Cell Stem Cell, Volume *24*

Supplemental Information

Hominoid-Specific Transposable Elements and KZFPs Facilitate Human Embryonic Genome Activation and Control Transcription in Naive Human ESCs

Julien Pontis, Evarist Planet, Sandra Offner, Priscilla Turelli, Julien Duc, Alexandre Coudray, Thorold W. Theunissen, Rudolf Jaenisch, and Didier Trono

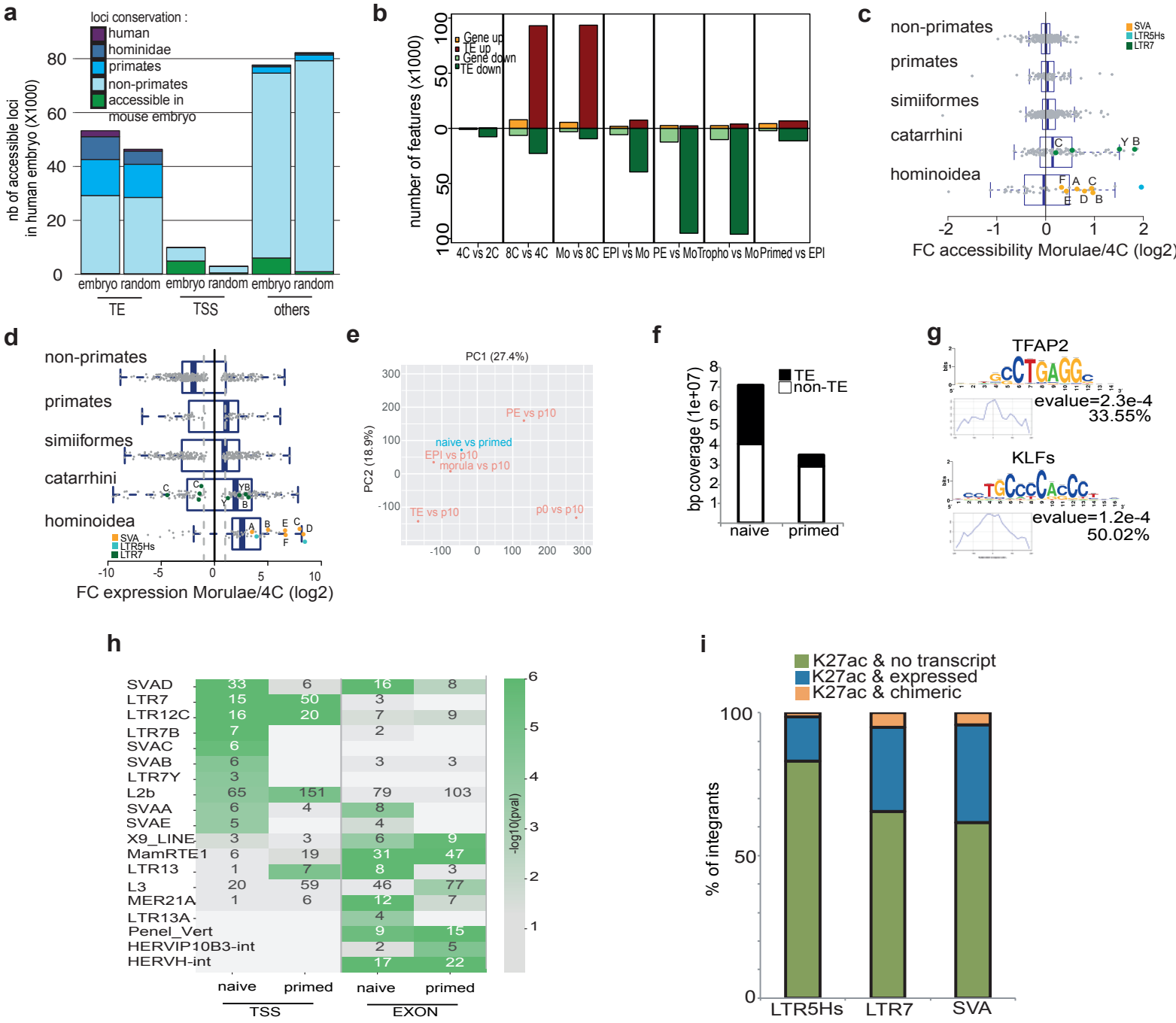

Figure S1 related to Figure 1. Evolutionarily recent TEs are activated during human EGA and in naïve hESC. (a) Genomic distribution of chromatin accessibility sites in human pre-implantation embryo. Observed DNase-seq enriched loci in human embryo (embryo) or averaged values from 10 randomly shuffled loci were overlapped with TSS of all coding genes (+/- 500 bp), TEs (> 50% overlap over a TE), or other genomic sites (others) as indicated. Loci were classified according to their conservation in other species as color-coded. Pale blue and green correspond respectively to loci conserved at the same syntenic location in the mouse and accessible during early murine embryogenesis (re-analyzed from (Gao et al., 2018)). (b) Early embryonic genome expression. Number of genes or TE integrants up- or down-regulated (fold change 2, adj. p-value < 0.05) during the indicated transitions in single RNA-seq data of human pre-implantation embryonic development: 2-cell (2C), 4-cell (4C), morula (Mo), epiblast (EPI), primitive endoderm (PE), trophectoderm (Tropho) and cell culture-derived ESCs after 10 passages (Primed). (c) TE subfamilies accessibility during EGA. Log2 fold-change of indicated TE subfamilies (classified by lineage restriction) using subfamily add-up of normalized read counts between 4C and morula. Green dots represent LTR7/HERVH integrants, with C, B and Y indicating the corresponding LTR7 subclasses, with and without their internal part. Cyan and orange dots represent LTR5Hs/HERVK and SVA subfamilies, respectively, with A, B, C, D, E and F designating SVA subclasses (re-analyzed from (Gao et al., 2018)). (d) TE subfamilies expression during EGA. Log2 fold-change of indicated TE subfamilies (classified by lineage restriction) using subfamily add-up of normalized read counts between 4C and morula. Green dots represent LTR7/HERVH integrants, with C, B and Y indicating the corresponding LTR7 subclasses, with or without their internal part. Cyan and orange dots represent LTR5Hs/HERVK and SVA subfamilies, respectively, with A, B, C, D, E and F designating SVA subclasses (re-analyzed from (Yan et al., 2013)). (e) Comparison of relative gene expression in pre-implantation embryo vs. naïve/primed hESCs. PCA analysis was performed on fold changes using single-cell RNA-seq data of human post-EGA pre-implantation embryonic development from (Yan et al., 2013) and the naive/primed hESC (Theunissen et al., 2016). Morula (morula), epiblast (EPI), primitive endoderm (PE), trophectoderm (TE) and cell culture-derived ESCs at harvest (p0) and after 10 passages (p10). Naïve cells were maintained in 4i/LA, KN/2i media and primed in hES media. (f) H3K27ac genomic coverage in naïve vs primed cells. Genomic coverage (bp overlap) of H3K27ac ChIP-seq peaks was computed in naïve (WIBR3 in 4i/LA media) and primed (WIBR3 in hES media) hESC over TEs and the rest of the genome. (g) Prominent naïve specific-accessibility sites. Naïve-specific ATAC-seq peaks (10-fold more than primed and p-value < 0.05) were analyzed with the RSAT motif discovery software. Most centered, represented and significant motifs are depicted. Percentage represents the number of accessible loci containing at least one motif. (h) TE-gene fusion transcripts. Heatmap representation of relative use of indicated TE subfamilies as alternative promoters or of their incorporation into exons in naïve and primed cells, with numbers corresponding to sum of integrants in each category and color scale to the relative over-representation (hypergeometric test) of that subfamily. TE subfamilies significantly over-represented (p-value < 0.05) in at least one category are plotted. (i) Histone acetylation and expression profiles of SVA/LTR5Hs/LTR7 integrants. All SVA/LTR5Hs/LTR7 sites enriched for H3K27ac in naïve hESCs were further subcategorized as transcribed (expressed), silent (no transcript) or part of a TE-gene fusion transcript (chimeric).

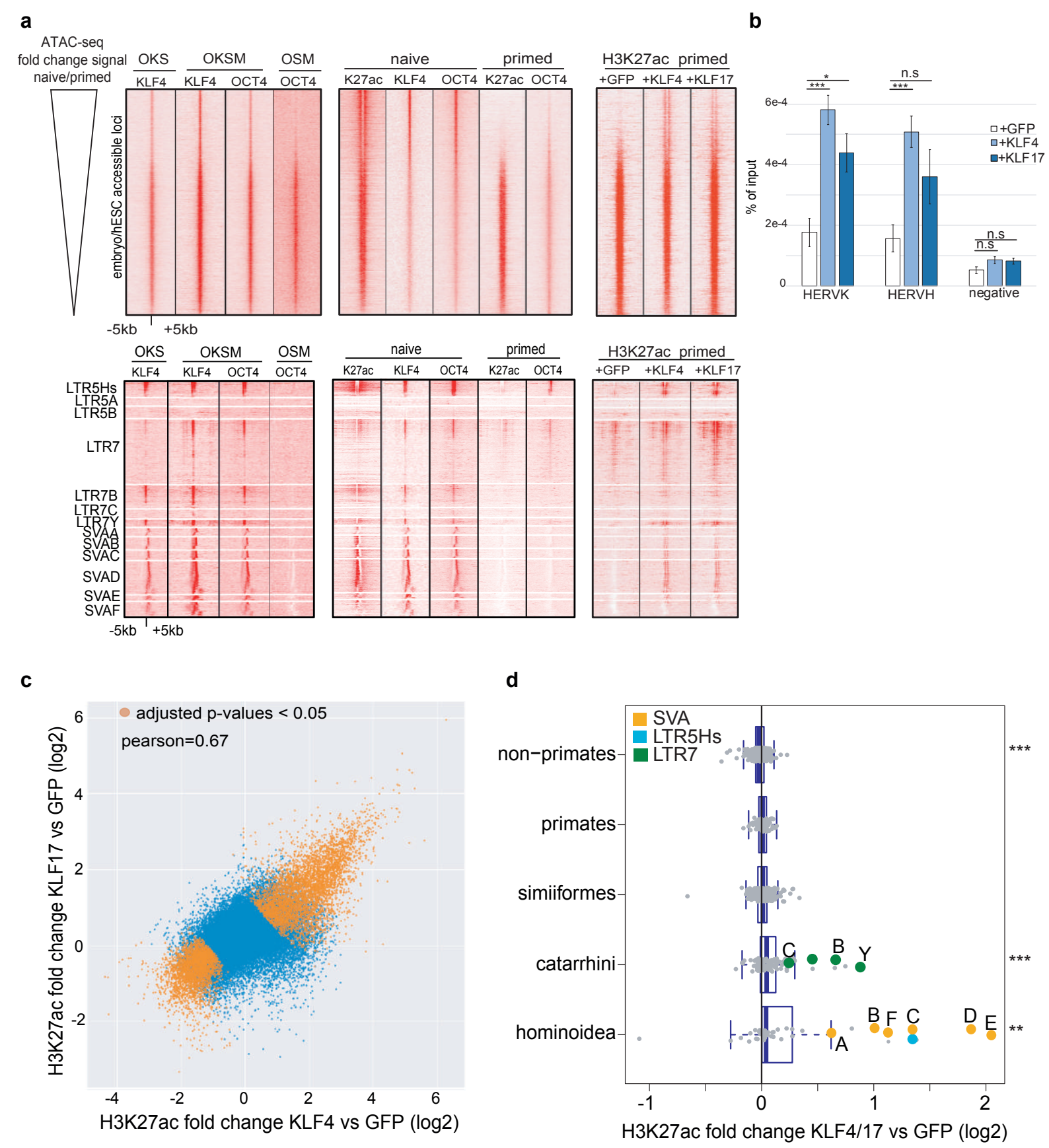

Figure S2 related to Figure 2. Krüppel-like factors are major inducers of EGA and naïve-specific TE enhancers. (a) Chromatin accessibility, H3K27ac enrichment and KLF4/OCT4 genomic recruitment. Upper panel, ChIP-seq raw data over 10kb window around H3K27ac-enriched and accessible chromatin genomic regions in pre-implantation embryo and hESCs (WIBR3 in 4i/LA, KN/2i or hES media) ordered in function of fold difference between naïve and primed ATAC-seq signals. Lower panel, ChIP-seq raw data over 10kb window around indicated TE loci. Left, OCT4 and KLF4 ChIP-seq peaks during early (2-4 days) OKSM/OSM-induced reprogramming of dermal fibroblast cells (Ohnuki et al., 2014) or OKS-induced of HAP1 cells (this study). Middle, OCT4 (Ji et al., 2016), KLF4 and H3K27ac ChIP-seq profiles (this study) in naïve and primed hESC (WIBR3). Right, H3K27ac ChIP-seq profiles in primed hES cells (H1) 5 days after transduction with GFP-, KLF4- or KLF17-expressing lentiviral vectors (this study). (b) KLFs induce OCT4 recruitment to TEs. OCT4 ChIP-qPCR was performed in primed hES cells (H1) 5 days after transduction with GFP-, KLF4- or KLF17-expressing lentiviral vectors. Error bars represent S.E.M of triplicate transduction while p-value was established with a t.test (*** ≤ 0.001, ** ≤ 0.01, * ≤ 0.05 and n.s > 0.05). (c) H3K27ac changes upon KLF4 and KLF17 overexpression in primed ESC cells. Scatter plot represents log2 fold change in H3K27ac signal between GFP and KLF4 (x-axis) and KLF17 (y-axis) -overexpressing H1 primed hESCs. Significant changes are highlighted in orange (adjusted p-value < 0.05); Pearson correlation was performed on all values. (d) KLFs-induced H3K27ac status of age-stratified human TEs in primed hESC using subfamily add-up of normalized read counts. *** for p-value ≤ 0.001 for the comparisons of each age category being different than 0 using t-test. Green dots represent LTR7/HERVH integrants, with C, B and Y indicating the corresponding LTR7 subclasses, with and without their internal part. Cyan and orange dots represent LTR5Hs/HERVK and SVA subfamilies, respectively, with A, B, C, D, E and F designating SVA subclasses.

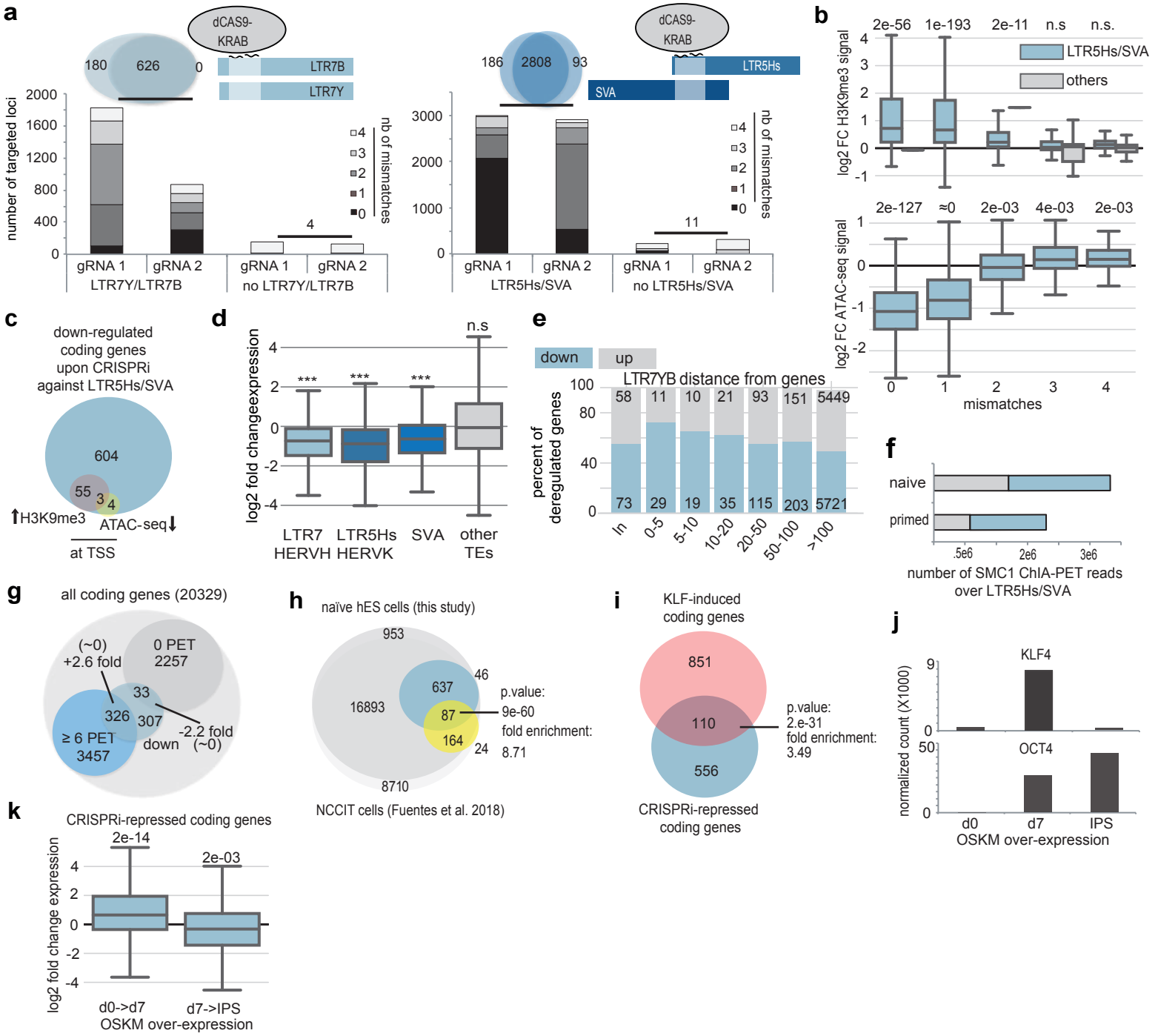

Figure S3 related to Figure 3. TEENhancers regulate gene expression in naïve hESCs. (a) Schematic representation of CRISPRi targeting of LTR7Y/B or SVA/LTR5Hs common region. Venn diagram depicts the range of TEs predicted to be targeted by either guide RNA, with bar plot representing the number of targeted genomic loci. Color scale of greys illustrates the number of mismatches between TE and guide RNA. (b) Fold-change H3K9me3 (top) and ATAC-seq (bottom) signals (average of replicates) over LTR5Hs/SVA integrants in control vs. LTR5Hs/SVA-targeting CRISPRi-modified naïve hESCs. Integrants were classified according to their complementarity to the guide RNA, from 0 to 4 mismatches. (c) Venn diagram of significantly down-regulated coding genes (blue circle), intersecting with fraction of all their known TSS exact coordinates displaying significantly (p-value < 0.05) decreased ATAC-seq signal (yellow circle) and increased H3K9me3 enrichment (pink circle). (d) Changes in TE expression in LTR7-directed CRISPRi-modified naïve hESC. Naïve hESCs were transduced with a dCAS9-KRAB lentiviral vector containing or not guide RNA against LTR7Y/B (two different guide RNA were used in duplicate each). Log2 fold change expression between the guide RNA expressing cells versus the empty condition of all expressed TE integrants belonging either to LTR7-HERVH, LTR5Hs-HERVK, SVA subfamilies or the other expressed TEs were plotted. P-values were established using one sample t.test (***: p.value ≤ 0.001; n.s: p.value > 0.05). (e) Impact of LTR7-targeting CRISPRi on gene expression. Percentage of up- and down-regulated genes (p-value <0.05 and >10% differences between paired replicates) at indicated distance from closest CRISPRi-targeted LTR7 integrant (In: TE within gene). (f) Cohesin ChIA-PET (SMC1) reads over LTR5Hs/SVA loci (+/-500bp, grey) and distal paired-end reads (blue). Data were re-analyzed from (Ji et al., 2016). (g) Number of genes displaying detected 3D interactions with SVA/LTR5Hs. We counted, in a 1kb window around all TSS, the number of reads (Paired-End Tags, PET) obtained from the other paired-end mapped reads overlapping SVA or LTR5Hs (+/-500b) re-analyzed from Cohesin ChIA-PET (SMC1) in naïve hESC (WIBR3 in 5iLA media, (Ji et al., 2016)). Interacting coding gene TSS were selected if ≥ 6 PET were detected and the non-interacted ones if no PET were detected. Venn diagram represents the intersection between SVA/LTR5Hs of the coding genes TSS down-regulated when using CRISPRi against LTR5Hs/SVA ("down" circle) and all genes interacting with SVA/LTR5Hs ("≥ 6 PET" circle). Permutation test was performed to obtain fold enrichments and p-values. (h) Overlap of LTR5Hs-controlled genes in naïve and teratocarcinoma cell line. Venn Diagram of genes expressed and deregulated genes upon LTR5Hs-tarteting CRISPR-modified NCCIT teratocarcinoma cells (Fuentes et al., 2018), using gene symbol annotation (blue) and naïve hESCs (this study, yellow). Hypergeometric test was used to compute the p-value. (i) LTR5Hs/SVA-targeted CRISPRi and KLFs overexpression regulate a common set of genes in hESC. Genes significantly (p-value < 0.05) up-regulated upon KLF4 and KLF17 overexpression in primed hESC (H1) were intersected with genes significantly down-regulated in LTR5Hs/SVA-CRISPRi-modified naïve hESC (WIBR3 in KN/2iL media). Hypergeometric test was used to compute the p-value. (j) KLF4 and OCT4 expression pattern during somatic reprogramming. Normalized read counts of KLF4 and OCT4 in human fibroblasts before (d0), after 7 days of OKSM overexpression (d7), and in induced-pluripotent cells (IPS) after 3 passages (re-analyzed from (Ohnuki et al., 2014)). (k) Expression profile of LTR5Hs/SVA-controlled genes during re-programming. Boxplots represent fold change expression of coding genes down-regulated by CRIPSRi targeting LTR5Hs/SVA in naïve hESC during indicated times of reprogramming of fibroblasts with OKSM (re-analyzed from (Ohnuki et al., 2014)).

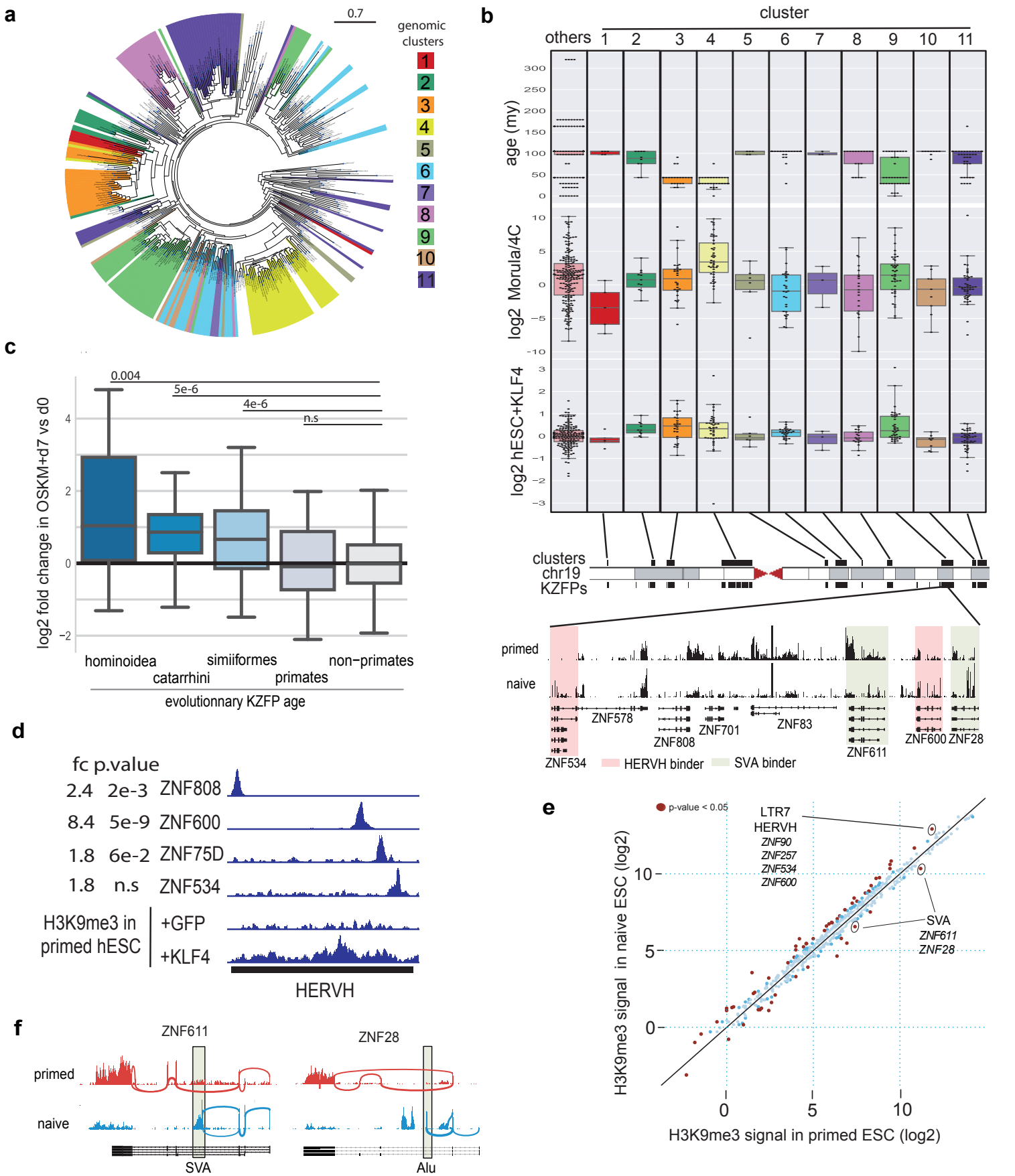

Figure S4 related to Figure 4. Evolutionary recent KZFPs tame TEENhancers during human early embryogenesis. (a) Phylogenic alignment of Human KFZPs. KZFP-coding ORFs (selected from UniProt database) were aligned, resulting in evolutionarily-related clusters. Colors correspond to genomic KZFP clusters depicted in (b). Scale represents the branch length as the expected number of substitutions per site computed by ete3. (b) Young KZFPs are embedded in clusters activated during EGA and by KLF4 overexpression in primed hESCs. Top, age distribution of KFZPs within indicated chromosome 19 clusters (1-11) or elsewhere in the genome (others); middle, KZFPs log2 fold change expression during EGA (morula vs 4-cell); bottom, KZFPs log2 fold change expression upon KLF4 overexpression in primed hESC (H1). Underneath is a screenshot of RNA-seq data from a cluster of TEENhancer-controlling KZFPs in naïve and primed hESCs, with magnification illustrating differential KZFP genes from this cluster, in naïve vs. primed hESCs. (c) Evolutionary young KZFP are activated by OKSM overexpression. Fold change expression of KZFP (classified by evolutionary age in x-axis) genes after 7 days post OKSM in human fibroblast (compare to “d0”). Data re-analyzed from (Ohnuki et al., 2014). (d) Young TEs display increased H3K9me3 enrichment upon KLF4 overexpression. Screenshot of a representative HERVH locus, illustrating its recruitment of several KZFPs. (Imbeault et al., 2017) and significantly increased level of H3K9me3 (adjusted p-value < 0.05) and KZFPs (with indicated fold changes and p-values) upon KLF4 over-expression in H1 cells. (e) Differential H3K9me3 signal (normalized read counts) over TE subfamilies in naïve vs. primed hESCs. Red dots correspond to p-values < 0.05. KZFPs having a positive and significant correlation of expression (between naïve and primed hESC) are depicted under their respective TE targeted subfamilies. (f) KZFPs are transcriptionally controlled by young transposable elements. Screenshot of sashimi plot of naïve and primed hESC transcriptome of ZNF611 and ZNF28 representing the gene expression with quantitative (line width) splice junctions.

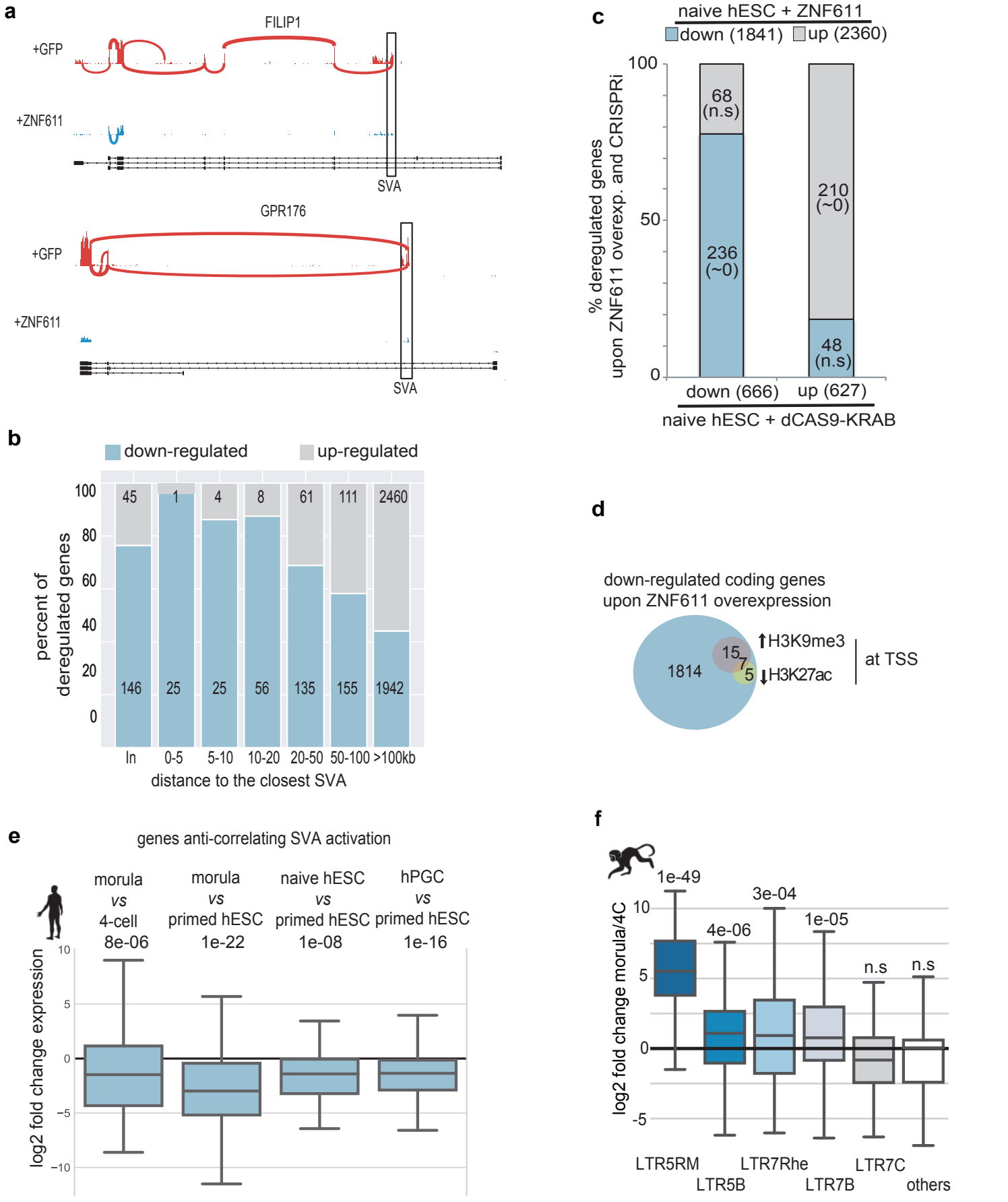

Figure S5 related to Figure 4 and 5. KZFPs and TEENhancers engineer species-specific early embryonic regulatory networks. (a) Two ZNF611-controlled SVA-driven chimeric transcripts. Screenshot of sashimi plot of naïve hESC transcriptome overexpressing GPF or ZNF611, depicting gene expression with relative use (reflected by line width) of splice junctions. (b) Percentage of genes up-or down-regulated (p-value <0.05) by ZNF611 overexpression in naïve hESC as a function of their distance (in kb) from targeted TEs. (In, TE inside gene). (c) Overlap between coding genes deregulated by both ZNF611 overexepression and LTR5Hs/SVA-targeting CRISPRi in naïve hESCs. The number in brackets refers to the total number of coding gene down-deregulated (down) or up-regulated (up) in each experiment (using p-value < 0.05), and p-value (in parenthesis, n.s, non-significant) was applied on these intersections with a hypergeometric test. (d) Venn diagram of the number of significantly deregulated coding genes, with indication of those with increased H3K9me3 or decreased H3K27ac signal (adjusted p.value < 0.05) at their TSS upon ZNF611 overexpression in naïve hESCs. (e) Fold-change expression of the two hundred genes repressed by both LTR5Hs/SVA-targeted CRISPRi and ZNF611 overexpression in naïve hESC (SVA-enhanced genes), comparing morula to 4C or primed hESC, and naïve hESCs and hPGCs to primed hESCs as indicated, reanalyzing data from (Tang et al., 2015; Theunissen et al., 2016; Yan et al., 2013). (f) TE expression during rhesus macaque EGA. Fold change expression of all expressed TE integrants from LTR7-HERVH or LTR5-HERVK subfamilies. Data were re-analyzed from (Wang et al., 2017) between 4-cell and morula stages of rhesus macaque.

symbol	Naive (log2 norm count)	Primed (log2 norm count)	fc Naive vs Primed (log2)	fc 4-cell vs Morula (log2)	fc Primed vs Morula (log2)
CTSF	11.38	5.60	5.78	6.68	5.59
KLF4	13.21	7.43	5.79	8.22	4.04
SEPT12	7.06	1.22	5.84	6.64	5.76
PRAMEF19	5.21	-0.68	5.88	10.72	12.75
APOBR	8.22	2.05	6.17	7.48	7.87
FUT3	6.94	0.72	6.22	10.89	10.87
BIN2	7.28	1.03	6.25	7.15	7.17
TUBB4A	16.27	9.98	6.28	9.67	3.97
CCL28	7.30	0.93	6.37	5.84	4.95
TINCR	11.22	4.74	6.47	9.68	7.55
FUT6	5.81	-0.68	6.48	7.00	6.17
TRIB3	14.08	7.54	6.54	9.07	3.27
LUZP4	6.22	-0.35	6.57	7.03	5.86
BTLA	7.08	-0.06	7.14	7.54	6.63
COX7B2	6.51	-0.68	7.18	9.46	8.58
INSM1	10.49	3.27	7.22	5.81	3.81
PRR23B	6.66	-0.99	7.65	6.82	6.41
RPL10L	7.21	-0.68	7.88	9.92	9.03
SUSD2	15.36	7.46	7.90	7.27	6.75
TPRX1	7.98	-0.04	8.02	12.61	13.67
ARGFX	7.66	-0.68	8.33	14.20	14.01
MAGEB2	10.62	2.17	8.44	8.21	7.41
KHDC1L	9.42	0.59	8.83	9.55	11.73
ZNF676	7.84	-0.99	8.83	6.64	1.65
ZNF208	7.96	-0.99	8.95	8.96	0.84
RTP3	10.20	0.89	9.31	6.96	7.65
ZNF528-AS1	8.33	-0.99	9.32	7.86	0.38
FAM151A	12.90	3.53	9.37	13.69	13.13
KLF17	10.96	1.30	9.66	11.25	12.99
ZFP42	14.33	4.67	9.67	10.14	-0.70
ZNF729	9.52	-0.68	10.20	8.26	5.26
ALPPL2	16.96	6.69	10.27	9.63	10.18
DNMT3L	16.52	5.65	10.87	9.14	7.78
POU5F1	16.86	15.53	1.33	3.57	-2.51

Table S1 related to Figure 2. Gene expression in naïve/primed hESC and in human embryo.

Ontology	P.value	Gene names
Transcription factor activity, sequence-specific DNA binding	6.9E-03	PPARD, ZFP42, ZNF232, ELK3, NFKB2, TBPL2, NPAS1, ZNF696, ZNF540, E4F1, ZNF607, CREBL2, ZNF449, ZNF641, KLF7, ZNF568, TAF4B, ZNF91, ZHX1, ZFP3, ZNF793, DDIT3, TAF13, ZNF432, PBX3, NFE2L3, ZNF571
Krueppel-associated box	2.7E-02	ZNF641, HKR1, ZNF566, ZNF100,ZNF729, ZNF486, ZNF568, ZNF91, ZNF540, ZNF793, ZNF432, ZNF607, ZNF571
Regulation of cell proliferation	7.5E-02	FA2H, TGFB3, EGLN3, TNK2, FANCA, RBBP9, TEC
Cell cycle arrest	7.5E-02	CDKN1C, PRKAA1, PPP1R15A, UHMK1, VASH1, DDIT3
Positive regulation of phagocytosis	7.7E-02	ABR, PYCARD, MERTK
Cell surface	5.1E-03	PVR, DEFB124, AIMP1, DEFB123, ITGA1, TGFB3, HILPDA, CDH5, LRPAP1, PROM1, WNT4, PRLR, VEGFA, FAM234A, RC3H2, SCARA5, LRP4, SLC9A1
Cell adhesion	5.9E-02	PVR, PTPRM, AIMP1, NRXN2, INPPL1, ITGA1, PTPRU, CDH5, LAMA1, RNASE10, GP5, DST, CDH24
Cell-cell adherens junction	6.7E-02	PVR, EFHD2, LIMA1, PTPRM, MACF1, ZC3HAV1, SNX5, CD2AP, PUF60, EHD4
Maintenance of cell polarity	7.7E-02	DST, SLC9A1
Defense response to virus	4.9E-02	TRIM56, AIMP1, ZC3HAV1, F2RL1, PYCARD, SAMHD1, DNAJC3
Wnt signaling pathway	3.0E-02	SENP2, WNT4, MACF1, PRKAA1, LGR6, DST, DDIT3, LRP4
Mitochondrial respiratory chain complex I assembly	7.9E-02	OXA1L, NDUFAF7, NDUFV1, NDUFAB1

Table S2 related to Figure 3. Gene Ontology of LTR5HS/SVA-controlled genes.

sgRNA against LTR5Hs/SVA	CTCCCTAATCTCAAGTACCC
sgRNA against LTR5Hs/SVA	TGTTTCAGAGAGCACGGGGT
sgRNA against LTR7YB	AAAGTACCTTCTTAAGGGTG
sgRNA against LTR7YB	AATCTCCCCCACCCTTAAGA
HERVK - ChIP/RT-qPCR	Fw:AGAGGAAGGAATGCCTCTTGCAG, Rv:TTACAAAGCAGTATTGCTGCCCGC
HERVH - ChIP/RTqPCR	Fw:GCAGCCTTTCCTTGGTGTTTAA, Rv:GCGTGGTCTGACACCTCTGA
Negative - ChIP	Fw:AAAGCTGGACTGGTGAATGC, Rv:TCAAAGGCTCATCTTTGCAG
ZFP42 - RTqPCR	Fw:GGAATGTGGGAAAGCGTTCGT, Rv:CCGTGTGGATGCGCACGT
LRP4 - RTqPCR	Fw:TGCAGTGAGTGCTCTTGGAG, Rv:TGCTGAGGGACAGTTCTCCT
PRODH - RTqPCR	Fw:CCCTGCTTCGGCACTACAG, Rv:GGGCCTGGTATTGCTTGTCC
ST6GAL1 - RTqPCR	Fw:AACTCTCAGTTGGTTACCACAGA, Rv:GGTGCAGCTTACGATAAGTCTT
MSTO1 - RTqPCR	Fw:CGAGCGACCGATTCCAAGG, Rv:CCAACTGTTTGTCCCTGTAGAG
C9ORF135 - RTqPCR	Fw:ATGGATAGCCTTGACAGATCCT, Rv:CCGGCTTCGAGTATTTCTTGTG
OVOL1 - RTqPCR	Fw:TGAACATGAGCCTTCGAGACT, Rv:CAAGGGTCACCTTCATCTTGG
GLS2 - RTqPCR	Fw:GCCTGGGTGATTTGCTCTTTT, Rv;CCTTTAGTGCAGTGGTGAACTT
BCDIN3D - RTqPCR	Fw:CTCGACGTGGGGTGTAACTC, Rv:GTTTCCCCGTCAGGTAGG
GAPDH - RTqPCR	Fw:CGAGATCCCTCCAAAATCAA, Rv:ATCCACAGTCTTCTGGGTGG
RPLP0 - RTqPCR	Fw:GCTTCCTGGAGGGTGTCC, Rv:GGACTCGTTTGTACCCGTTG

Table S3 related to Figure 3 and 4. Oligonucleotides.