LETTER

doi:10.1038/nature14308

Intrinsic retroviral reactivation in human preimplantation embryos and pluripotent cells

Edward J. Grow[1], Ryan A. Flynn[2], Shawn L. Chavez[3,4,5], Nicholas L. Bayless[6], Mark Wossidlo[1,3,4], Daniel J. Wesche[3], Lance Martin[2], Carol B. Ware[7], Catherine A. Blish[8], Howard Y. Chang[2], Renee A. Reijo Pera[1,3,4,9] & Joanna Wysocka[3,10,11]

Endogenous retroviruses (ERVs) are remnants of ancient retroviral infections, and comprise nearly 8% of the human genome[1]. The most recently acquired human ERV is HERVK(HML-2), which repeatedly infected the primate lineage both before and after the divergence of the human and chimpanzee common ancestor[2,3]. Unlike most other human ERVs, HERVK retained multiple copies of intact open reading frames encoding retroviral proteins[4]. However, HERVK is transcriptionally silenced by the host, with the exception of in certain pathological contexts such as germ-cell tumours, melanoma or human immunodeficiency virus (HIV) infection[5–7]. Here we demonstrate that DNA hypomethylation at long terminal repeat elements representing the most recent genomic integrations, together with transactivation by OCT4 (also known as POU5F1), synergistically facilitate HERVK expression. Consequently, HERVK is transcribed during normal human embryogenesis, beginning with embryonic genome activation at the eight-cell stage, continuing through the emergence of epiblast cells in preimplantation blastocysts, and ceasing during human embryonic stem cell derivation from blastocyst outgrowths. Remarkably, we detected HERVK viral-like particles and Gag proteins in human blastocysts, indicating that early human development proceeds in the presence of retroviral products. We further show that overexpression of one such product, the HERVK accessory protein Rec, in a pluripotent cell line is sufficient to increase IFITM1 levels on the cell surface and inhibit viral infection, suggesting at least one mechanism through which HERVK can induce viral restriction pathways in early embryonic cells. Moreover, Rec directly binds a subset of cellular RNAs and modulates their ribosome occupancy, indicating that complex interactions between retroviral proteins and host factors can fine-tune pathways of early human development.

Given the substantial contribution of transposons to the human genome and their emerging roles in shaping host regulatory networks[8,9], understanding the dynamic expression and function of these genomic elements is important for dissecting both human- and primate-specific aspects of gene regulation and development. We used published single-cell RNA-sequencing (RNA-seq) data sets to analyse the expression of major transposon classes at various stages of human preimplantation embryogenesis[10], a developmental period associated with dynamic changes in DNA methylation and transposon expression[11]. This analysis revealed two major clusters, one primarily consisting of repeats that begin to be transcribed at the onset of embryonic genome activation (EGA), which in humans occurs around the eight-cell stage, and a second cluster of repeats, whose transcripts can be detected in the embryo before EGA, indicating maternal deposition (Extended Data Fig. 1a). Within each cluster, more discrete stage-specific changes in repeat transcription could be observed, such that analysis of the repetitive transcriptome alone was able to distinguish pre- and post-EGA cells, as well as eight-cell/morula cells from blastocyst cells (Extended Data Fig. 1a). For example, HERVK and its regulatory element, long terminal repeat (LTR)5HS, were both induced in eight-cell stage embryos, morulae, and continued to be expressed in epiblast cells of blastocysts (Fig. 1a–c and Extended Data Fig. 1a). We further observed that although HERVK was expressed in blastocyst outgrowths (passage 0 human embryonic stem (ES) cells), it was downregulated by passage 10 (Fig. 1d). In contrast, transcripts of another HERV, HERVH, and of its regulatory element LTR7, were detected before EGA and throughout preimplantation development, including in all blastocyst lineages and human ES cells (Extended Data Fig. 1a–c).

Recent studies have reported conditions for capturing a human naive pluripotent state *in vitro*[12–16], and we used RNA-seq to analyse the repetitive transcriptome of ELF1, a cell line derived from an eight-cell-stage human embryo under naive culture conditions, and compared it to the repeat expression in ELF1 cells matured *in vitro* into a primed state[14]. Surprisingly, although many transposon classes (for example, HERVH and LINE1-HS) were highly expressed in both cell states, only a few showed differential levels between the two (Fig. 1e). In particular, transcripts corresponding to HERVK proviruses and their regulatory elements, LTR5HS (but not the older LTR5a or LTR5b; see later), were among the most strongly induced in naive versus primed ELF1 cells (Fig. 1e and Extended Data Fig. 1d). Similar results were obtained by analysing available transcriptomes of primed H1 human ES cells and naive 3iL cells derived from them, as well as of primed H9 human ES cells and those 'reset' to the naive state by *NANOG* and *KLF2* transgene expression[12,15](Fig. 1e). Therefore, naive-state-specific upregulation of HERVK is consistent across multiple genetic backgrounds, derivation methods or culture conditions.

From an evolutionary perspective, HERVK is especially interesting, as it is the most recently acquired HERV from which multiple insertions have retained protein-coding potential[17] (Extended Data Fig. 2a). While HERVK is present in all Old World primates, nearly a third of its proviruses in the human genome represent human-specific insertions, and 48% of those show polymorphisms in the human population, suggesting that HERVK was active within the last 200,000 years[18] (Extended Data Fig. 2a). All human-specific and human-polymorphic HERVK elements are regulated by a specific LTR subgroup, LTR5HS, whereas insertions representing older integrations typically have regulatory elements of the LTR5a or LTR5b subtype[4] (Extended Data Fig. 2a). Interestingly, during human preimplantation development and in the naive state, transcripts originating from LTR5HS, but not LTR5a or LTR5b, are preferentially expressed

[1]Department of Genetics, Stanford University School of Medicine, Stanford, California 94305, USA. [2]Howard Hughes Medical Institute and Program in Epithelial Biology, Stanford University School of Medicine, Stanford, California 94305, USA. [3]Institute for Stem Cell Biology & Regenerative Medicine, Stanford University School of Medicine, Stanford University, Stanford, California 94305, USA. [4]Department of Obstetrics and Gynecology, Stanford University School of Medicine, Stanford University, Stanford, California 94305, USA. [5]Division of Reproductive and Developmental Sciences, Oregon National Primate Research Center, Oregon Health & Science University, Beaverton, Oregon 97006, USA. [6]Stanford Immunology, Stanford University School of Medicine, Stanford, California 94305, USA. [7]Department of Comparative Medicine, University of Washington, Seattle, Washington 98195-8056, USA. [8]Department of Medicine, Stanford University School of Medicine, Stanford, California 94305, USA. [9]Department of Cell Biology and Neurosciences, Montana State University, Bozeman, Montana 59717, USA. [10]Department of Chemical and Systems Biology, Stanford University School of Medicine, Stanford, California 94305, USA. [11]Department of Developmental Biology, Stanford University School of Medicine, Stanford, California 94305, USA.

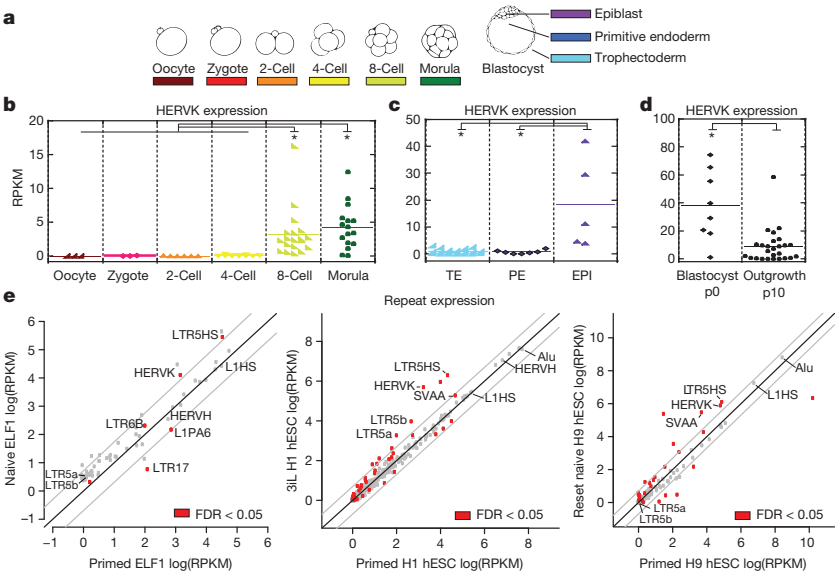


Figure 1 | Transcriptional reactivation of HERVK in human preimplantation embryos and naive human ES cells. **a**, Schematic of human preimplantation development. **b**, HERVK expression in single cells of human embryos at indicated stages. Solid line indicates mean. Oocyte ($n = 3$), zygote ($n = 3$), 2-cell ($n = 6$), 4-cell ($n = 11$), 8-cell ($n = 19$), morula ($n = 16$). **b–d**, Data are taken from ref. 10. *P value < 0.05, non-paired Wilcoxon test. RPKM, reads per kilobase per million. **c**, HERVK expression in single cells of human blastocysts, grouped by lineage. Solid line indicates mean. Trophectoderm (TE; $n = 18$), primitive ectoderm (PE; $n = 7$), epiblast (EPI; $n = 5$). **d**, HERVK expression in single cells of blastocyst outgrowths (passage (p)0) or human ES cells at passage (p)10. Solid line indicates mean. p0 ($n = 8$), p10 ($n = 26$). **e**, Analysis of the repetitive transcriptomes of three, genetically matched naive/primed human ES cell pairs. Left, naive/primed ELF1 human ES cells (data from this study) ($n = 3$ biological replicates for both conditions). Middle, 3iL/primed H1 human ES cells (data are taken from ref. 12) ($n = 3$ biological replicates for both conditions). Right, naive/primed H9 human ES cells (data are taken from ref. 15) ($n = 3$ biological replicates for both conditions). Significant repeats indicated in red at false discovery rate (FDR) < 0.05, DESeq. hESC, human ES cells.

(Fig. 1e), and we observed an upregulation of human-specific proviruses compared to evolutionarily older elements (Fig. 2a). We hypothesized that this differential regulation can be explained by *cis*-regulatory change in LTR5HS. Indeed, sequence analysis uncovered an OCT4 motif at position 693–699 base pairs (bp) of LTR5HS, which was conserved across diverse LTR5HS sequences, but not present in LTR5a/LTR5b, despite their overall high (~88%) sequence homology with LTR5HS (Fig. 2b and Extended Data Fig. 2a). To test whether OCT4 binding contributes to the transcriptional activation of LTR5HS, we used pluripotent NCCIT human embryonic carcinoma (EC) cells, which express OCT4, but, in contrast to human ES cells, are permissive for HERVK expression[5,19] (Extended Data Fig. 2b–d). Chromatin immunoprecipitation with quantitative polymerase chain reaction (ChIP-qPCR) analysis of human EC cells showed preferential occupancy of OCT4, p300 and histone marks of active chromatin at LTR5HS elements, as compared to LTR5a/LTR5b (Fig. 2c). In contrast, we did not detect OCT4 or p300 binding at LTR5HS in primed human ES cells (Extended Data Fig. 2f). Consistent with a functional role in HERVK activation, knockdown of *OCT4* or *SOX2*, but not of *NANOG*, led to a significant decrease in viral transcripts in human EC cells (Extended Data Fig. 2e and Extended Data Fig. 3a). Furthermore, the activity of transcriptional reporters driven by LTR5HS was impaired by mutations in the OCT4 motif (Fig. 2d and Extended Data Fig. 3b).

The aforementioned observations are consistent with transactivation by OCT4 being a driver of LTR5HS regulatory activity, but do not explain the differential transcriptional status of HERVK in primed versus naive human ES cells and in human EC cells, as all three express OCT4. We hypothesized that DNA methylation may contribute an additional layer of regulation, and indeed we observed HERVK hypomethylation of solo and proviral LTR5HS (but not the Gag open reading frame (ORF)) in human EC cells and naive ES cells, as compared to primed human ES cells and human induced pluripotent stem cells (iPSCs) (Fig. 2e and Extended Data Fig. 3c, d). Strong and preferential demethylation of LTR5HS was also observed in recently published DNA methylation maps from human preimplantation embryos, whereas HERVK coding sequences remained more highly methylated[11]. Importantly, treatment of primed human ES cells with a DNA methylation inhibitor, 5-aza-2′-deoxycytidine, for 24 h induced HERVK transcription, with 8–12-fold upregulation of an early transcript encoding an accessory protein, Rec (Fig. 2f). In addition, inhibition of DNA methylation, together with overexpression of *OCT4* and *SOX2*, jointly facilitated HERVK transcription in HEK293 cells (Fig. 2g and Extended Data Fig. 3e), indicating that DNA hypomethylation and transactivation by OCT4 synergistically promote HERVK expression.

A defining characteristic of HERVK is that multiple proviruses have retained ORFs encoding full-length retroviral proteins[4]. Consequently, HERVK reactivation in pathological conditions has been associated with the presence of HERVK proteins[5–7], prompting us to examine whether retroviral proteins are also present in human embryos. We used a well-characterized monoclonal antibody recognizing the HERVK Gag precursor and its proteolytically processed form Capsid, which detects cytoplasmic signal with a characteristic punctate pattern in human EC cells and in a subset of naive ELF1 cells, but shows no staining in primed human ES cells and loss of signal in human EC cells after *gag* short interfering RNA (siRNA) knockdown (Extended Data Fig. 4a–d). In human blastocysts, Gag/Capsid staining was also detected in dense cytoplasmic puncta resembling those seen in human EC cells and naive ELF1 cells (Fig. 3a and Extended Data Fig. 4a, d, e), with all analysed blastocysts ($n = 19/19$) showing a robust signal. Several HERVK-positive human EC cell lines have been shown to produce viral-like particles (VLPs)[20]. Remarkably, heavy metal staining transmission electron microscopy (TEM) of blastocysts revealed the presence of cytoplasmic, electron-dense particles of approximately 100 nm in diameter—the reported size of reconstructed HERVK VLPs—with electron-lucent cores[21,22] (Fig. 3b). Additionally, human blastocyst cells also contained cytosolic vesicles enclosing 50 or more smaller, highly electron-dense particles of approximately 75 nm in size, which resembled the immature VLPs also seen in human EC cells (Fig. 3c and Extended Data Fig. 5a). The presence of HERVK-derived particles in human blastocysts was further supported by immuno-gold TEM staining, which detected VLPs (or vesicles with multiple VLPs) labelled by Gag/Capsid antibodies either within embryonic cells or on the cell surface, similar to those seen in immuno-gold TEM staining of human EC cells (Fig. 3d, e and

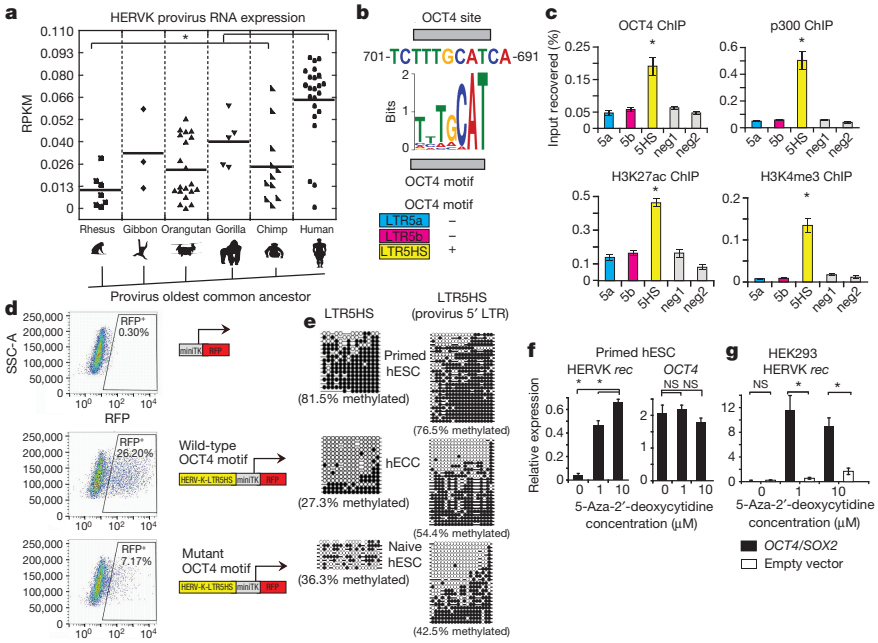


Figure 2 | Transactivation by OCT4 and DNA hypomethylation of LTR5HS synergistically regulate HERVK transcription. **a**, Expression of different HERVK proviral sequences, grouped according to the oldest common ancestor, as defined previously[4]. *P value < 0.05, non-paired Wilcoxon test. Solid line indicates mean. RNA-seq data set used for the analysis was from 3iL naive H1 cells[12]; $n = 3$ biological replicates. **b**, Conserved OCT4 site in LTR5HS with position weight matrix of the corresponding motif shown for comparison (top). Presence/absence of OCT4 motif in distinct LTR5 sequences is indicated (bottom); more detailed sequence information is in Extended Data Fig. 2a. **c**, ChIP-qPCR analyses from human EC cells (NCCIT) using antibodies indicated on top of each graph. Signals were quantified using primer sets specific to LTR5HS (5HS), LTR5a (5a) and LTR5b (5b) consensus sequences or two 'negative' intergenic, non-repetitive regions (neg1, neg2). *P value < 0.05 compared to negative control, one sided t-test. $n = 4$ biological replicates, error bars are ± 1 standard deviation (s.d.). **d**, Flow cytometry analysis of human EC cells with integrated LTR5HS fluorescent reporters, either wild type (middle) or with OCT4 motif mutation (bottom). Red fluorescent protein (RFP)-positive population was gated using side-scatter area (SSC-A) and cells with integrated negative control reporter (top) containing minimal thymidine kinase (miniTK) promoter. Shown is a representative result of two independent experiments. **e**, Bisulfite conversion quantification of LTR5HS 5-methyl-cytosine levels measured using LTR5HS-specific primer pairs anchored in the LTR5HS consensus sequence (left) or provirus-specific 5′ LTR5HS (right) for human EC cells (hECC; NCCIT), human ES cells (hESC; H9) or naive human ES cells (ELF1). Filled circles depict modified cytosines, open circles depict unmodified cytosines. Human EC cells (NCCIT) and naive human ES cells (ELF1) are less methylated than primed human ES cells (H9). $P < 0.05$, non-paired Wilcoxon test. **f**, qPCR with reverse transcription (RT–qPCR) analysis of human ES cells (H9) treated with indicated concentrations of 5-aza-2′-deoxycytidine for 24 hours. *P value < 0.05, one-sided t-test. $n = 3$ biological replicates, error bars ± 1 s.d. **g**, RT–qPCR analysis of HERVK *rec* RNA levels in HEK293 cells treated with indicated concentrations of 5-aza-2'-deoxycytidine, followed by transfection with *OCT4/SOX2* expression constructs. *P value < 0.05, one-sided t-test; NS, not significant. $n = 4$ biological replicates, error bars ± 1 s.d.

Extended Data Fig. 5b); control blastocyst staining showed no signal from secondary antibody (Extended Data Fig. 5c). Altogether, these data demonstrate that human preimplantation development proceeds in the presence of retroviral proteins and VLPs (summarized in Extended Data Fig. 5d).

Recent studies highlight the ability of TEs to contribute regulatory sequences to mammalian genomes[9,23,24]. For example, MERV-L elements in mouse have been reported to function as alternative promoters, driving expression of many two-cell stage-specific chimaeric transcripts[23]. However, we did not detect robust evidence for HERVK-associated chimaeric transcription (Extended Data Fig. 6a, b and Supplementary Table 1), suggesting that LTR5HS is unlikely to contribute promoter activity to nearby host genes. Alternatively, LTR sequences derived from ERVs could be co-opted to act as long-distance enhancers for the host[24]. In agreement with such a possibility, LTR5HS elements were marked by p300 and H3K27ac (Fig. 2c), while genes located in their vicinity showed a strong bias for naive-state-enriched expression, regardless of their upstream or downstream position in relation to the LTR5HS (Extended Data Fig. 6c–e). However, we cannot rule out that this result could be a consequence of preferential HERVK integration near genes active in the naive state.

HERVK encodes a small accessory protein, Rec, homologous to HIV Rev, which binds to and promotes nuclear export and translation of viral RNAs[25]. *rec*, an early viral transcript derived through alternative splicing of the *env* gene (Extended Data Fig. 2a), is expressed in naive ES cells and human blastocysts, and is rapidly induced in primed human ES cells exposed to 5-aza-2′-deoxycytidine (Extended Data Fig. 7a and Fig. 2f). We hypothesized that Rec-mediated nuclear export of viral RNAs into the cytoplasm might ultimately lead to the induction of innate antiviral responses, which typically rely on cytosolic detection of viral RNA/DNA and protein. We noted a striking induction of messenger RNA encoding an interferon-induced viral restriction factor IFITM1 (ref. 26; also known as FRAGILIS2) in human epiblast cells[10], as well as upregulation of *IFITM1* transcripts and surface protein levels in human naive versus primed human ES cells (Extended Data Fig. 7b, c, f and Supplementary Table 6). Furthermore, expression of a *rec* transgene in human EC cells was sufficient to elevate surface-localized IFITM1 protein levels (Fig. 4a). This was at least in part mediated through an effect on *IFITM1* mRNA transcription or stability, as *Rec* overexpression or knockdown had, respectively, increased or decreased *IFITM1* mRNA levels (Extended Data Fig. 7d). Of note, although the minimal components of the JAK/STAT interferon pathway are present in human EC cells, many other interferon-induced genes are not upregulated or expressed, indicating that HERVK triggers a precise antiviral response in host cells (Supplementary Table 2). To test whether HERVK expression provides viral resistance, we infected control wild-type human EC cells, control human EC cells expressing a green fluorescent protein (*GFP*) transgene, or two independent clonal Rec human EC cell lines (Rec-hECCs) with influenza H1N1(PR8) virus. Interestingly, the Rec-hECCs exhibited substantially attenuated infec-

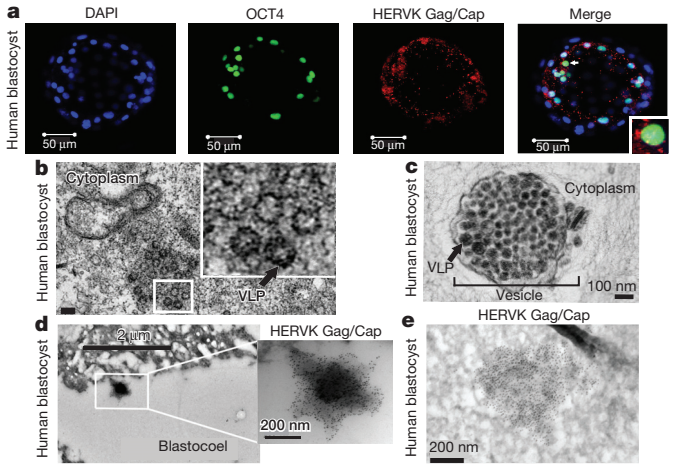


Figure 3 | Human blastocysts contain HERVK proteins and viral-like particles. **a**, Immunofluorescence of human blastocysts (days post-fertilization (DPF) 5–6) stained with 4′,6-diamidino-2-phenylindole (DAPI; blue), OCT4 antibody (green), and HERVK Gag/Capsid antibody (red). Images show a representative example ($n = 19$ embryos). Scale bar = 50 µm. White arrow indicates an OCT4$^+$ nucleus, surrounded by cytoplasmic Gag/Capsid (Cap), which is shown with higher magnification in an inset. **b**, Heavy metal staining TEM of a human blastocyst. Arrow indicates putative VLP (found in $n = 2/3$ blastocysts, DPF 5–6). Higher magnification of indicated region is shown in inset. Scale bar = 200 nm. **c**, Heavy metal staining TEM of human blastocyst. Arrow indicates putative immature VLP, bracket indicates vesicle filled with putative VLP (found in $n = 2/3$ blastocysts, DPF 5–6). Scale bar = 100 nm. **d**, **e**, Immuno-TEM of human blastocysts with Gag/Capsid staining; region of higher magnification is boxed. Representative examples of budding (**d**) and cell-internal (**e**) particles are shown; $n = 3$ blastocysts (DPF 5–6), $n = 3$ labelled particles in two embryos.

tion levels as compared to the control GFP-hECCs (Fig. 4b) or wild-type human EC cells (Extended Data Fig. 7e).

Retroviral accessory proteins often masterfully manipulate host cell factors to achieve optimal replicative efficiency. To examine whether, beyond reported binding to HERVK 3′ LTRs[25,27], Rec can also associate with cellular RNAs, we performed tandem affinity purification iCLIP-seq in human EC cells expressing Flag–GFP (GFP) or Flag–GFP-tagged *rec* transgene (Extended Data Fig. 8a, b). We did not detect associated RNA in the control Flag–eGFP purifications, indicating low nonspecific RNA recovery of our assay (Extended Data Fig. 8b). In contrast, parallel Rec purifications from two Flag–GFP Rec expressing clones yielded ultraviolet-crosslinked RNAs, sequencing of which demonstrated that *in vivo*, Rec robustly binds LTR5HS, but only in the region previously defined as containing the highly structured Rec-responsive element[25,28] (Fig. 4c and Extended Data Fig. 8b, c). In addition, Rec directly interacts with ~1,600 host mRNAs, preferentially in their 3′ untranslated regions (UTRs), a positional preference analogous to that observed in the viral RNA (Fig. 4d, e, Extended Data Fig. 9a and Supplementary Table 3). We did not detect specific RNA sequence motifs enriched at Rec-bound sites; however, multiple examined Rec iCLIP targets were predicted to fold into stable secondary structures (Extended Data Fig. 9b). This is reminiscent of the interaction of Rec with its HERVK LTR response element, which is mediated by RNA secondary structure, rather than a discrete specific binding site[28]. We also observed Rec association with mRNAs encoding surface receptor molecules and ligands (for example, *FGFR1*, *FGF13*, *FGFR3*, *KLRG2*, *IGF1R*, *FZD7*, *GDF3*) and chromatin regulators (for example, *DNMT1*, *CHD4*) (Extended Data Fig. 9a and Supplementary Table 3).

Given that Rec binding to viral RNAs promotes their nuclear export and translation, we next examined if endogenous mRNAs bound by Rec are also more efficiently targeted to ribosomes[22,25]. Ribosome profiling of Rec-hECCs, in comparison to wild-type human EC cells, revealed both increases and decreases in ribosomal occupancy, with differential enrichment of 941 mRNAs, of which 134 were also Rec iCLIP targets, representing a significant overlap (P value < 0.05, hypergeometric test) (Fig. 4f and Supplementary Table 5). Notably, mRNAs bound by Rec in 3′ UTRs or coding sequences were more likely to be upregulated in their ribosomal occupancy than expected by chance (P value < 0.05, hypergeometric test), but we did not observe such enrichment for mRNAs bound in their 5′ UTRs. We also noticed that several Rec-bound transcripts (for example, *RPL22*, *RPL31*, *RPS13*, *RPS20*, *EIF4G1*) encoding ribosome components and translation regulators had increased occupancy in Rec-hECCs,

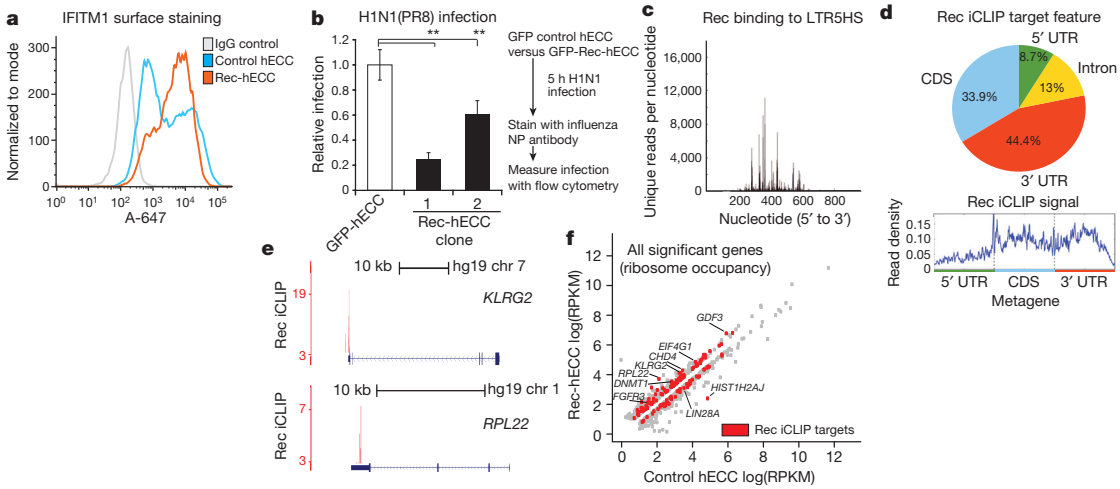


Figure 4 | HERVK accessory protein Rec upregulates viral restriction pathway and engages cellular mRNAs. **a**, Flow cytometry histograms of IFITM1 surface staining in control human EC cells or Rec-hECC cells; histogram of negative control cells stained with isotype IgG$^+$ Alexa-647 secondary antibody (A-647) is shown for comparison. Shown is a representative result of two independent experiments. **b**, H1N1(PR8) influenza infection of control GFP-hECC cells or two clonal lines of Rec-hECCs. Control cells were set as 100%, shown is aggregate results from two independent experiments, $n = 8$ total biological replicates for each condition. Error bars are ±1 s.d. **P value < 0.005, one-sided t-test. **c**, Rec iCLIP reads mapped to the LTR5HS sequence, $n = 2$ biological replicates. **d**, Distribution of Rec binding sites on endogenous mRNAs (top) and aggregate Rec iCLIP-seq signal on a metagene (bottom), $n = 2$ biological replicates. CDS, coding DNA sequence. **e**, Distribution of Rec iCLIP reads at representative target mRNAs *KLRG2* (top), *RPL22* (bottom); *y*-axis, iCLIP score, at cut-off = 3 (see Methods for details). **f**, Ribosome profiling signal for all significant genes (FDR < 0.05 Cuffdiff) in wild-type human EC cells versus Rec-hECCs, $n = 4$ biological replicates. Rec iCLIP targets are coloured in red.

potentially contributing to additional indirect translational effects of Rec overexpression (Fig. 4e, f and Supplementary Table 5).

Altogether, our results demonstrate that early human development is accompanied by the stage-specific transcriptional activation of HERVK, translation of its ORFs, and assembly of VLPs (Extended Data Fig. 10a). Beyond preimplantation development, we predict that HERVK reactivation occurs in human primordial germ cells (PGCs), which are also characterized by the presence of OCT4 and genome-wide DNA hypomethylation[29]. HERVK protein products have the potential to engage host machinery, as exemplified here by modulation of cellular mRNAs by Rec. This fine-tuning of cellular functions by HERVK proteins may contribute to human-specific or even individual-specific aspects of early development, as the retroviral ORFs are preferentially expressed from the human-specific proviruses, many of which are polymorphic in the human population[4,18]. Finally, our data raise the intriguing possibility that HERVK provides an immunoprotective effect for human embryos against different classes of viruses sensitive to the IFITM1-type restriction. Although *IFITM* family members were first described as interferon-induced genes, they are also classical naive-state and PGC markers in the mouse, which nonetheless appear to be dispensable for development[30]. These observations suggest that IFITM1-mediated restriction may be a evolutionarily conserved mechanism protecting both embryos and germ cells from either reinfection from infectious ERVs or exogenous viral infection (Extended Data Fig. 10a).

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 4 March 2014; accepted 10 February 2015.
Published online 20 April 2015.

1. Stoye, J. P. Studies of endogenous retroviruses reveal a continuing evolutionary saga. *Nature Rev. Microbiol.* **10,** 395–406 (2012).
2. Belshaw, R. *et al.* Long-term reinfection of the human genome by endogenous retroviruses. *Proc. Natl Acad. Sci. USA* **101,** 4894–4899 (2004).
3. Barbulescu, M. *et al.* Many human endogenous retrovirus K (HERVK) proviruses are unique to humans. *Curr. Biol.* **9,** 861–868 (1999).
4. Subramanian, R. P., Wildschutte, J. H., Russo, C. & Coffin, J. M. Identification, characterization, and comparative genomic distribution of the HERVK (HML-2) group of human endogenous retroviruses. *Retrovirology* **8,** 90 (2011).
5. Herbst, H., Sauter, M. & Mueller-Lantzsch, N. Expression of human endogenous retrovirus K elements in germ cell and trophoblastic tumors. *Am. J. Pathol.* **149,** 1727–1735 (1996).
6. Muster, T. *et al.* An endogenous retrovirus derived from human melanoma cells. *Cancer Res.* **63,** 8735–8741 (2003).
7. Contreras-Galindo, R. *et al.* Human endogenous retrovirus K (HML-2) elements in the plasma of people with lymphoma and breast cancer. *J. Virol.* **82,** 9329–9336 (2008).
8. Pace, J. K. & Feschotte, C. The evolutionary history of human DNA transposons: evidence for intense activity in the primate lineage. *Genome Res.* **17,** 422–432 (2007).
9. Kunarso, G. *et al.* Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nature Genet.* **42,** 631–634 (2010).
10. Yan, L. *et al.* Single-cell RNA-seq profiling of human preimplantation embryos and embryonic stem cells. *Nature Struct. Mol. Biol.* **20,** 1131–1139 (2013).
11. Smith, Z. D. *et al.* DNA methylation dynamics of the human preimplantation embryo. *Nature* **511,** 611–615 (2014).
12. Chan, Y.-S. *et al.* Induction of a human pluripotent state with distinct regulatory circuitry that resembles preimplantation epiblast. *Cell Stem Cell* **13,** 663–675 (2013).
13. Gafni, O. *et al.* Derivation of novel human ground state naive pluripotent stem cells. *Nature* **504,** 282–286 (2013).
14. Ware, C. B. *et al.* Derivation of naive human embryonic stem cells. *Proc. Natl Acad. Sci. USA* **111,** 4484–4489 (2014).
15. Takashima, Y. *et al.* Resetting transcription factor control circuitry toward ground-state pluripotency in human. *Cell* **158,** 1254–1269 (2014).
16. Theunissen, T. W. *et al.* Systematic identification of culture conditions for induction and maintenance of naive human pluripotency. *Cell Stem Cell* **15,** 471–487 (2014).
17. Hohn, O., Hanke, K. & Bannert, N. HERVK(HML-2), the best preserved family of HERVs: endogenization, expression, and implications in health and disease. *Front. Oncol.* **3,** 246 (2013).
18. Shin, W. *et al.* Human-specific HERVK insertion causes genomic variations in the human genome. *PLoS ONE* **8,** e60605 (2013).
19. Boller, K. *et al.* Evidence that HERVK is the endogenous retrovirus sequence that codes for the human teratocarcinoma-derived retrovirus HTDV. *Virology* **196,** 349–353 (1993).
20. Bieda, K., Hoffmann, A. & Boller, K. Phenotypic heterogeneity of human endogenous retrovirus particles produced by teratocarcinoma cell lines. *J. Gen. Virol.* **82,** 591–596 (2001).
21. Dewannieux, M. *et al.* Identification of an infectious progenitor for the multiple-copy HERVK human endogenous retroelements. *Genome Res.* **16,** 1548–1556 (2006).
22. Lee, Y. N. & Bieniasz, P. D. Reconstitution of an infectious human endogenous retrovirus. *PLoS Pathog.* **3,** e10 (2007).
23. Macfarlan, T. S. *et al.* Embryonic stem cell potency fluctuates with endogenous retrovirus activity. *Nature* **487,** 57–63 (2012).
24. Chuong, E. B., Rumi, M. A. K., Soares, M. J. & Baker, J. C. Endogenous retroviruses function as species-specific enhancer elements in the placenta. *Nature Genet.* **45,** 325–329 (2013).
25. Löwer, R., Tönjes, R. R., Korbmacher, C., Kurth, R. & Löwer, J. Identification of a Rev-related protein by analysis of spliced transcripts of the human endogenous retroviruses HTDV/HERVK. *J. Virol.* **69,** 141–149 (1995).
26. Brass, A. L. *et al.* The IFITM proteins mediate cellular resistance to influenza A H1N1 virus, West Nile virus, and dengue virus. *Cell* **139,** 1243–1254 (2009).
27. Hanke, K. *et al.* Staufen-1 interacts with the human endogenous retrovirus family HERVK(HML-2) Rec and Gag proteins and increases virion production. *J. Virol.* **87,** 11019–11030 (2013).
28. Magin-Lachmann, C. *et al.* Rec (formerly Corf) function requires interaction with a complex, folded RNA structure within its responsive element rather than binding to a discrete specific binding site. *J. Virol.* **75,** 10359–10371 (2001).
29. Gkountela, S. *et al.* The ontogeny of cKIT^{+} human primordial germ cells proves to be a resource for human germ line reprogramming, imprint erasure and *in vitro* differentiation. *Nature Cell Biol.* **15,** 113–122 (2013).
30. Lange, U. C. *et al.* Normal germ line establishment in mice carrying a deletion of the *Ifitm/Fragilis* gene family cluster. *Mol. Cell. Biol.* **28,** 4688–4696 (2008).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank P. Bieniasz for the HERVK-con plasmid, P. Lovelace for assistance with FACS, M. Teruel for recombinant *Giardia* DICER, J. Perrino for TEM assistance, T. Swigut for ideas and input on data analysis, B. Gu for assistance with bisulfite sequencing, A. Moore for assistance with influenza experiments, J. Skowronski and members of the Wysocka laboratory for invaluable comments on the manuscript. This work was supported by the National Institutes of Health (NIH) P01 GM099130 and R01 GM112720 (J.W.); Stanford Genome Sciences Training Program and National Science Foundation Graduate Research Fellowship Program (E.J.G.); NIH F30 (1F30CA189514-01) (R.A.F.); NIH DP2AI11219301 (C.A.B.); Smith Family Stanford Graduate Fellowship (N.L.B.); CIRM RB4-05763 and NIH P50-HG007735 (H.Y.C.); CIRM RB3-02209, March of Dimes 6-FY10-351 and U01 HL100397 (R.A.R.P.) grants and NIH equipment grants 1S10RR02933801 (Institute for Stem Cell Biology and Regenerative Medicine FACS core) and 1S10RR02678001 (electron microscopy core facility).

Author Contributions E.J.G. and J.W. conceived the project, designed experiments and wrote the manuscript, with input from all authors. E.J.G. carried out the majority of the experiments and data analyses. S.L.C., M.W. and E.J.G. performed human blastocyst handling and immunoflourescence with expertise and resources provided by R.A.R.P. R.A.F., L.M. and H.Y.C. performed and analysed iCLIP experiments. R.A.F. provided assistance with ribosome profiling experiments and analysis. N.L.B. and C.A.B. contributed influenza infection experiments. C.B.W. provided human naïve cells and reagents. D.J.W. performed expression analysis of LTR5HS-associated genes.

Author Information Sequencing data sets generated for this study are deposited under in the Gene Expression Omnibus under accession number GSE63570. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.W. (wysocka@stanford.edu).

METHODS

DNA and RNA isolation at reverse transcription. Genomic DNA was isolated using phenol:chloroform:isoamyl (100:100:1; PCI) (Invitrogen). Briefly, cells were digested in 10 mM Tris-HCl (pH = 8.0), 0.1 M EDTA, 0.5% SDS for 37 °C for 1 h, then proteinase K was added to final concentration of 100 μg ml^{-1} and then incubated for 3 h at 50 °C. DNA was PCI extracted, ethanol precipitated, and resuspended in TE. RNA was extracted using Trizol (Invitrogen) according to the manufacturer's instructions. DNase treatment with Turbo DNase (Ambion) was performed for 30 min at 37 °C, PCI extracted, ethanol precipitated, and re-suspended in water. Reverse transcription was performed with SuperScript III (Invitrogen) using ~500 ng of DNAase treated total RNA following the manufacturer's instructions. No reverse transcriptase controls were performed where necessary.

Cell lines and culture. NCCIT and HEK293 cells were obtained from ATCC. NCCIT cells were maintained in 10% FBS (Omega), 1× Glutamax-I supplement (100× stock, Invitrogen), 1× non-essential amino acids (100× stock, Invitrogen), and basal media RPMI 1640 (Hyclone). HEK293 cells were maintained in 10% FBS, 1× NEAA, 1× glutamax in DMEM-high glucose (Hyclone). Human ES cells (H9 line female, Wi-Cell) were used at passage 60–67 and were expanded in feeder-free, serum-free medium (mTeSR-1) from StemCell technologies. HESC HSF-1 (male) and HSF-8 (male) human ES cells were used at passage 20–28, cultured as described earlier and their characterization is described elsewhere[31]. Cells were passaged 1:7 every 5–6 days by incubation with accutase (Invitrogen) and the resultant small cell clusters (50–200 cells) were subsequently re-plated on tissue culture dishes coated overnight with growth-factor-reduced matrigel (BD Biosciences). ELF1 naive human ES cells were obtained from C.W. and cultured as previously described[14], with 10 ng ml^{-1} human recombinant LIF (R&D). Cell cultures were routinely tested and found negative for mycoplasma infection (MycoAlert, Lonza).

ChIP. ChIP assays were performed from approximately 10^7 cells per experiment, according to previously described protocol with slight modifications[32,33]. Briefly, cells were crosslinked with 1% formaldehyde for 10 min at room temperature and formaldehyde was quenched by addition of glycine to a final concentration of 0.125 M. Chromatin was sonicated to an average size of 0.5–2 kb, using Bioruptor (Diagenode). 50–75 μl of protein G dynal beads (Invitrogen) were used to capture 3–5 μg of antibody in phosphate citrate buffer pH 5.0 (2.4 mM citric acid, 5.16 mM Na_2HPO_4) for 30 min at 27 °C. Antibody bead complexes were rinsed two times with PBS and added to sonicated chromatin and rotated at 4 °C overnight. Ten per cent of chromatin was reserved as 'input' DNA. Magnetic beads were washed and chromatin eluted, followed by reversal of the crosslinkings and DNA purification. Resultant ChIP DNA was dissolved in TE.

Flow cytometry. Cells were trypsinized and analysed on a CS&T-calibrated BD FACS Aria II SORP flow cytometer on a 561 nm laser line for turboRFP, with 582/15BP. For IFITM1 flow cytometry, cells were allowed to recover after trypsinization for 2 h at 37 °C in media. Then, 2.5 × 10^5 cells were washed with PBS/10% FBS/0.1% sodium azide and stained with 1:100 IFITM1 antibody (rabbit pAb, ProteinTech, #50556193) for 30 min at 4 °C. Washed cells were then incubated with chick, anti-mouse A647 secondary for 30 min at room temperature. Control stainings using rabbit IgG (santa cruz) and anti-mouse A647 were also performed.

Bisulfite sequencing. EpiTect Plus Bisulphite conversion kit (Qiagen) was used to bisulfite convert 1 μg genomic DNA as per manufacturer's instructions. Approximately 20 ng of BS-treated DNA was used as a template for 35–40 cycles with Platinum taq (Invitrogen, 10966) as per manufacturer's instructions. A-tailed PCR fragments were gel purified and inserted into pGEM-T. 5′ LTR provirus= specific BS-PCR was conducted with primers including NcoI and NotI sites to facilitate cloning into pGEM-T. Approximately 15 clones were Sanger sequenced for both forward and reverse strands. BiQ software was used to align and quantify CpG methylation.

Protein extraction and immunoblotting. Proteins were extracted using previously described protocols[33]. Briefly, cells were resuspended in buffer A (10 mM HEPES, pH 7.9, 10 mM KCl, 1.5 mM MgCl, 0.34 M sucrose, 10% glycerol) and fresh protease inhibitors (Complete EDTA-free, Roche), 1 mM PMSF and 0.1% Triton-X 100 were added. Cytoplasmic extract was further clarified by centrifugation at 13,000 r.p.m. at 4 °C for 10 min, and total protein concentration was assayed with Bradford reagent (Biorad). Equal amounts of protein were run on SDS–PAGE gels and then transferred onto Hybond ECL membranes (Amersham). Membranes were blocked using 5% milk, PBS, 0.1% Tween-20 for 1 h at 27 °C. Primary antibodies (see Supplementary Table 10) were used in blocking solution overnight at 4 °C. Horseradish peroxidase (HRP)-conjugated secondary antibodies were used and chemoluminescence was assayed using Lumi-light plus (Roche).

qPCR. All primers used in qPCR analyses are shown in Supplementary Table 10. qPCR was performed using SensiFAST SYBR No-Rox Kit (Bioline) in a Light Cycler 480II machine (Roche), using technical triplicates. ChIP-qPCR signals were calculated as percentage of input and unless, indicated, qRT–PCR signal was normalized to 18S rRNA. Standard deviations were measured from the averages of the technical repeats for each biological replicates and represented as error bars ±1 s.d.

Plasmid and constructs. HERVK LTR5_HS sequence from HERVK-con[22] was cloned upstream of a miniTK promoter driving turboRFP and inserted into piggy-back transposon (SystemBio). Motif mutations for OCT4 or SOX2 were produced by replacing the respective motif with a NotI site. 2.5 μg of reporter vector along with 0.5 μg of piggy-back transposase were transfected into cells using 18 μl lipofectamine2000 (Invitrogen) in 6-well plates. 400 μg ml^{-1} G418 (Amresco) was used to select for integrants. Cells were analysed >10 days later to minimize signal from non-integrated reporter expression. Complementary DNAs encoding *OCT4* or *SOX2* were cloned into pcDNA containing carboxy-terminal or amino-terminal Flag–haemagglutinin (HA) tags, respectively. The same LTR regulatory regions were cloned into pGL3 firefly luciferase reporters, and constructs were co-transfected with *Renilla* luciferase for perform dual luciferase assays. SV40 promoter/enhancer firefly luciferase was used a positive control. Transgene constructs for Rec expression in NCCIT cells were used with *eif1a* promoter, N-terminal Flag–eGFP-tagged Rec cloned into a piggy-back construct with a puromycin selectable marker. Control construct using Flag–eGFP alone (vector only control) was also used in parallel. Transgene constructs were cotransfected with piggy-back transposase plasmid to generate stable lines. Clones were selected and expanded. Flag–eGFP–Rec cone #1 has ~30× endogenous expression of *rec* mRNA (as measured by qPCR) and Flag–eGFP–Rec clone #2 has ~14× endogenous expression of *rec* mRNA (qPCR), data not shown.

siRNA knockdown. siRNA was generated using baculovirus-produced *Giardia* DICER as described[34]. Briefly, 1 μg of PCR product was *in vitro* transcribed using Megascript T7 (Ambion) and digested using DICER at 37 °C for 16 h. siRNA was purified using Purelink RNA mini Kit (Ambion) and the absence of >22 nucleotides RNA was verified using gel electrophoresis and ethidium bromide staining. NCCIT cells were plated onto Matrigel-coated 24-well plates, transfected using 1.5 μl of RNAi-max (Invitrogen) in optimem (Gibco) with 25 nM siRNA concentrations for 4 h before addition of fresh media. siRNA knockdowns were performed for three consecutive days, cells were harvested 24 h after final transfection. Two independent siRNA pools were generated for *OCT4*, *NANOG* and *SOX2*, one each for *turboRFP* (non-targeting control) and *rec*, which overlaps the *env* ORF. Primers used to generate double-stranded RNA (dsRNA) templates are listed in Supplementary Table 10.

Human embryo source and procurement. Human embryos were obtained as previously described[35]. Approximately 25 supernumerary human blastocysts from successful IVF cycles, subsequently donated for non-stem-cell research, were obtained with written informed consent from the Stanford University RENEW Biobank. De-identification was performed according to the Stanford University Institutional Review Board-approved protocol #10466 entitled 'The RENEW Biobank' and the molecular analysis of the embryos was in compliance with institutional regulations. Approximately 25% of the embryos were from couples that used donor gametes and the most common cause of infertility was unexplained at 35% of couples. No protected health information was associated with any of the embryos.

Human embryo thawing and culture. Human embryos cryopreserved at the blastocyst stage were thawed by a two-step rapid thawing protocol using Quinn's Advantage Thaw Kit (CooperSurgical) as previously described[35,36]. In brief, either cryostraws or vials were removed from the liquid nitrogen and exposed to air before incubating in a 37 °C water bath. Once thawed, embryos were transferred to a 0.5 mol l^{-1} sucrose solution for 10 min followed by a 0.2 mol l^{-1} sucrose solution for an additional 10 min. The embryos were then washed in Quinn's advantage medium with HEPES (CooperSurgical) plus 5% serum protein substitute (CooperSurgical) and each transferred to a 25 μl microdrop of Quinn's advantage blastocyst medium (CooperSurgical) supplemented with 10% serum protein substitute under mineral oil (Sigma). The embryos were cultured at 37 °C with 6% CO_2, 5% O_2 and 89% N_2 under standard human embryo culture conditions in accordance with current clinical IVF practice. Embryos used in this study were DPF 5–6.

Immunofluorescence. Cells were grown on Matrigel-coated glass coverslips, fixed using EM-grade 4% PFA (Electron Microscopy Sciences) for 15 min at 27 °C, washed three times with PBS, blocked and permeablized with 1% BSA, 0.3% Triton-X 100 in PBS (antibody buffer) supplemented with 5% serum for species-matched secondary antibody for 1 h at 27 °C. Primary antibodies were resuspended in antibody buffer and incubated at 4 °C overnight. Washes were performed three times using 0.1% Triton-X 100 in PBS, and secondary antibodies were added for 1 h at 27 °C in the dark. Cells were mounted using Prolong-fade gold (Invitrogen) with DAPI and imaged on Zeiss LSM 700 confocal.

For embryo immunostaining, the zona pellucida (ZP) was removed from each embryo by treatment with acidified Tyrode's solution (Millipore) and ZP-free embryos were washed in PBS plus 0.1% BSA and 0.1% Tween-20 (PBS-T; Sigma-Alrdrich) before fixation in 4% paraformaldehyde for 20 min at room temperature. Once fixed, the embryos were washed three times in PBS-T to remove any residual fixative and permeabilized in 1% Triton X-100 (Sigma-Aldrich) for 1 h at room temperature. Following permeabilization, the embryos were washed three times in PBS-T and then blocked in 4% of chicken or goat serum in PBS-T overnight at 4 °C. The embryos were incubated with primary antibodies in PBS-T with 1% serum sequentially for 1 h each at room temperature at the following dilutions: 1:200 OCT4, 1:100 Gag/Capsid. Primary signals were detected using the appropriate 488- or 647-conjugated Alexa Fluor secondary antibody (Invitrogen) at a 1:250 dilution at room temperature for 1 h in the dark and subsequently DAPI stained. Immunofluorescence was visualized by sequential imaging, whereby the channel track was switched each frame to avoid cross-contamination between channels, using a Zeiss LSM510 Meta inverted laser scanning confocal microscope. The instrument settings, including the laser power, pinhole and gain, were kept constant for each channel to facilitate semi-quantitative comparisons between embryos.

DNA demethylation treatment. HEK293 cells were plated on Matrigel-coated 24-well plates, and treated with 0, 1 or 10 μM 5-aza-2′-deoxyctidine (Calbiochem) freshly prepared every 24 h. Cells were then transfected with 1 μg each of pcDNA3.1-OCT4 and pcDNA3.1-SOX2 expression plasmids. Media was changed 24 h later, and cells were harvested 3 days after transfection for RNA analysis. Human ES cells (H9) were grown as described earlier, except mTeSR was supplmented with Rock inhibitor (y-27632, Sigma) at 5 μM, and treated with 0, 1 or 10 μM 5-aza-2′-deoxyctidine (Calbiochem) for 24 h.

RNA-seq library construction. Libraries were constructed as described[33], using ~10 μg of total RNA followed by poly-A selection with oligo-dT beads, ligation and ten cycles of PCR with NEBnext kit oligonucleotides, and sequenced using Illumina Hi-Seq2000 at the Stanford Sequencing Facility or ELIM Bio.

Sequence analysis. For RNA-seq repeat analysis of data from embryo and human ES cell libraries (for Fig. 1 and Extended Data Fig. 1), FASTQ files were aligned to repbase consensus sequences (downloaded from RepBase) with bowtie using the command "bowtie -q -p 8 -S -n 2 -e 70 -l 28-maxbts 800 -k 1 -best". These bowtie parameters ensure that only the best alignment (highest scores) is reported, furthermore only one alignment per read is reported, that is, these settings do not allow multiple-matching. For Fig. 2a analysis of HERVK proviruses, RNA-seq reads were aligned to hg19 using the same parameters described earlier, and the overlap between the manually curated HERVK provirus data set[5] is reported. For RefSeq analysis for RNA-seq libraries generated for this paper (ELF1 naive or primed human ES cells; from human EC cell siRNA RNA-seq, or Rec-hECC versus wild-type human EC cell experiments), reads were processed using DNAnexus software to obtain read counts and RPKM. Reads were counted and where indicated normalized to repeat length and library size using RPKM. Differential expression in RNA-seq experiments described earlier was performed using DESeq, with reported FDR using Benjamini–Hochberg correction.

Interferon-induced gene set analysis. Genes were defined as interferon induced if they were induced fivefold in interferon-treated cells/tissues for experimentally deposited data sets found in Interferome database[37] (http://interferome.its.monash.edu.au/interferome/home.jspx).

LTR5HS-associated gene analysis. RefSeq genes were classified as associated or not associated with LTR5HS (downloaded from UCSC genome browser table) using Great Analysis Software (Bejerano laboratory, Stanford University) with a cut-off of 100 kb distance from the TSS. These classified RefSeq genes were then compared using the RPKM and DESeq analysis as described earlier. Differential enrichment of LTR5HS-associated transcripts in naive/primed upregulated versus naive/primed downregulated was analysed using non-paired Wilcoxon test, and significance is reported at P value < 0.05. Higher average naive/primed RPKM of LTR5HS-associated versus non-LTR5HS-associated genes was tested using non-paired Wilcoxon test.

Chimaeric transcript identification. One-hundred base-pair paired-end RNA-seq reads generated with ELF1 naive versus primed human ES cells (see earlier) were analysed using a published pipeline[22]. Briefly, Cufflinks software was used to perform *de novo* identification of transcript models. These transcript models were then used to identify splice junctions in which one side of the transcript model overlapped the GTF file (for hg19 from UCSC), cataloguing known genes and long noncoding RNAs (lincRNAs), and the other side of the transcript model aligned to hg19 classified as a repeat (UCSC genome browser, repeat track). Transcripts that fulfilled these criteria were classified as chimaeric transcripts, and are reported in Supplementary Table 1.

Clustering. Hierarchical clustering was performed using Gene-e software (http://www.broadinstitute.org/cancer/software/GENE-E/index.html) using K-means clustering of $\log_2$-transformed RPKM.

Statistical tests. A list of the statistical tests, multiple-hypothesis testing corrections, and normality criteria for parametric tests is reported in Supplementary Table 7.

Electron microscopy. Samples were fixed using 4% PFA and 0.01% glutaraldehyde for 15 min at 27 °C. Routine heavy metal staining was conducted where indicated. Immuno-TEM with 1:100 dilution of anti-HERVK Gag/Capsid using overnight incubation at 4 °C and labelling was visualized using 5 nm gold-labelled anti-mouse secondary antibody. Secondary only controls demonstrated specificity of the antibody for this application. TEM was performed at the Electron Microscopy core at Stanford University using a Jeol JEM-1400 electron microscope.

iCLIP and data analysis. The iCLIP method was performed as described before with the specific modifications below[38]. Flag–GFP–Rec (FG-Rec)-expressing NCC cells were UV-C crosslinked to a total of 0.3 J cm^{-2}. Each iCLIP experiment was normalized for total protein amount, typically 1 mg, and partially digested with RNaseI (Life Technologies) for 10 min at 37 °C and quenched on ice. FG-Rec was isolated with anti-Flag agarose beads (Sigma) for 3 h at 4 °C on rotation. Samples were washed sequentially in 1 ml for 5 min each at 4 °C: 2× high stringency buffer (15 mM Tris-HCl pH 7.5, 5 mM EDTA, 2.5 mM EGTA, 1% Triton X-100, 1% Na-deoxycholate, 120 mM NaCl, 25 mM KCl), 1× high salt buffer (15 mM Tris-HCl pH 7.5, 5 mM EDTA, 2.5 mM EGTA, 1% Triton X-100, 1% Na-deoxycholate, 1 M NaCl), 1× NT2 buffer (50 mM Tris-HCl pH 7.5, 150 mM NaCl, 1 mM $MgCl_2$, 0.05% NP-40). Purified FG-Rec was then eluted from anti-Flag agarose beads using competitive Flag peptide elution. Each sample was resuspended in 500 μl of Flag elution buffer (50 mM Tris-HCl pH 7.5, 250 mM NaCl, 0.5% NP-40, 0.1% Na-deoxycholate, 0.5 mg ml^{-1} Flag peptide) and rotated at 4 °C for 30 min. The Flag elution was repeated once for a total of 1 ml elution. FG-Rec was then captured using anti-GFP antibody (Life Technologies, A-11122) conjugated to Protein A dynabeads (Life Technologies) for 3 h at 4 °C on rotation. Samples were then washed as described previously in the anti-Flag agarose beads. 3′-End RNA dephosphorylation, 3′-end single-stranded RNA (ssRNA) ligation, 5′ labelling, SDS–PAGE separation and transfer, autoradiograph, RNP isolation, Proteinase K treatment and overnight RNA precipitation took place as previously described[38]. The 3′-ssRNA ligation adaptor was modified to contain a 3′ biotin moiety as a blocking agent. The iCLIP library preparation was performed as described elsewhere[38,39]. Final library material was quantified on the BioAnalyzer High Sensitivity DNA chip (Agilent) and then sent for deep sequencing on the Illumina HiSeq 2500 machine for 1× 75 bp cycle run. iCLIP data analysis was performed as previously described[39]. For analysis of repetitive noncoding RNAs, custom annotation files were built from the Rfam database. For analysis of endogenous retroviral elements, custom annotation files were built from the repbase database. iCLIP reads were filtered for quality, barcode split, PCR-duplicate removed, trimmed (5′ and 3′ends), and mapped for unique matches under parameters previously described[38,39]. Bioinformatic pipeline used for iCLIP data analysis is described previously[39]. Briefly, RT stops were used to map nucleotide resolution of Rec binding, and only nucleotides supported with three independent RT stops in two replicates (with at least one RT stop in each replicate) were reported as binding events, and are reported in Supplementary Table 3.

Ribosome profiling. Human EC cells (NCCIT) were cultured as described earlier. Total RNA was extracted using Trizol (Life Technologies) and used as input material for the ARTseq Ribosome Profiling Kit—Mammalian (Epicentre) following the manufacturer's protocol with the following modifications. The 3′ RNA ligation adaptor and cDNA synthesis primers from the iCLIP protocol were for library construction. Final library material was quantified as in the iCLIP experiments and sequenced on the Illumina HiSeq 2500 machine for 1× 75 bp cycle run. Sequencing reads were preprocessed (quality filter, PCR duplicate removal, and trimming) as in the iCLIP protocol. Mapping was performed using an established pipeline previously described[40]. Briefly, reads were aligned to 45 s rDNA repeat sequence with bowtie to remove residual rRNA reads from libraries. Non-aligning reads (non-rRNA) were then aligned to hg19 with TopHat2 and differential expression was identified using default parameters for CuffDiff/Cufflinks software with significance at FDR < 0.05.

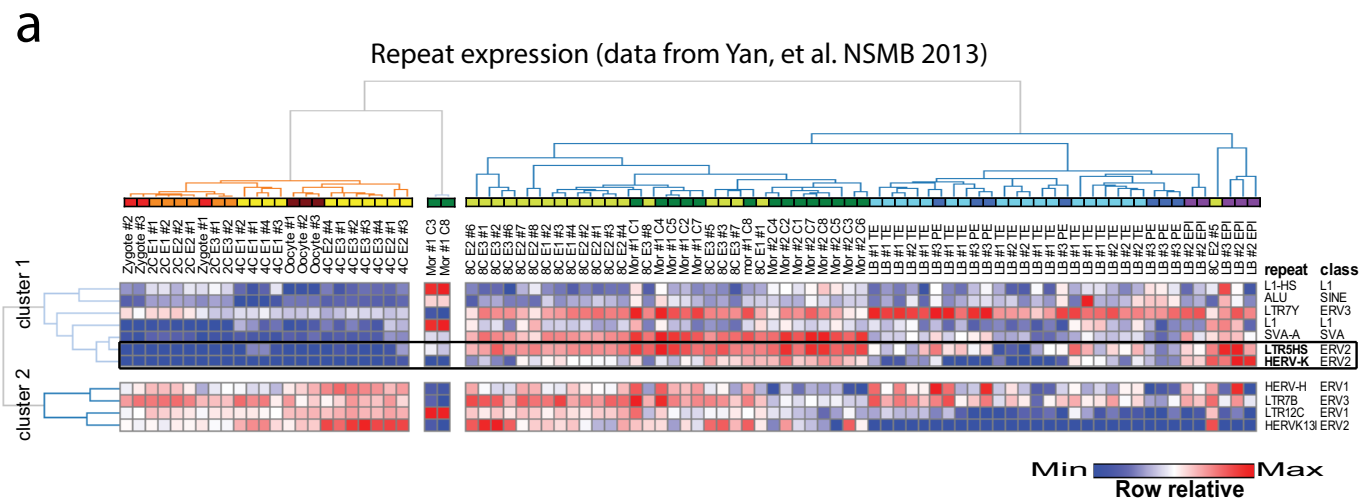
Influenza infection experiments. Human EC cells (NCCIT) were plated in duplicate (1.5×10^5 cells per well) on a 96-well flat-bottom plate in 100 μl Virus Diluent (DMEM, Gibco, supplemented with 1% BSA, 1× antibiotics and 20 mM HEPES). Cells were incubated at 37 °C and 5% CO_2 for 1.5 h. Wild-type human EC cells and REC-hECCs were then infected with virus (influenza A/H1N1/PR8/1934, diluted 1:10 into 100 μl virus diluent, increasing total volume to 200 μl). Cells were incubated at 37 °C for 1 h. FBS (Hyclone) was added to the wells to a final concentration of 10% FBS. Cells were incubated at 37 °C for 5 h. 20 mM EDTA (20 μl) was added to all wells and mixed thoroughly to stop infection. Cells were washed with 200 μl 1× PBS (Hyclone), re-suspended in

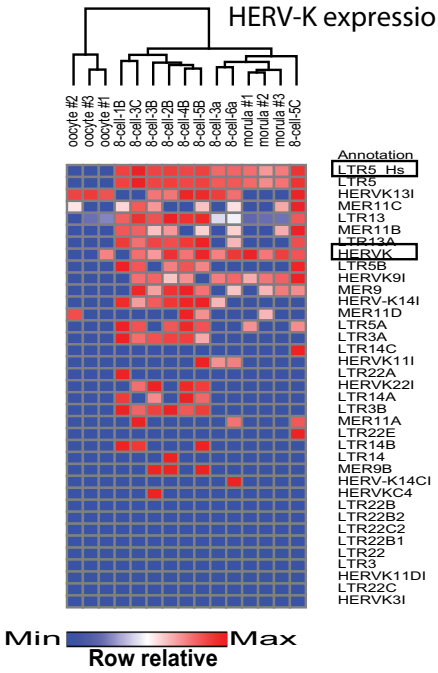
100 µl 1× BD FACS Lysing Solution (BD Biosciences) and stored at −80 °C for later processing.

For staining and analysis, cells were thawed in 37 °C for 20 min. One-hundred microlitres FACS wash (1× HyClone DPBS with 2% FBS) was added to each well and plate was centrifuged. Cell pellets were re-suspended in 200 µl BD FACS Permeabilizing Solution II (BD Biosciences). Cells were incubated at room temperature in the dark for 10 min. Plate was centrifuged and cells were washed twice with 200 µl FACS wash. Cells were stained with primary antibody (mouse anti-influenza A nucleoprotein, C43 clone, Abcam) diluted to 2 µg ml^{-1}. Cells were incubated in the dark at room temperature for 30 min and washed twice. Cell pellets were resuspended in 2 µg ml^{-1} of secondary antibody (chicken anti-mouse Alexa647, Invitrogen) in 50 µl FACS wash and incubated in the dark at room temperature for 30 min. Cells were washed twice and cell pellets were resuspended in 1% PFA (Electron Microscopy Sciences). Cells were analysed on the MACSQuant Analyzer (Miltenyi Biotec). MACSQuant Calibration Beads (Miltenyi Biotec) were used for calibration of the cytometer. Compensation controls were run using 1:1 mixture of CompBead Plus Anti-mouse Igκ (BD) and negative control beads. Single stained cellular controls were run in parallel to infected and uninfected samples. Data were analysed by FlowJo 9.7.6 (TreeStar). Cells were gated to exclude dead cells and debris. Infection levels were background subtracted using uninfected wells, and normalized to infection levels in GFP-hECC for each run.

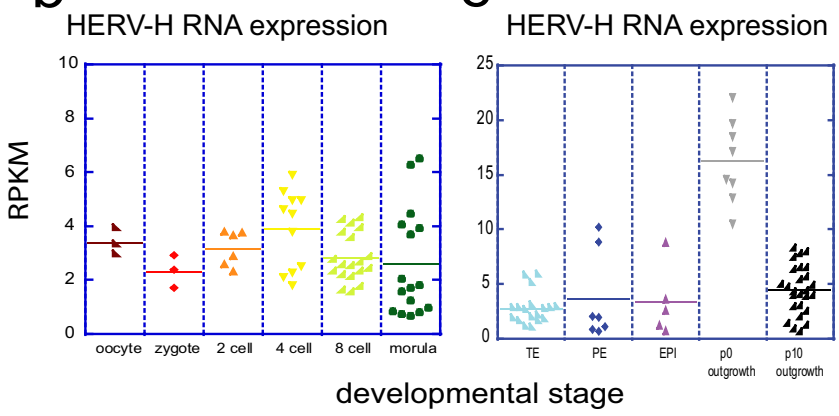
RNA-seq data sets. Data sets used in this study can be accessed from: Array Express Database (accession number E-MATB-2031)[12]; Gene Expression Omnibus (accession number GSE36552)[10]; Gene Expression Omnibus (accession number GSE44183)[41]; Array Express (accession number E-MTAB-2857)[15]. Sequencing data sets generated for this study are deposited under in the Gene Expression Omnibus under accession number GSE63570, and are summarized in Supplementary Table 8.

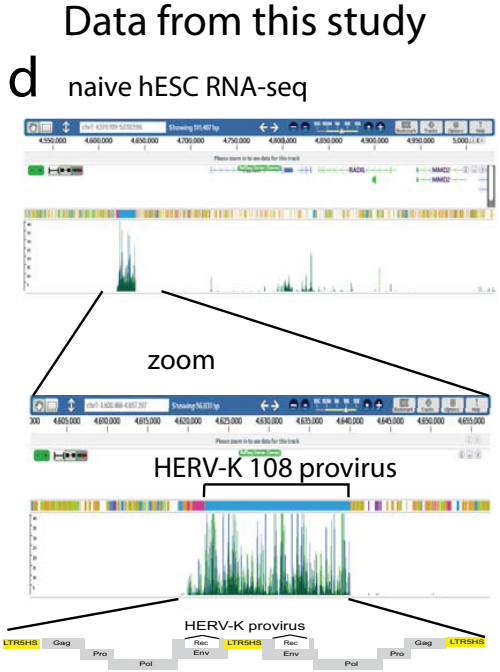
31. Chavez, S. L., Meneses, J. J., Nguyen, H. N., Kim, S. K. & Pera, R. A. R. Characterization of six new human embryonic stem cell lines (HSF7, -8, -9, -10, -12, and -13) derived under minimal-animal component conditions. *Stem Cells Dev.* **17,** 535–546 (2008).
32. Boyer, L. A. *et al.* Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* **122,** 947–956 (2005).
33. Peng, J. C. *et al.* Jarid2/Jumonji coordinates control of PRC2 enzymatic activity and target gene occupancy in pluripotent cells. *Cell* **139,** 1290–1302 (2009).
34. Myers, J. W. & Ferrell, J. E. Jr. in *RNA Silencing* (ed. Carmichael, G. G.) 93–196 (Humana Press, 2005).
35. Chavez, S. L. *et al.* Dynamic blastomere behaviour reflects human embryo ploidy by the four-cell stage. *Nature Commun.* **3,** 1251 (2012).
36. Wong, C., Chen, A. A., Behr, B. & Shen, S. Time-lapse microscopy and image analysis in basic and clinical embryo development research. *Reprod. Biomed. Online* **26,** 120–129 (2013).
37. Rusinova, I. *et al.* INTERFEROME v2. 0: an updated database of annotated interferon-regulated genes. *Nucleic Acids Res.* **41,** D1040–D1046 (2013).
38. Huppertz, I. *et al.* iCLIP: protein-RNA interactions at nucleotide resolution. *Methods* **65,** 274–287 (2014).
39. Flynn, R. A. *et al.* Dissecting noncoding and pathogen RNA–protein interactomes. *RNA* **21,** 135–143 (2015).
40. Ingolia, N. T., Lareau, L. F. & Weissman, J. S. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* **147,** 789–802 (2011).
41. Xue, Z. *et al.* Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. *Nature* **500,** 593–597 (2013).



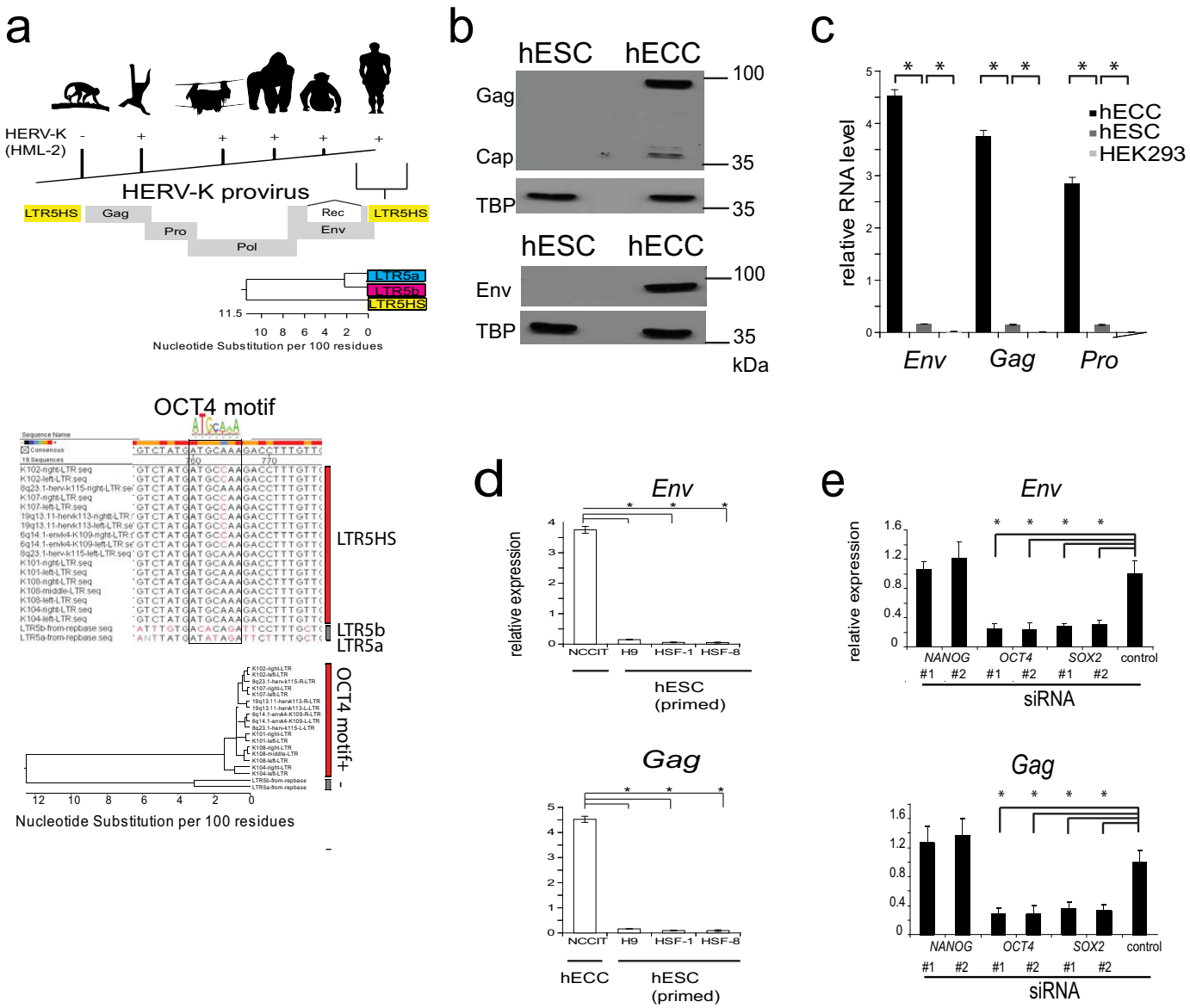


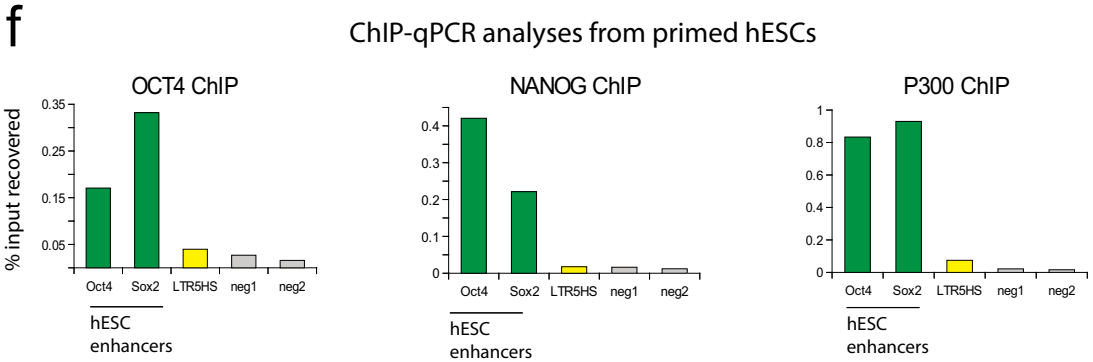
Data from Yan, et al. Cell Stem Cell 2013.



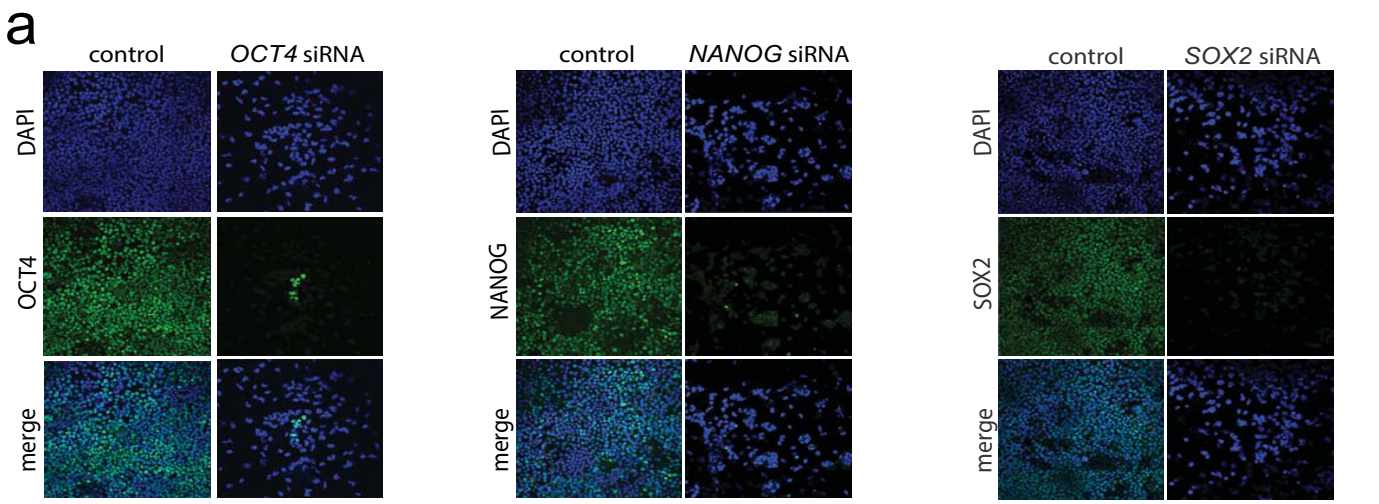


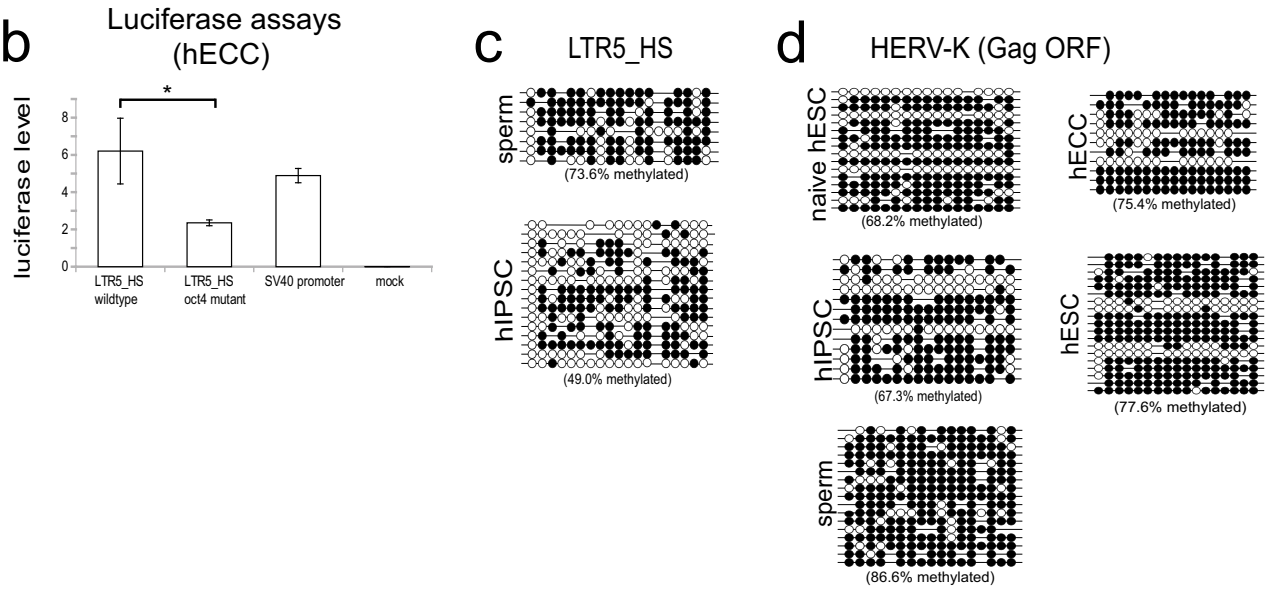
Extended Data Figure 1 | Additional single-cell RNA-seq data analyses from preimplantation human embryos (supporting Fig. 1). a, Heat map and hierarchical *K*-means clustering of highly expressed (average RPKM > 6 across 89 embryo libraries) repetitive elements in single cells of human preimplantation embryos at indicated developmental stages (top) and HERVK expression (bottom) using indicated data sets. **b**, HERVH expression (RPKM) in single cells of human embryos at indicated preimplantation stages. Solid line indicates mean. RNA-seq data are taken from ref. 10. **c**, HERVH expression (RPKM) in single cells of human blastocysts, grouped by lineage. Solid line indicates mean. Oocyte ($n = 3$), zygote ($n = 3$), 2-cell ($n = 6$), 4-cell ($n = 11$), 8-cell ($n = 19$), morula ($n = 16$), TE ($n = 18$), PE ($n = 7$), EPI ($n = 5$), p0 ($n = 8$), p10 ($n = 26$). RNA-seq data set was from ref. 10. **d**, Genome browser snapshot showing 100 bp PE-RNA-seq reads from ELF1 naive human EScells aligning at the HERVK 108 provirus on chromsome 7.

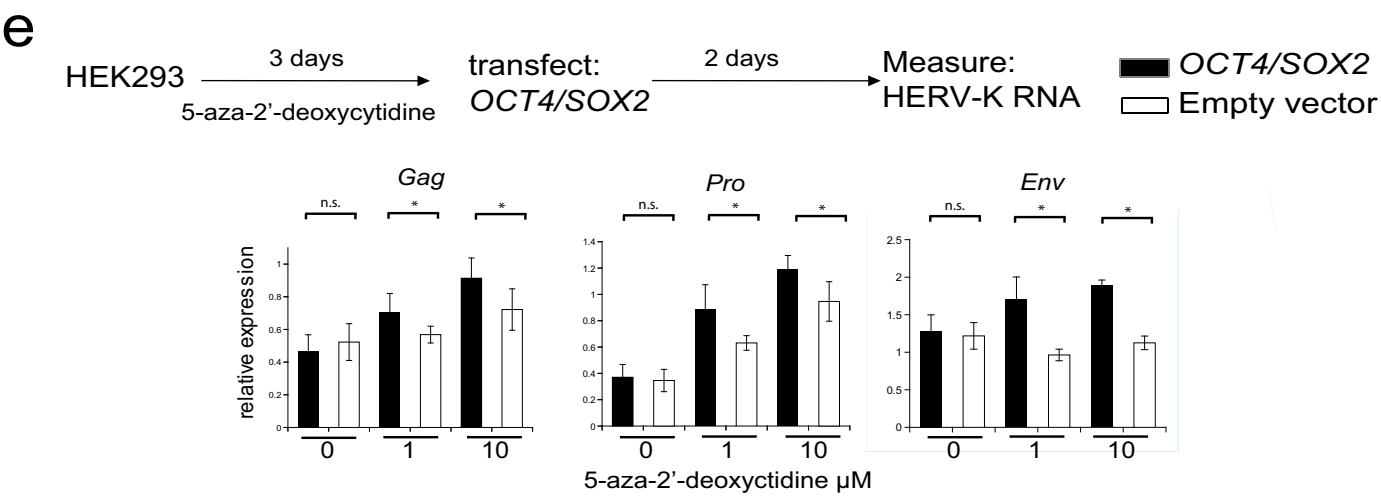




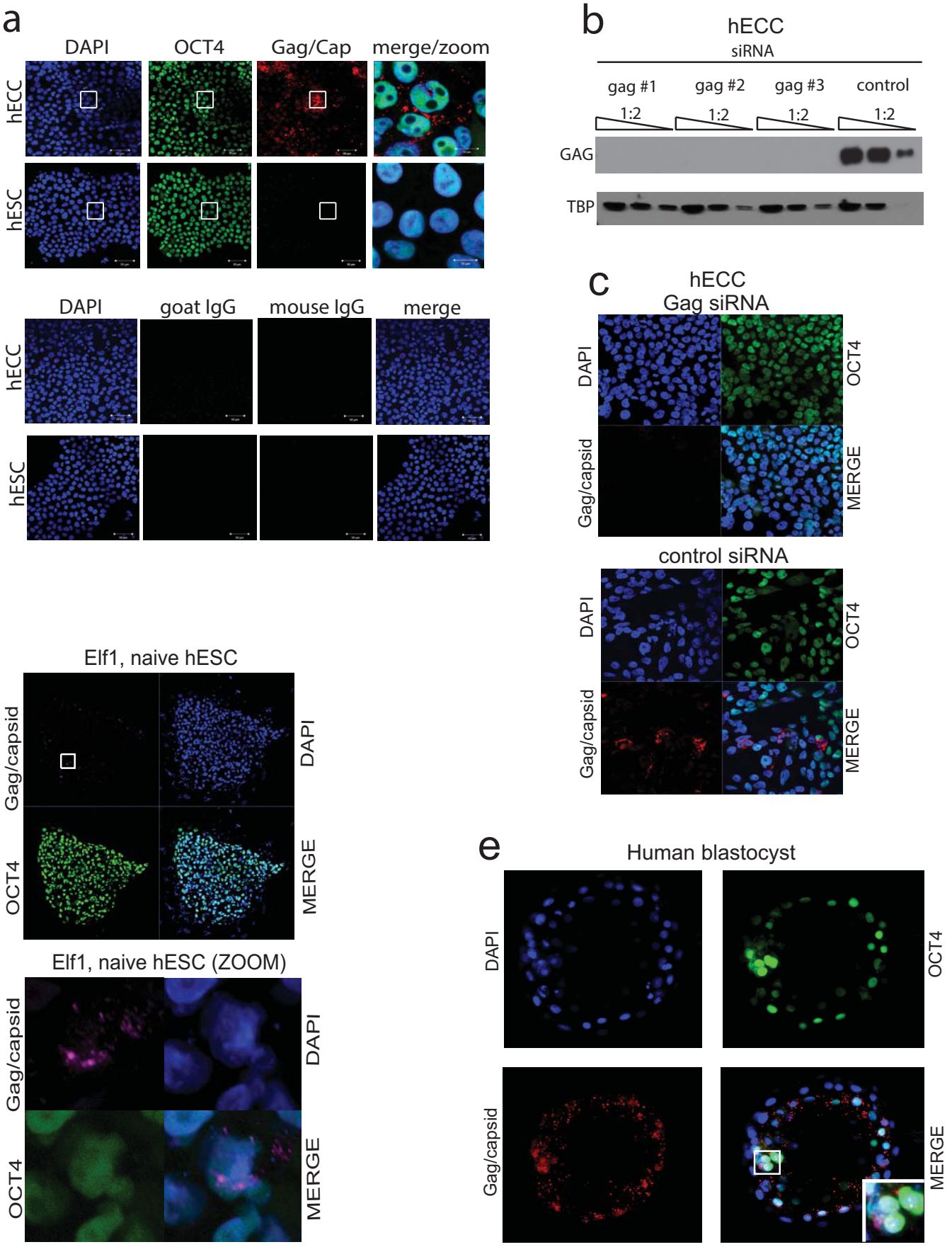
Extended Data Figure 2 | LTR5 alignments, HERVK expression data in cell lines, and control ChIP-qPCR analyses in primed human ES cells (supporting Fig. 2). a, Top, presence of HERVK(HML-2) sequences in Old World primates, but absence in New World primates. Middle, schematic of HERVK proviral genome; all human-specific insertions contain LTR5HS. Bottom, phylogenetic relationship of HERVK LTR subclasses showing high degree of sequence similarity. Pro, protease; Pol, polymerase; Gag, group-specific antigen; Env, envelope. Bottom, ClustLW multiple sequence alignment of indicated HERVK LTR sequences (top), region around OCT4 motif is boxed, phylogenetic tree (bottom) indicating presence/absence of OCT4 motif. **b**, HERVK protein expression in human EC cells and human ES cells. Protein extracts from human EC cells (NCCIT) and human ES cells (H9) were analysed by immunoblotting with an antibody detecting HERVK Gag precursor and the processed Capsid (top), or the glycosylated, unprocessed form of the HERVK envelope protein Env (bottom). Tata-binding protein (TBP) was used as a loading control. Shown is a representative result of three independent experiments. **c**, RT–qPCR analysis of HERVK RNA expression in human EC cell line NCCIT, human ES cell line H9, and HEK293 cells. Three distinct qPCR amplicons, corresponding to *env*, *gag* and *pro* are shown. Samples were normalized to 18S ribosomal RNA levels. **P* value < 0.05, one-sided *t*-test. Error bars are ± 1 s.d., $n = 3$ biological replicates. **d**, HERVK *gag* or *env* expression in male human ES cell lines HSF-1, HSF-8, female human ES cell line H9 and human EC cell line NCCIT. **P* value < 0.05, one sided *t*-test compared to control siRNA, $n = 3$ biological replicates. Error bars are ± 1 s.d. **e**, RT–qPCR analysis of HERVK transcripts after siRNA knockdown of NANOG, OCT4 or SOX2 in human EC cells (NCCIT). Signals were normalized to 18S rRNA. **P* value < 0.05, one sided *t*-test compared to control siRNA, $n = 3$ biological replicates. Error bars are ± 1 s.d. **f**, ChIP-qPCR analyses of human ES cells (H9) with indicated antibodies. Signals were interrogated with primer sets for positive control regions (active human ES cell *OCT4* and *SOX2* enhancers), LTR5HS, or non-repetitive, intergenic negative regions, as indicated at the bottom. Shown is a representative result of two biological replicates.



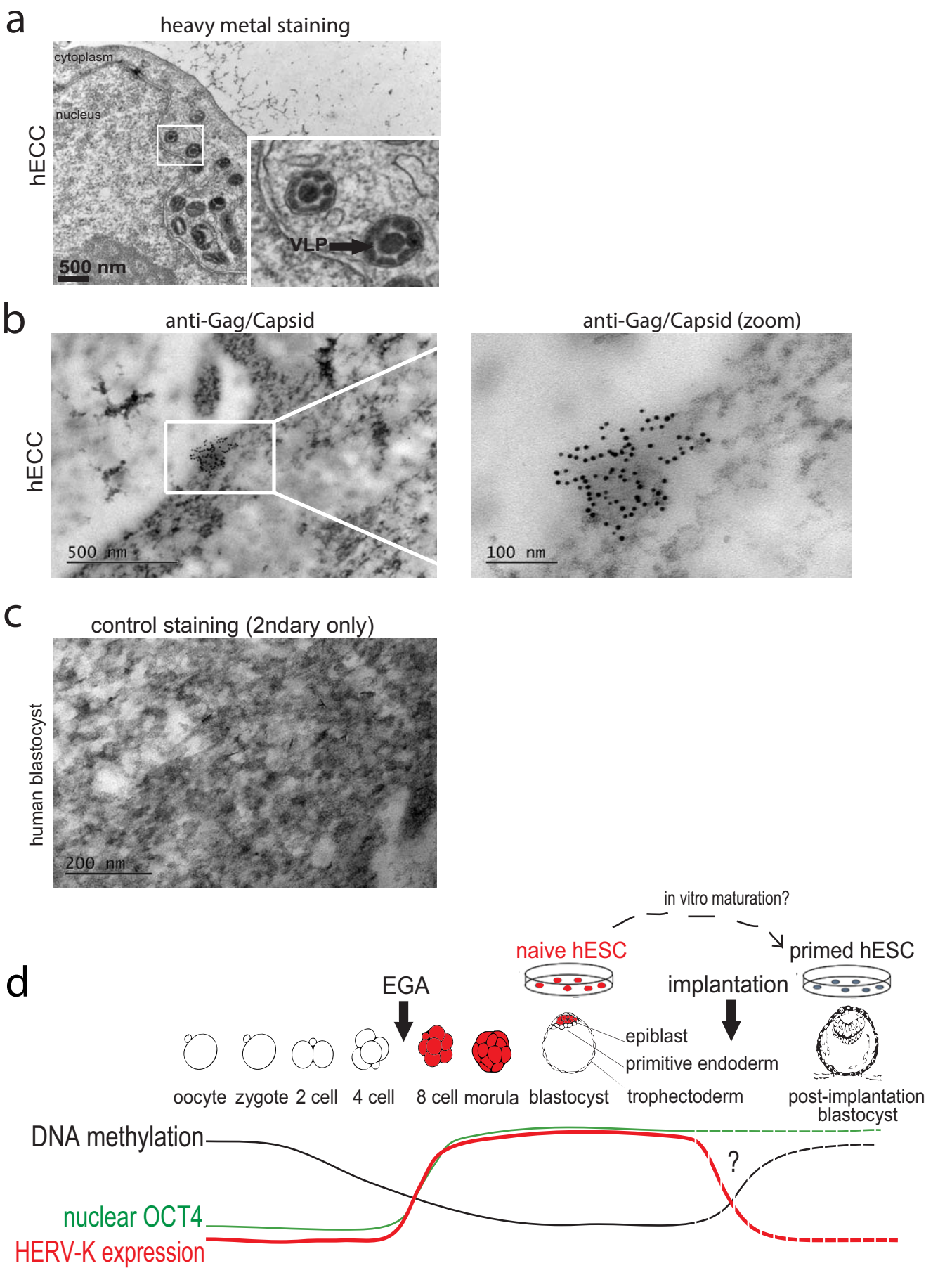




Extended Data Figure 3 | HERVK regulation by OCT4 and DNA methylation (supporting Fig. 2). a, Transcription factor knockdown in human EC cells (NCCIT). Cells were transfected with siRNA pools targeting indicated transcription factors and protein depletion was measured by immunofluoresence with respective antibodies in comparison with control, mock-transfected cells. DAPI (blue), OCT4 (green, left), NANOG (green, middle), SOX2 (green, right). Shown is one of three representative fields of view at ×20 magnification. **b**, Dual luciferase assays with indicated reporter constructs in human EC cells (NCCIT) showing that mutation of OCT4 site decreases reporter activity. $N = 3$ biological replicates, error-bars ±1 s.d. *P value < 0.05, one-sided t-test. SV40 enhancer/promoter construct was used as a positive control. **c**, Bisulfite sequencing for indicated cell types (WT33 human IPSC) analysing consensus LTR5HS-specific amplicon as in Fig. 2e. **d**, Bisulfite sequencing analysis of HERVK proviral consensus amplicon containing 3′ end of LTR, primer binding site, and 5′ region of Gag ORF (see Extended Data Fig. 2a) in indicated cell types: ELF1 naive, human ES cell, WT33 human IPSC, NCCIT human EC cell, or H9 human ES cell. **e**, RT–qPCR analysis of HERVK RNA levels in HEK293 cells treated with indicated concentrations of 5-aza-2′-deoxycytidine for 3 days, followed by transfection with *OCT4/SOX2* expression constructs and RNA collection 48 h after transfection. qPCR primer sets were designed to three independent amplicons of HERVK. *P value < 0.05, one-sided t-test. $n = 4$ biological replicates, error bars ±1 s.d.

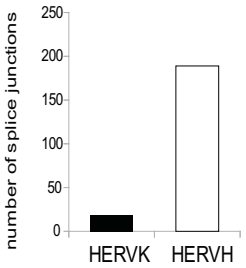


Extended Data Figure 4 | HERVK Gag/Capsid antibody validation and staining (supporting Fig. 3). a, Immunofluorescence analysis of human EC cells (NCCIT) and human ES cells (H9) stained with DAPI (blue), OCT4 (green), Gag/Capsid (red), or IgG control (bottom). White boxes indicate regions shown in higher magnification/merge (right). Shown are representative fields of three independent experiments. **b**, Sensitivity of HERVK Gag/Capsid antibody immunoblot signal to HERVK knockdown. Human EC cells were transfected with one of three independent siRNA pools targeting HERVK Gag or with a control, non-targeting pool (synthesized against RFP) and total protein was analysed by immunoblotting with anti-Env and anti-Gag/Capsid antibodies. 1:2 serial dilution of total protein was loaded, as indicated. Blots were stripped and re-probed with TBP as a loading control. Shown is a representative result of two independent experiments. **c**, Sensitivity of HERVK Gag/Capsid antibody immunofluorescence signal to siRNA knockdown of Gag/Capsid (top) or control siRNA targeting RFP (bottom). Shown is a representative result of three fields of view. Magnification: 20X **d**, Immunoflourescence of naive ELF1 human ES cells with antibodies against OCT4 (green), HERVK Gag/Capsid (pink), DAPI in blue. Region marked with white box on left is shown with larger magnification (bottom). Magnification = 20x, 40x respectively. **e**, Another representative example of immunoflourescence of human blastocysts with DAPI (blue), OCT4 (green) and Gag/Capsid (red) shown ($n = 19$ blastocysts; DPF 5–6). Original magnification, ×40.

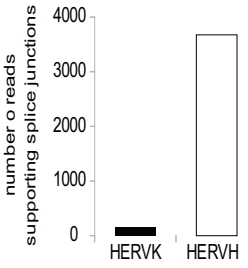


Extended Data Figure 5 | TEM analyses of human EC cells and control embryo staining (supporting Fig. 3). a, TEM analysis of human EC cells (NCCIT) with heavy metal staining; arrow indicates VLPs. Boxed region is shown with higher magnification in an inset. Scale bar = 500 nm. Shown is a representative example of two independent experiments. **b**, TEM immuno-gold labelling of human EC cells (NCCIT) with Gag/Capsid antibodies. Shown is a representative example from two independent experiments. **c**, Secondary antibody only control for immuno-gold labelling of human blastocysts. Shown is a representative example from eight fields of view. **d**, Model figure summarizing HERVK transcriptional regulation in human embryos and *in vitro* cultured pluripotent cells. Dashed lines indicate inference of OCT4, DNA methylation and HERVK level changes at implantation from those observed between naive and primed human ES cells, in the absence of data from actual postimplantation human embryos.

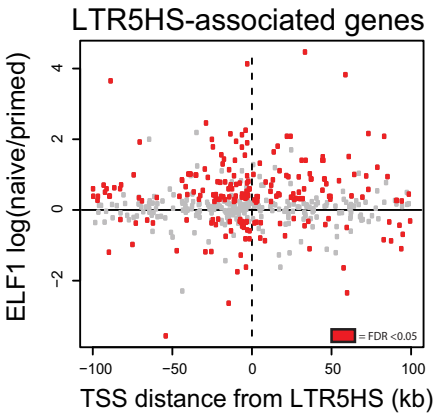
a Chimeric transcript identification (Cufflinks analysis ELF1 naive hESC)

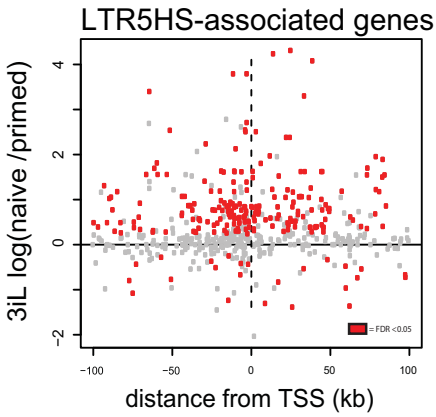


b Chimeric transcript identification (Cufflinks analysis ELF1 naive hESC)

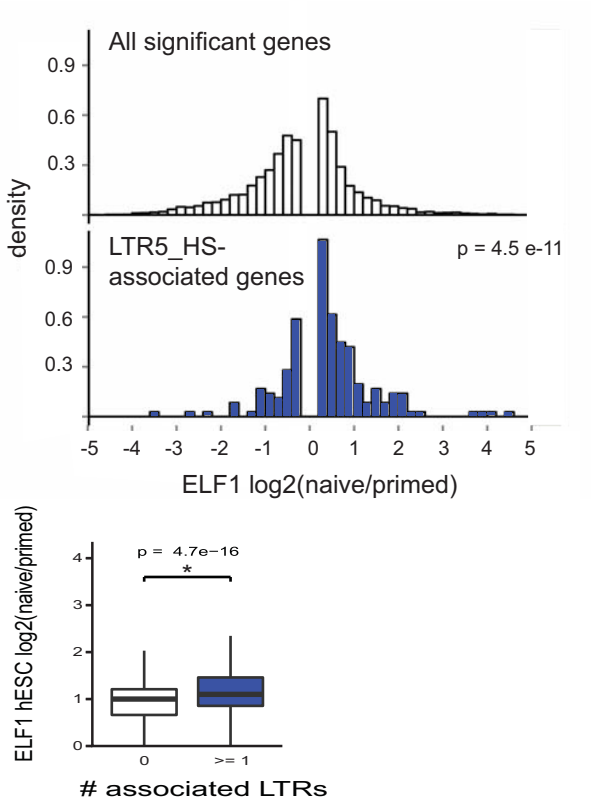


c

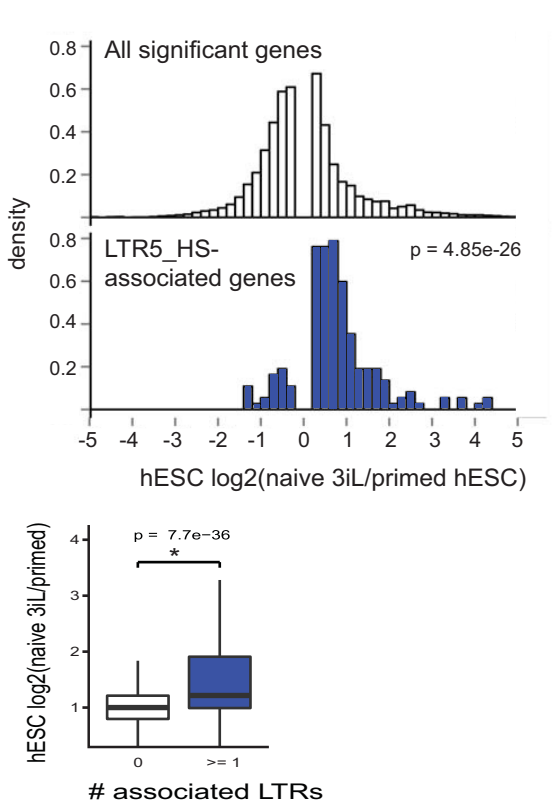




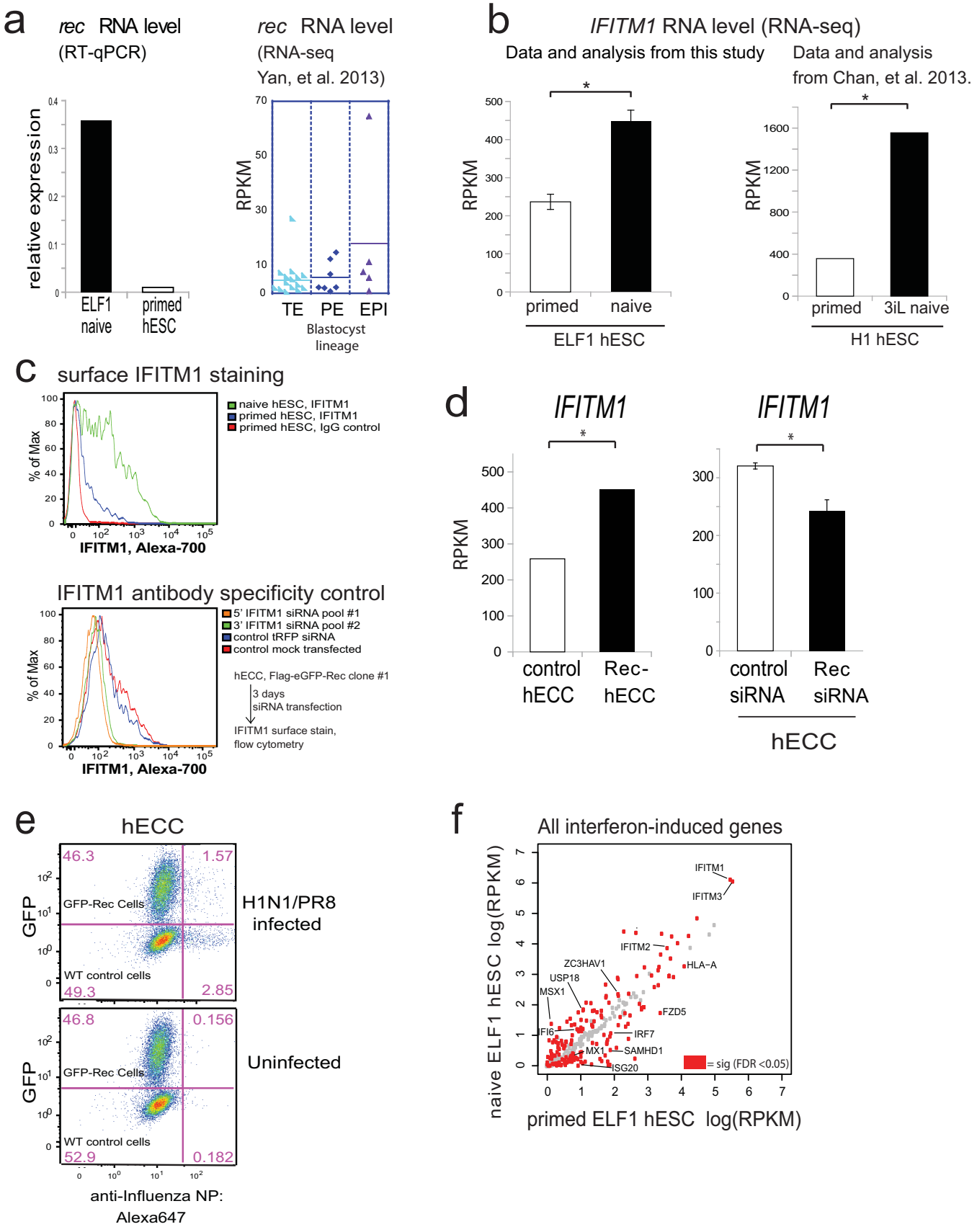
d



e

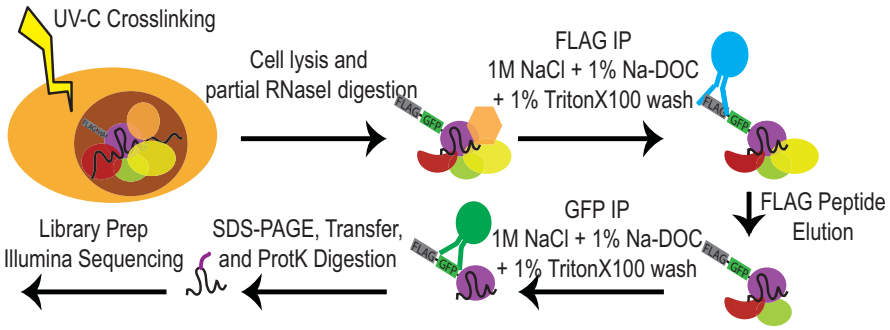


Extended Data Figure 6 | Correlation of HERVK LTR5HS elements with gene expression (supporting Fig. 4). a, Number of splice junctions identified linking indicated HERV class to annotated ReqSeq genes. Analysis was done using RNA-seq data set from ELF1 naive human ES cells, $n = 3$ biological replicates. **b**, Number of reads supporting chimaeric transcripts from indicated HERV class in ELF1 naive human ES cells, $n = 3$ biological replicates. **c**, Expression of LTR5HS linked genes plotted as a function of distance to the gene's transcription start site (TSS). *x*-axis: distance of TSS to the nearest LTR5HS in kb; *y*-axis: fold change in expression of the linked gene in ELF1 naive versus primed human ES cells (this study, left) or expression of the linked gene in 3iL versus primed H1 human ES cells (right, ref. 12). **d**, Top, histograms showing expression of all genes that significantly change in expression between naive and primed ELF1 human ES cells (top histogram, white) or significantly changed genes that are LTR5HS associated (bottom histogram, blue); expression values from naive versus primed ELF1 human ES cell RNA-seq data sets (FDR < 0.05 DESeq). Fischer's exact test gives stated *P* value, indicating enrichment of LTR5HS-linked genes in naive upregulated category. Bottom, quantification of average expression of LTR5HS-linked (blue) or unlinked (white) genes. Non-paired Wilcoxon test with stated *P* value indicating that genes linked to 1 or more LTR5HS have significantly higher mean expression. **e**, Top, histograms showing expression of all genes that significantly change in expression between 3iL and primed H1 human ES cells (top histogram, white) or significantly changed genes that are LTR5_HS associated (bottom histogram, blue); expression values from RNA-seq data sets reported previously[12], FDR < 0.05 DESeq. Fischer's exact test gives stated *P* value indicating enrichment of LTR5HS-linked genes in naive upregulated category. Bottom, quantification of average expression of LTR5HS-linked (blue) or unlinked (white) genes. Non-paired Wilcoxon test with stated *P* value indicating that genes linked to 1 or more LTR5HS have significantly higher mean expression.

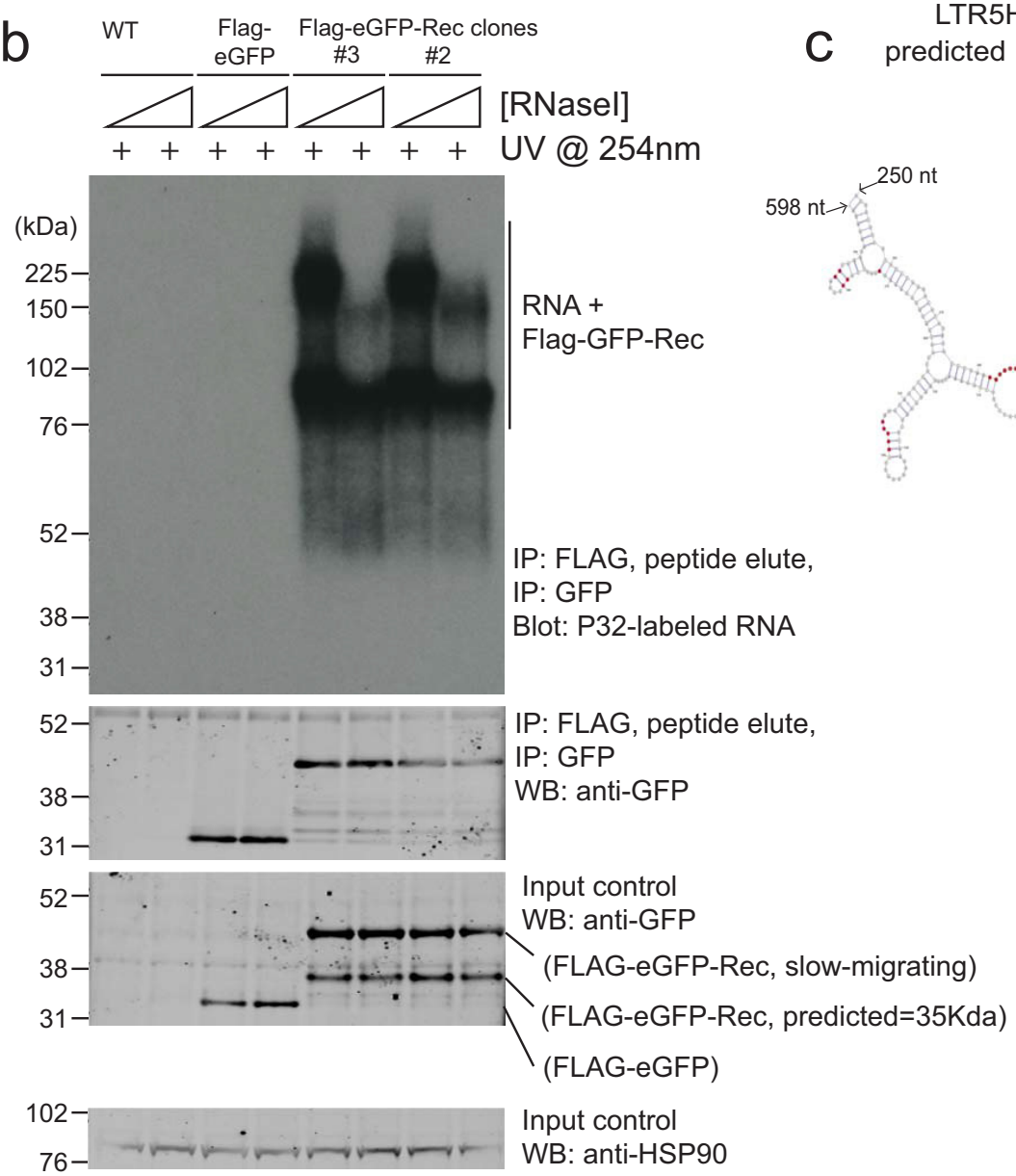


Extended Data Figure 7 | *rec* and *IFITM1* expression in naive human ES cells, and effect of Rec expression on H1N1(PR8) infection (supporting Fig. 4). **a**, Left, RT–qPCR analysis of HERVK *rec* expression levels in ELF1 naive human ES cells ($n = 3$ biological replicates) or H9 primed human ES cells (one biological replicate). Normalized to 18S rRNA. Right, Rec RNA levels in indicated blastocyst lineages. Solid line indicates mean; data are from ref. 10. **b**, RNA-seq quantification of *IFITM1* RNA levels in naive or primed ELF1 human ES cells (left, this study) or 3iL human ES cells versus primed H1 human ES cells from ref. 12 (right). $n = 3$ biological replicates for each condition, error bars are ± 1 s.d. Asterisk indicates significance at FDR < 0.05, DESeq. **c**, Flow cytometry for surface-localized IFITM1 staining in the indicated H9 human ES cells or naive ELF1 human ES cells (top) or, as a control for IFITM1 antibody specificity, knockdown of *IFITM1* with two independent *IFITM1* siRNA pools compared to control siRNA-treated cells in Flag–eGFP–Rec-hECCs (bottom). **d**, Left, *IFITM1* expression in control human EC cell versus Rec-hECC (NCCIT) RNA-seq data sets. $n = 2$ biological replicates. Significance = FDR < 0.05, DESeq. Right, *IFITM1* expression in control siRNA versus Rec siRNA-treated human EC cells (NCCIT) RNA-seq. $n = 3$ biological replicates, error bars are ± 1 s.d. Significance = FDR < 0.05, DESeq. **e**, Flow-cytometry profiles for indicated cell types in H1N1(PR8) infected (top) or non-infected (bottom) wild-type (WT) control human EC cells or Flag–GFP–Rec-hECCs, clone #1. Shown is one representative example of four independent experiments showing a co-plating experiment in which GFP-Rec cells and wild-type control (GFP negative) cells are infected in the same well, stained in the same tube and identified by GFP fluorescence after gating for FSC and SSC. **f**, Scatterplot of ELF1 naive versus primed human ES cell RNA-seq showing all interferon-induced genes, with differentially regulated genes (FDR < 0.05 DESeq, $n = 3$ biological replicates each) highlighted in red. There is a significant overlap between differentially regulated genes and interferon-induced genes as measured by a hypergeometric test (P value < 0.05).

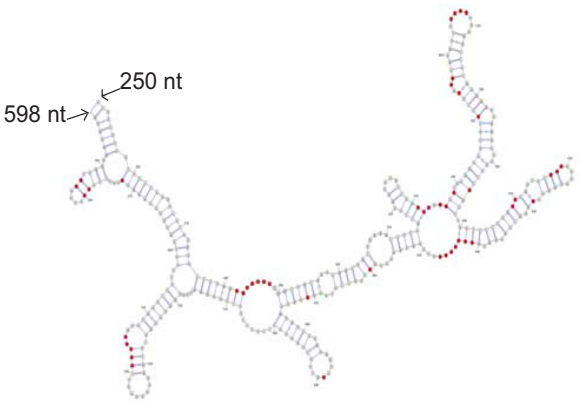
a Diagram of iCLIP-seq procedure



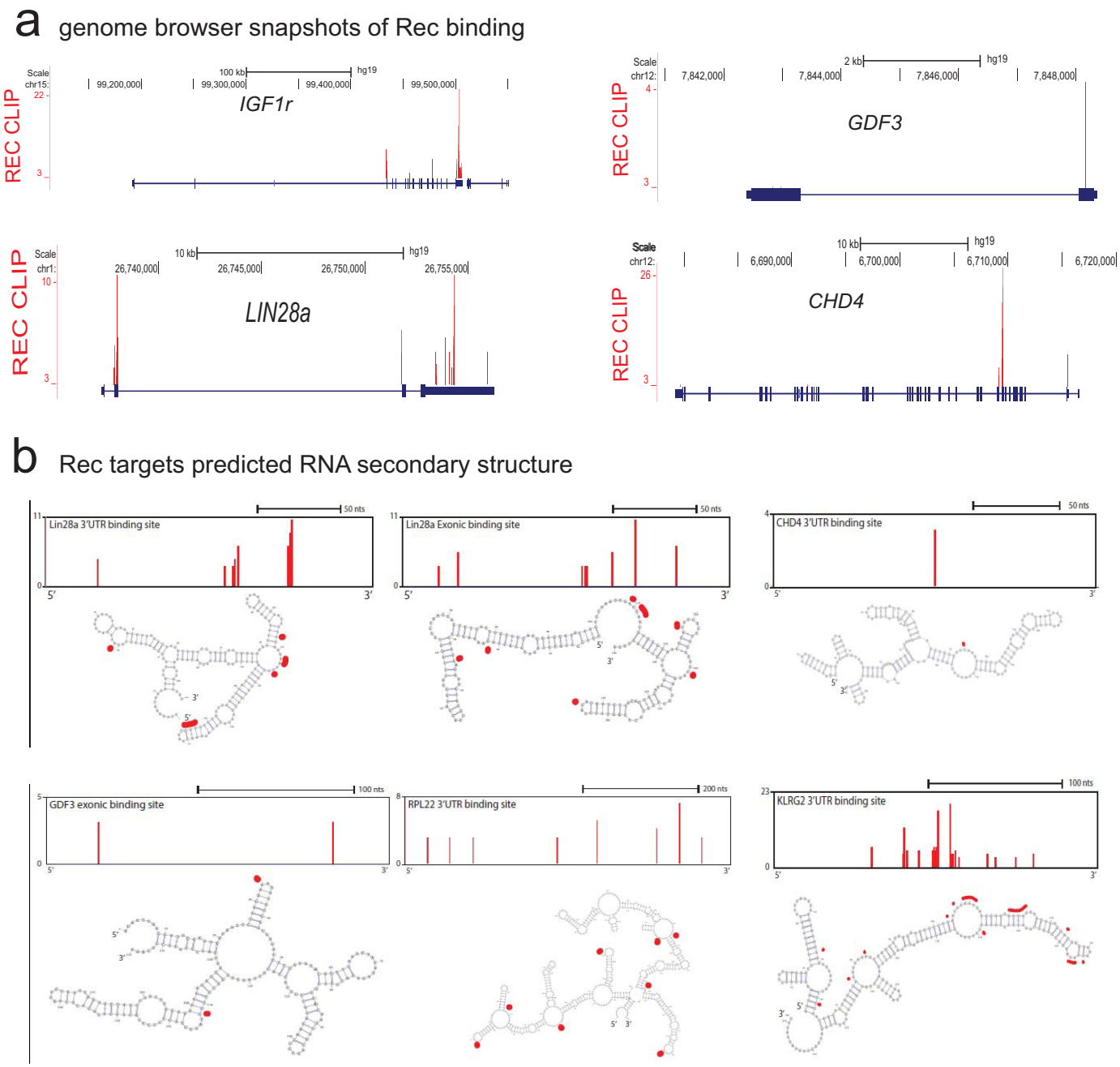
b



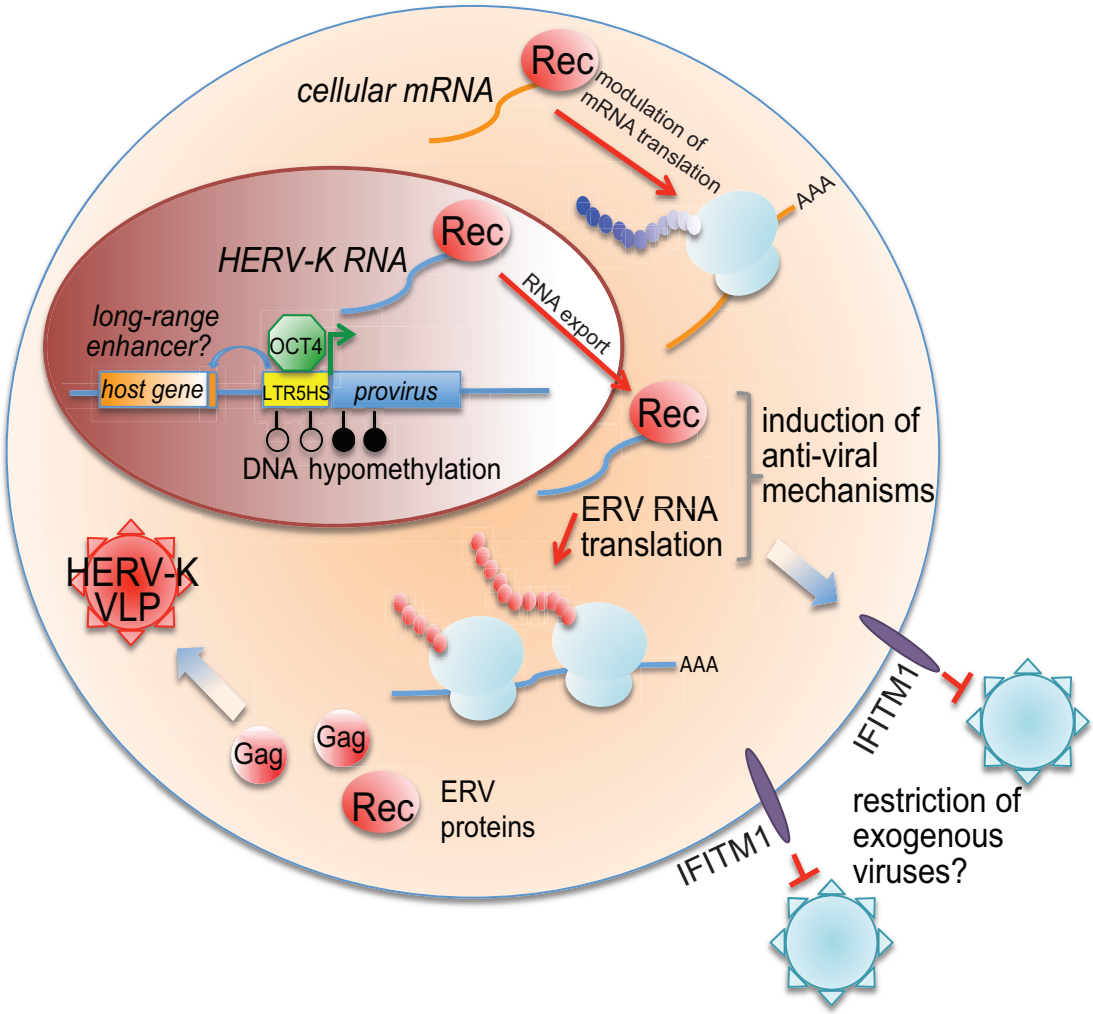
c LTR5HS RRE
predicted RNA structure



Extended Data Figure 8 | iCLIP analysis of Rec-associated RNAs (supporting Fig. 4). a, Diagram of iCLIP-seq procedure (see Methods for details). Briefly, cells are crosslinked using ultraviolet, lysed and digested with RNase to trim RNAs. Sequential immunopurification is performed using Flag M2, peptide elution, and GFP immunoprecipitation (IP). After stringent washing, RNAs are recovered and either radiolabelled (shown in **b**) or reverse transcribed and prepared for Illumina HTPS libraries. **b**, Autoradiogram of labelled RNAs (top) recovered from ultraviolet-crosslinked cells using sequential Flag–eGFP immunoprecipitation from: wild-type human EC cells (lanes 1, 2), Flag–eGFP control human EC cells (lanes 3, 4), or two independent Rec-hECC transgenic lines (lanes 5–8), separated on an SDS–polyacrylamide gel electrophoresis (SDS–PAGE) gel. Free Rec protein runs as a ~35 kDa band, while Rec protein crosslinked to RNA molecules show lower electrophoretic mobility. Please note that: (1) Rec-bound RNAs are resistant to even high concentrations of RNaseI, probably indicating extensive secondary RNA structures, and (2) low/no background of contaminating RNAs in control immunoprecipitation from wild-type human EC cells or Flag–eGFP control human EC cells. Western blots with anti-GFP antibody were also performed to confirm the presence of tagged protein in Flag–eGFP control and Flag–eGFP–Rec cells, both in input and immunoprecipitation fractions (middle). HSP90 was used as a loading control (bottom). **c**, Computationally predicted (using mFold) secondary structure of LTR5HS sequence around the Rec response element (identified experimentally *in vitro* previously[25]). Single nucleotide resolution Rec ultraviolet-crosslinking sites determined by iCLIP are shaded in red; $n = 2$ biological replicates.



Extended Data Figure 9 | Rec target mRNA analysis (supporting Fig. 4). **a**, Genome browser representations of the Rec iCLIP read ($n = 2$ biological replicates) distribution at indicated mRNA targets. **b**, Computationally predicted (using mFold) secondary structures of indicated Rec iCLIP-seq targets. Single-nucleotide resolution Rec ultraviolet-crosslinking sites determined by iCLIP are shaded in red; to orient the reader, browser representation of the folded fragment is shown above each respective cartoon.



Extended Data Figure 10 | Model of HERVK regulation and function.