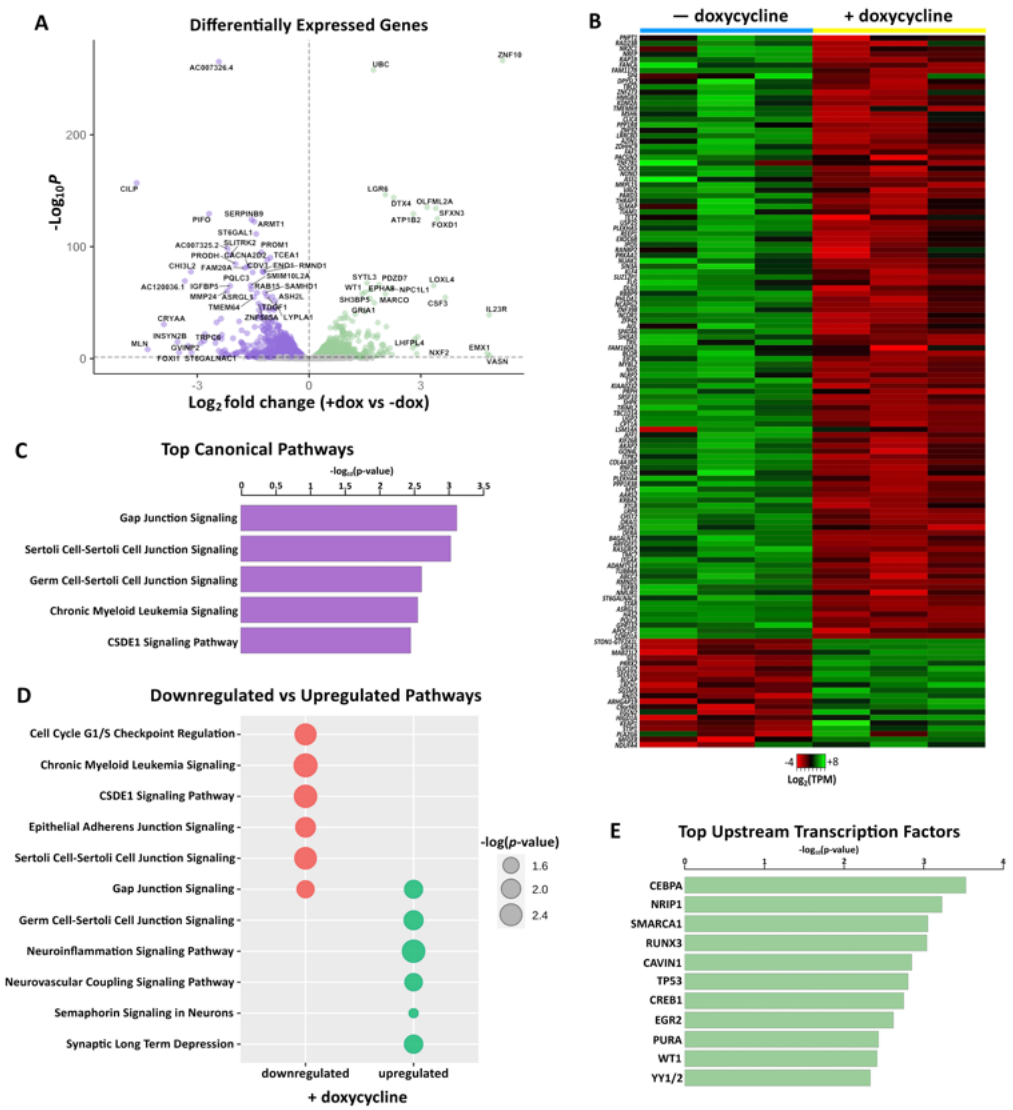


Figure 4 – CRISPRi-mediated repression of conventionally de-repressed SVAs results in aberrant gene expression. (A) Volcano plot showing genes differentially expressed in SCi-NCCITs after doxycycline treatment (purple = downregulated genes; green = upregulated genes). (B) Heatmap of 131 genes that are differentially expressed post-doxycycline treatment and also represent the nearest gene to an SVA. (C) Top canonical pathways predicted by IPA for the 131 genes differentially expressed after doxycycline treatment. (D) Canonical pathways predicted by IPA for the 111 downregulated (red) and 20 upregulated (green) genes. (E) Top upstream regulators/transcription factors predicted by IPA for the 131 genes differentially expressed after doxycycline treatment.

Figure 5

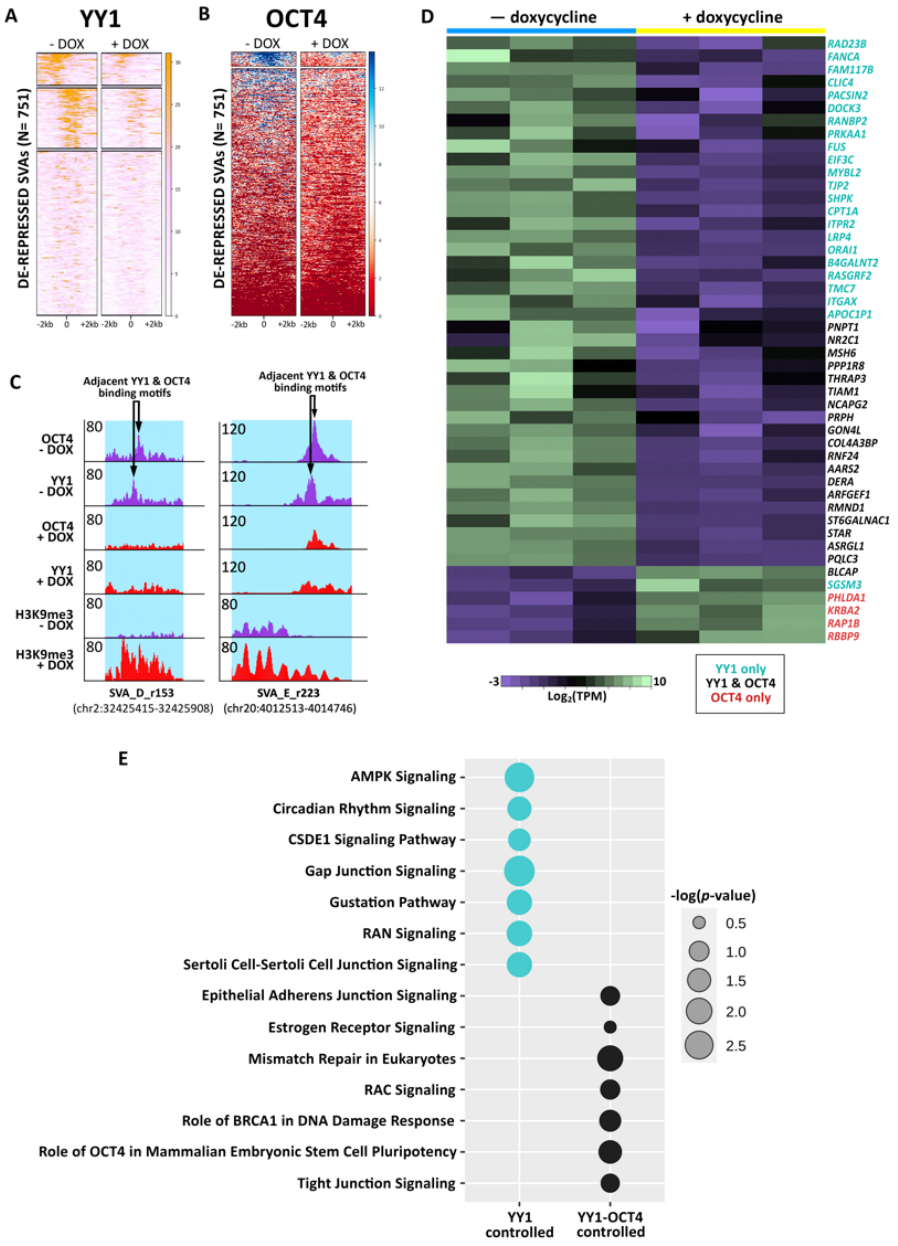
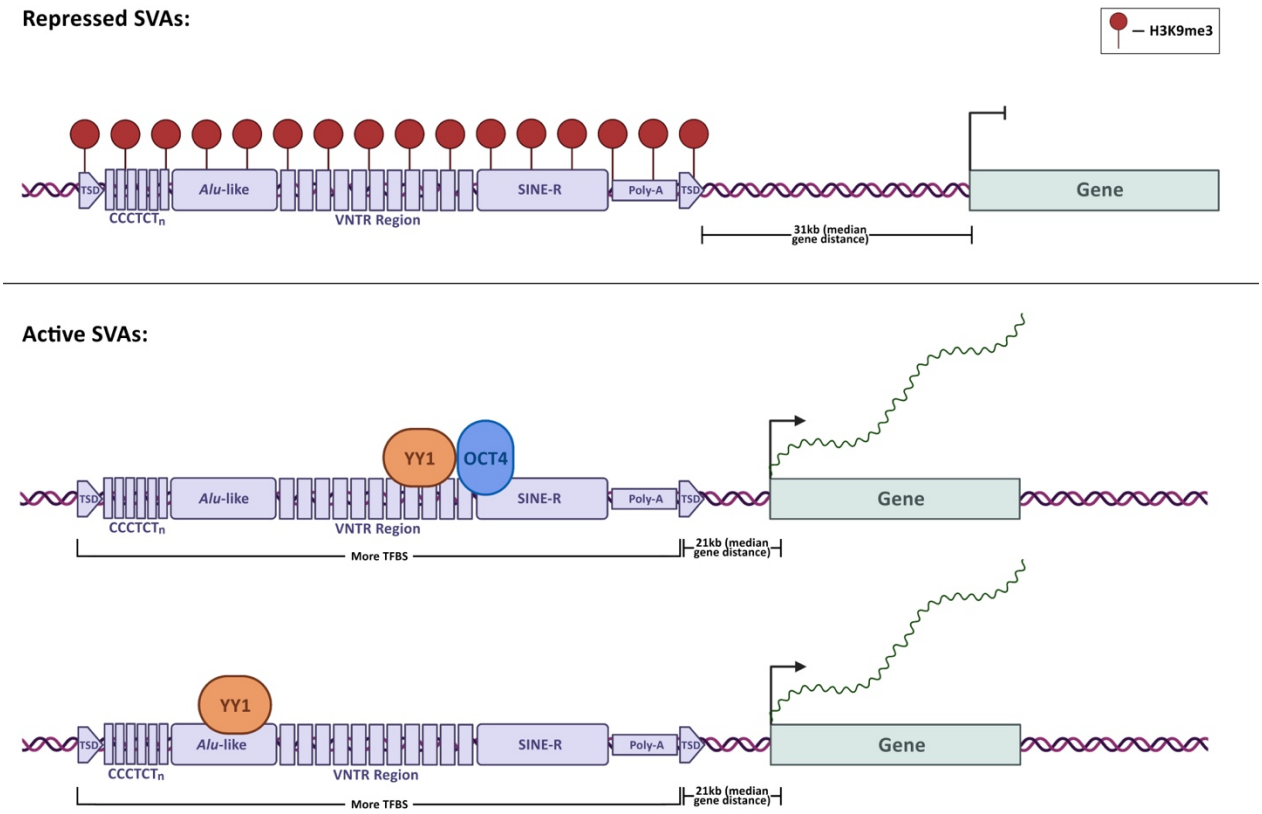


Figure 5 – Individual and sequential binding of YY1 and OCT4 contributes to SVA regulatory activity. (A) YY1 ChIP-seq signal at the 751 de-repressed SVAs before and after doxycycline treatment. (B) OCT4 ChIP-seq signal at the 751 de-repressed SVAs before and after doxycycline treatment. (C) Genome browser screenshot displaying the sequential binding of YY1 and OCT4 at a truncated, de-repressed SVA-D and a full-length, de-repressed, human-specific SVA-E. Upon doxycycline induction, the binding of both transcription factors is lost, while H3K9me3 signal is gained. (D) Heatmap of 47 genes that are differentially expressed upon doxycycline treatment and regulated by YY1 and/or OCT4 (Black: YY1 and

809 OCT4, Teal: YY1 only, Red: OCT4 only). (E) Canonical pathways predicted by IPA for the SVA-regulated
810 genes that are differentially expressed upon doxycycline treatment and regulated by YY1-only (teal) and
811 YY1-OCT4 (black).
812
813
814 **Figure 6**
815



816
817 **Figure 6** – Model for SVA co-option in human pluripotent stem cell gene regulation. This figure was created
818 with BioRender.com.