HHS Public Access

Author manuscript

Nature. Author manuscript; available in PMC 2015 December 11.

Published in final edited form as:

Nature. 2015 June 11; 522(7555): 221–225. 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**[3,4,10], **Daniel Wesche**[3], **Lance Martin**[2], **Carol Ware**[7], **Catherine A. Blish**[8], **Howard Y. Chang**[2], **Renee A. Reijo Pera**[3,4,9,12], and **Joanna Wysocka**[3,10,11,*]

[1]Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305, USA

[2]Howard Hughes Medical Institute and Program in Epithelial Biology, Stanford University School of Medicine, Stanford, CA 94305, USA

[3]Institute for Stem Cell Biology & Regenerative Medicine, Stanford University School of Medicine, Stanford University, Stanford, CA 94305, USA

[4]Department of Obstetrics and Gynecology, Stanford University School of Medicine, Stanford University, Stanford, CA 94305

[5]Division of Reproductive and Developmental Sciences, Oregon National Primate Research Center, Oregon Health & Science University. Beaverton, OR 97006, USA

[6]Department of Immunology, Stanford University School of Medicine, Stanford, CA 94305, USA

[7]Department of Comparative Medicine, University of Washington, Seattle, WA 98195-8056, USA

[8]Department of Medicine, Stanford University School of Medicine, Stanford, CA 94305, USA

[9]Department of Cell Biology and Neurosciences, Montana State University, Bozeman, MT 59717

[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

[12]Department of Cell Biology and Neurosciences, Montana State University, Bozeman, MT 59717

Summary

Endogenous retroviruses (ERVs) are remnants of ancient retroviral infections, which comprise nearly 8% of the human genome[1]. The most recently acquired human ERV is HERV-K (HML-2),

*Correspondence and requests for reagents should be addressed to J. W. wysocka@stanford.edu.

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 majority of the experiments and data analyses. S.L.C., M.W. and E.J.G. performed human blastocyst handling and IF, with expertise and resources provided by R.A.R.P. R.A.F., L.M., H.Y.C. performed and analyzed iCLIP experiments. R.A.F. provided assistance with ribosome profiling experiments and analysis. N.L.B and C.B. contributed influenza infection experiments. D.W. performed expression analysis of LTR5HS-associated genes.

Author information: The authors declare no competing financial interests.

which repeatedly infected the primate lineage both before and after the divergence of humans and chimpanzees[2,3]. Unlike most other human ERVs, HERV-K retained multiple copies of intact open reading frames (ORFs) encoding retroviral proteins[4]. However, HERV-K is transcriptionally silenced by the host with exception of certain pathological contexts, such as germ cell tumors, melanoma, or HIV infection[5–7]. Here we demonstrate that DNA hypomethylation at LTR elements representing the most recent genomic integrations, together with transactivation by OCT4, synergistically facilitate HERV-K expression. Consequently, HERV-K is transcribed during normal human embryogenesis beginning with embryonic genome activation (EGA) at the 8-cell stage, continuing through the emergence of epiblast cells in pre-implantation blastocysts, and ceasing during hESC derivation from blastocyst outgrowths. Remarkably, HERV-K viral-like particles and Gag proteins are detected in human blastocysts, indicating that early human development proceeds in the presence of retroviral products. We further show that overexpression of one such product, HERV-K 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 HERV-K can induce viral restriction pathways in early embryonic cells. Moreover, Rec directly binds a subset of cellular RNAs and modulates their ribosome occupancy, arguing that complex interactions between retroviral proteins and host factors can fine-tune regulatory properties of early human development.

Given the substantial contribution of transposable elements (TEs) to human genome and their emerging roles in shaping host's regulatory networks[8,9], understanding dynamic expression and function of TEs is important for dissecting both human- and primate-specific aspects of gene regulation and development. We utilized published single-cell RNA-seq datasets to analyze expression of major TE classes at various stages of human preimplantation embryogenesis[10], a developmental period associated with dynamic changes in DNA methylation and TE expression[11]. This analysis revealed two major clusters, one consisting of repeats that begin to be transcribed at the onset of embryonic genome activation (EGA), which in humans occurs around the 8-cell stage, and a second cluster of repeats, whose transcripts can be detected in the embryo prior to EGA, indicating maternal deposition (Extended Data Fig. 1a). Within each cluster, more discreet 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 lineages of the blastocyst (Extended Data Fig. 1a). For example, human endogenous retrovirus HERV-K and its regulatory element, LTR5HS, were both induced in 8-cell stage embryos, morulae, and continued to be expressed in epiblast (EPI) cells of the blastocysts (Fig. 1 a, b, c and Extended Data Fig. 1a). We further observed that although HERV-K was expressed in blastocyst outgrowths (passage 0 hESC), it was downregulated by passage 10 (Fig. 1d). In contrast, transcripts of another HERV, HERV-H, and its regulatory element LTR7, were detected prior to EGA and throughout preimplantation development, including all blastocyst lineages and hESCs (Extended Data Fig. 1a, b, c).

Recent studies have reported conditions for capturing a human naïve pluripotent state *in vitro*[12–16], and we used RNA-seq to analyze the repetitive transcriptome of ELF1, a cell line derived from an 8-cell stage human embryo under naïve culture conditions, and compared it to the repeat expression in ELF1 cells matured *in vitro* into a primed state[14]. Surprisingly,

although many TE classes (e.g. HERV-H 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 HERV-K proviruses and their regulatory elements, LTR5HS (but not the older LTR5a or LTR5b; see below), were among the most strongly induced in naïve vs. primed ELF1 cells (Fig. 1e, Extended Data Fig. 1d). Similar results were obtained by analyzing available transcriptomes of primed H1 hESC and naïve 3iL cells derived from them, as well as of primed H9 hESC and those 'reset' to the naïve state by NANOG/KLF2 transgene expression[12,15](Fig. 1e). Therefore, naïve-state specific upregulation of HERV-K is consistent across multiple genetic backgrounds, derivation methods or culture conditions.

From an evolutionary perspective, HERV-K 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 HERV-K 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 HERV-K was active within the last 200,000 years[18](Extended Data Fig. 2a). All human-specific and human-polymorphic HERV-K 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 naïve state, transcripts originating from LTR5HS, but not LTR5a or LTR5b are preferentially expressed (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 692–699bp 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 if OCT4 binding contributes to the transcriptional activation of LTR5HS, we used pluripotent NCCIT human embryonic carcinoma cells (hECCs), which express OCT4, but in contrast to hESCs, are permissive for HERV-K expression[5,19](Extended Data Fig. 2b–d). ChIP-qPCR analysis of hECCs showed preferential occupancy of OCT4, p300 and histone marks of active chromatin at LTR5HS elements, as compared to the LTR5a/LTR5b (Fig. 2c). In contrast, we did not detect OCT4 or p300 binding at LTR5HS in hESCs (Extended Data Fig. 2f). Consistent with a functional role in HERV-K activation, knockdown of OCT4 or SOX2, but not of NANOG led to a significant decrease in viral transcripts in hECCs (Extended Data Fig. 2e, 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 HERV-K in primed versus naïve hESCs and hECCs, as all three express OCT4. We hypothesized that DNA methylation may contribute an additional layer of regulation, and indeed we observed HERV-K hypomethylation of solo and proviral LTR5HS (but not the Gag ORF) in hECCs and naïve cells, as compared to conventionally grown hESCs and hiPSCs (Fig. 2e, 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 HERV-K coding sequences remained more highly methylated[11]. Importantly, treatment of primed hESCs with a DNA methylation inhibitor 5-aza-2'-deoxycytidine for 24 hours induced HERV-K 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/SOX2, jointly facilitated HERV-K transcription in HEK293 cells (Fig. 2g and Extended Data Fig. 3e), indicating that DNA hypomethylation and transactivation by OCT4 synergistically promote HERV-K expression.

A defining characteristic of HERV-K is that multiple proviruses have retained ORFs encoding full-length retroviral proteins[4]. Consequently, HERV-K reactivation in pathological conditions has been associated with the presence of HERV-K proteins[5–7], prompting us to examine if retroviral proteins are also present in human embryos. We used a well-characterized monoclonal antibody recognizing HERV-K Gag precursor and its proteolytically processed form Capsid, which detects cytoplasmic signal with a characteristic punctate pattern in hECCs and a subset of naïve ELF1 cells, but shows no staining in hESC and loss of signal in *Gag* siRNA knockdown hECCs (Extended Data Fig. 4a,d,b,c). In human blastocysts, Gag/Capsid staining was also detected in dense cytoplasmic puncta resembling those seen in hECCs and naïve ELF1 cells (Fig. 3a and Extended Data Fig. 4a,d,e), with all analyzed blastocysts (n=19/19) showing robust signal.

Germ cell tumors and certain HERV-K-positive hECC lines have been shown to produce viral-like particles (VLPs)[20]. Remarkably, heavy metal staining/transmission electron microscopy (TEM) of blastocysts revealed presence of cytoplasmic, electron-dense particles of approximately 100 nm in diameter—the reported size of reconstructed HERV-K 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 hECCs (Fig. 3c and Extended Data Fig. 5a). The presence of HERV-K-derived particles in human blastocysts was further supported by immuno-gold TEM staining, which detected VLPs (or vesicles with multiple VLPs) labeled by Gag/Capsid antibodies either within embryonic cells or on the cell surface, similar to those seen in immuno-gold TEM staining of hECCs (Fig. 3d,e and 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 the mouse have been reported to function as alternative promoters, driving expression of many '2-C' specific chimeric transcripts[23]. However, we did not detect robust evidence for HERV-K associated chimeric 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 naïve

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 HERV-K integration near genes active in the naïve state.

HERV-K encodes a small accessory protein Rec, homologous to the 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 naïve cells, human blastocysts, and rapidly induced in primed hESC 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 anti-viral responses, which typically rely on cytosolic detection of viral RNA and protein. We noted a striking induction of mRNA encoding an interferon-induced viral restriction factor *IFITM1*[26] (also known as *FRAGILIS2*) has been reported in human epiblast cells[10], as well as upregulation of *IFITM1* transcripts and surface protein levels in human naïve versus primed hESCs (Extended Data Fig. 7b,c,f and Supplementary Table 6). Furthermore, expression of a Rec transgene in hECCs was sufficient to elevate surface-localized IFITM1 protein levels (Fig. 4a). This was at least in part mediated through effect on *IFITM1* mRNA transcription/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 hECCs, many other interferon induced genes are not upregulated or expressed, indicating that HERV-K triggers a precise antiviral response in host cells (Supplementary Table 2). To test whether HERV-K expression provides viral resistance, we infected control wild type hECCs, control hECCs expressing a GFP transgene, or two independent clonal Rec-hECC lines with influenza H1N1(PR8) virus. Interestingly, Rec-hECC exhibited substantially attenuated infection levels as compared to the control GFP-hECC (Fig. 4b) or wildtype hECCs (Extended Data Fig. 7e).

Retroviral accessory proteins often masterfully manipulate host cell factors to achieve optimal replicative efficiency. To examine if, beyond reported binding to HERV-K 3' LTRs[25,27], Rec can also associate with cellular RNAs, we performed tandem affinity purification iCLIP-seq in hECCs expressing FLAG-eGFP or FLAG-eGFP-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-eGFP-Rec transgenic lines yielded UV-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 ∼1600 host mRNAs, preferentially in their 3' UTRs, a positional preference analogous to that observed in the viral RNA (Fig. 4d,e and Extended Data Fig. 9a, 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 Rec's interaction with its HERV-K 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 (e.g. *FGFR1, FGF13, FGFR3, KLGR2, IGFR1, FZD7, GDF3*) and chromatin regulators (e.g. *DNMT1, CHD4*) (Extended Data Fig. 9a, 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 overexpressing hECCs (Rec-hECCs), in comparison to wild type hECCs, 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 (hypergeometric test, p-value <0.05), but we did not observe such enrichment for mRNAs bound in their 5' UTRs. We also noticed that several Rec-bound transcripts encoding ribosome components and translation regulators (e.g. *RPL22, RPL31, RPS13, RPS20, EIF4G1*) had increased occupancy in Rec-hECCs, 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 HERV-K, translation of its ORFs, and assembly of VLPs (Extended Data Fig. 10a). Beyond preimplantation development, we predict that HERV-K reactivation occurs in human primordial germ cells (PGCs), which are also characterized by the presence of OCT4 and genome-wide DNA hypomethylation[29]. HERV-K 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 HERV-K 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 HERV-K provides an immunoprotective effect for human embryos against different classes of viruses sensitive to the IFITM1-type restriction. Although IFITM (a.k.a. *FRAGILIS*) proteins were first described as interferon-induced genes, they are also classical naïve 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 an evolutionarily conserved mechanism protecting both embryos and germ cells from either exogenous viral infection or reinfection from infectious ERVs (Extended Data Fig. 10a).

Full Methods

DNA and RNA isolation at reverse transcription

Genomic DNA was isolated using phenol:chloroform:isoamyl (100:100:1) (Invitrogen). Briefly, cells were digested in 10mM Tris-HCl (pH=8.0), 0.1M EDTA, 0.5% SDS for 37C for 1 hour, then proteinase K was added to final concentration of 100ug/mL and then incubated for 3 hour at 50C. DNA was PCI extracted, ethanol precipitated, and resuspended in TE. RNA was extracted using Trizol (Invitrogen) according to manufacturers instructions. DNAse treatment with Turbo DNAse (Ambion) was performed at 30 min for 37C, PCI

extracted, ethanol precipitated, resuspend in water. Reverse transcription was performed with SuperScript III (Invitrogen) using ∼500 ng of DNAase treated total RNA following manufacture instructions. No reverse transcriptase controls were performed where necessary.

Cell lines and culture

NCCIT, HEK293 cells were obtained from ATCC. NCCIT cells were maintained in 10% FBS (Omega), 1× Glutamax-I supplement (100× stock, Invitrogen), 1xNon-essential amino acids (100x stock, invitrogen), and basal media RPMI 1640 (Hyclone). HEK293 cells were maintained in 10% FBS, 1x NEAA, 1x glutamax in DMEM-high glucose (Hyclone). hESCs (H9 line, 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) hESC were used at passage 20–28, cultured as described above and their characterization is described elsewhere[31]. Cells were passaged 1:7 every 5–6 days by incubation with accutase (Invitrogen) and 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 naïve hESC were obtained from Dr. Carol Ware and cultured as previously described[14], with 10ng/mL human recombinant LIF (R&D). Cell cultures were routinely tested and found negative for mycoplasma infection (MycoAlert, Lonza).

Chromatin Immunoprecipitation

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). 50–75uL of protein G dynal beads (Invitrogen) were used to capture 3–5ug of antibody in phosphate citrate buffer pH=5.0 (2.4 mM citric acid, 5.16 mM Na2HPO4) for 30 min at 27C. Antibody bead complexes were rinsed 2x with PBS and added to sonicated chromatin and rotated at 4C overnight. 10% 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 analyzed on CS&T calibrated BD FACS Aria II SORP flow cytometer on 561 nm laser line for turboRFP, with 582/15BP. For IFITM1 flow cytometry, cells were allowed to recover after trypsinization for 2 hours at 37C 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 4C. 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 Bisulfite conversion kit (Qiagen) was used to bisulfite convert 1 ug genomic DNA as per manufacturer instructions. ~20 ng of BS-treated DNA was used as a template for 35–40 cycles with Platinum taq (Invitrogen, 10966) as per manufacturer 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 subjected to Sanger sequencing for both forward and reverse strands. BiQ software was used to align and quantify CpG methylation.

Protein extraction/immunblotting

Proteins were extracted using previously described protocols[33]. Briefly, cells were resuspended in buffer A (10mM Hepes, pH=7.9, 10mM Kcl, 1.5mM MgCl, 0.34M sucrose, 10% glycerol) and fresh protease inhibitors (Complete EDTA-free, Roche), 1mM PMSF, and 0.1% Triton-X 100 were added. Cytoplasmic extract was further clarified by centrifugation at 13,000 RPM at 4C for 10 minutes, 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 hour at 27C. Primary antibodies (using dilutions listed in a antibodies section) were used in blocking solution overnight at 4C. 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 SD.

Plasmid and constructs

HERV-K LTR5_HS sequence from HERV-K-con[22] was cloned upstream of miniTK promoter driving turbo RFP and inserted into piggy-back transposon (SystemBio). Motif mutations for OCT4 or SOX2 were produced by replacing the respective motif with NotI site. 2.5ug of reporter vector along with 0.5ug of piggy-back transposase were transfected into cells using 18 uL lipofectamine2000 (Invitrogen) in 6-well plates. 400Ug/mL G418 (Amresco) was used to select for integrants. Cells were analyzed >10 days later to minimize signal from nonintegrated reporter expression. cDNAs encoding OCT4 or SOX2 were cloned into pcDNA containing C-terminal or N-terminal Flag-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 ∼30x endogenous expression of Rec mRNA (as measured by qPCR) and Flag-eGFP-Rec clone #2 has ∼14x endogenous expression of Rec mRNA (qPCR), data not shown.

siRNA knockdown

siRNA was generated using baculovirus produced giardia Dicer as described[34]. Briefly 1 ug of PCR product was *in vitro* transcribed using Megascript T7 (Ambion) and digested using dicer at 37C for 16 hours. siRNA was purified using Purelink RNA mini Kit (Ambion), absence of >22nt RNA was verified using gel electrophoresis and ethidium bromide staining. NCCIT cells were plated onto matrigel coated 24-well plates, transfected using 1.5 uL of RNAi-max (Invitrogen) in optimem (Gibco) with 25nM siRNA concentrations for 4 hours before addition of fresh media. siRNA knockdowns were performed for three consecutive days, cells were harvested 24 hours after final transfection. Two independent siRNA pools were generated for *OCT4*, *NANOG*, *SOX2*, one each for turboRFP (non-targeting control) and *Rec*, which overlaps the *Env* ORF. Primers used to generate 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 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, Trumbull, CT) 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−1sucrose solution for 10 min followed by a 0.2 mol l−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 either Quinn’s advantage cleavage medium (CooperSurgical) or Quinn’s advantage cleavage medium (CooperSurgical) supplemented with 10% serum protein substitute under mineral oil (Sigma, St Louis, MO). The embryos were cultured at 37 °C with 6% CO2, 5% O2 and 89% N2 under standard human embryo culture conditions in accordance with current clinical IVF practice. Embryos used in this study were days post fertilization (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 27C, washed 3x 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 for 1 hour at 27C. Primary antibodies were resuspended in antibody buffer and incubated at 4C overnight. Washes were performed 3x using 0.1% Triton-X 100 in PBS, and secondary antibodies were added for 1 hour at 27C 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 (RT). 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 hour at RT. 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 w/ primary antibodies in PBS-T with 1% serum sequentially for 1 hour each at RT 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 RT for 1 hour 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-quantitiative comparisons between embryos.

DNA demethylation treatment

HEK293 cells were plated on matrigel coated 24-well plates, and treated with 0, 1, or 10 micromolar 5-aza-2'-deoxyctidine (Calbiochem) freshly prepared every 24-hours. Cells were then transfected with 1 microgram each of pcDNA3.1-OCT4 and pcDNA3.1-SOX2 expression plasmids. Media was changed 24 hours later, and cells were harvested 3 days after transfection for RNA analysis. HESC (H9) were grown as described above, except mTeSR was supplmented with Rock-inhibitor (y-27632, Sigma) at 5 micromolar, and treated with 0, 1, or 10 micromolar 5-aza-2'-deoxyctidine (Calbiochem) for 24 hours.

RNA-seq datasets

Chan, *et al* 2013: Array Express Database (E-MATB-2031). Yan, *et al* 2013: GEO (GSE36552). Xue, *et al* 2013: GEO (GSE44183). Takashima, *et al*. 2014: in Array Express (E-MTAB-2857). Sequencing datasets generated for this study are deposited under the GEO accession GSE63570, and summarized in Supplementary Table #8.

RNA-seq library construction

Libraries were constructed as described[33],using ∼10 micrograms of total RNA followed by poly-A selection with oligo-dT beads, ligation and 10 cycles of PCR with NEBnext kit oligos, and sequenced using Illumina Hi-Seq2000 at the Stanford Sequencing Facility or ELIM Bio (Hayward, CA).

Sequence analysis

For RNA-seq repeat analysis of data from embryo and hESC libraries (for Fig. 1, Extended Data Fig. 1), FASTQ files were aligned to repbase consensus sequences 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, i.e. these settings do not allow multiple-matching. For Fig. 2b analysis of HERV-K proviruses, RNA-seq reads were aligned to hg19 using the same parameters described above, and the overlap between the manually curated HERV-K provirus dataset[5] is reported. For RefSeq analysis for RNA-seq libraries generated for this paper (ELF1 naïve or primed hESC; from hECC siRNA-RNA-seq, or Rec-hECC versus wildtype hECC 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 above was performed using DESeq, with reported FDR using Benjamini-Hochberg correction.

Interferon induced gene set analysis

Genes were defined as interferon induced if 5-fold induced in interferon treated cells/tissues for experimentally deposited data sets found in Interferome database[40] (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 lab, Stanford University) with a cut-off of 100 Kb distance from TSS. These classified Refseq genes were then compared using the RPKM and DESeq analysis as described above. Differential enrichment of LTR5HS associated transcripts in naïve/primed upregulated versus naïve/primed downregulated was analyzed using non-paired Wilcoxon Test, and significance is reported at p-value <0.05. Higher average naïve/primed RPKM of LTR5HS-associated versus non-LTR5HS associated genes was tested using non-paired Wilcoxon Test.

Chimeric transcript identification

100bp paired-end RNA-seq reads generated with ELF1 naïve versus primed hESC (see above) were analyzed 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) cataloging known genes and 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 chimeric 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 log2 transformed RPKM.

Statistical Tests

A list of the statistical tests, multiple-hypothesis testing correction, and normality criteria for parametric tests are 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-HERV-K Gag/Capsid using overnight incubation at 4°C, and labeling was visualized using 5nm gold-labeled anti-mouse secondary. Secondary only controls demonstrated specificity of the antibody for this application. TEM was performed at the Electron Microscopy core at Stanford University using Jeol JEM-1400 electron microscope.

iCLIP and data analysis

The iCLIP method was performed as described before with the specific modifications below[37] . FLAG-GFP-Rec (FG-Rec) expressing NCC cells were UV-C crosslinked to a total of 0.3J/cm^2. Each iCLIP experiment was normalized for total protein amount, typically 1mg, and partially digested with RNaseI (Life Technologies) for 10 minutes at 37°C and quenched on ice. FG-Rec was isolated with antiFLAG agarose beads (Sigma) for 3 hours at 4°C on rotation. Samples were wash sequentially in 1mL for 5min each at 4°C: 2x high stringency buffer (15mM Tris-HCl pH7.5, 5mM EDTA, 2.5mM EGTA, 1% TritonX-100, 1% Nadeoxycholate, 120mM NaCl, 25mM KCl), 1x high salt buffer (15mM Tris-HCl pH7.5, 5mM EDTA, 2.5mM EGTA, 1% TritonX-100, 1% Na-deoxycholate, 1M NaCl), 1x NT2 buffer (50mM Tris-HCl pH7.5, 150mM NaCl, 1mM $MgCl_2$, 0.05% NP-40). Purified FG-Rec was then eluted off antiFLAG agarose beads using competitive FLAG peptide elution. Each sample was resuspended in 500μL of FLAG elution buffer (50mM Tris-HCl pH7.5, 250mM NaCl, 0.5% NP-40, 0.1% Na-deoxycholate, 0.5mg/mL FLAG peptide) and rotated at 4°C for 30 minutes. The FLAG elution was repeated once for a total of 1mL elution. FG-Rec was then captured using antiGFP antibody (Life Technologies, A-11122) conjugated to Protein A dynabeads (Life Technologies) for 3 hour at 4°C on rotation. Samples were then wash as previously in the antiFLAG agarose beads. 3'-end RNA dephosphorylation, 3'-end ssRNA ligation, 5' labeling, SDS-PAGE separation and transfer, autoradiograph, RNP isolation, ProteinaseK treatment, and overnight RNA precipitation took place as previously described[37]. 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[37,39]. 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 1×75bp 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[37,39]. Bioinformatic pipeline used for iCLIP data analysis is described in[39]. Briefly, RT stops were used to map nucleotide resolution of Rec binding, and only nucleotides supported with 3 independent RT stops in two replicates (with at least 1 RT stop in each replicate) were reported as binding events, and reported in Supplementary Table #3.

Ribosome Profiling

hECC (NCCIT) cells were cultured as described above. Total RNA was extracted using Trizol (Life Technologies) and used as input material for the ARTseq Ribosome Profiling Kit – Mammalian (Epicentre) following the manufactures 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 1×75bp 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[38]. Briefly, reads were aligned to 45s rDNA repeat sequence with bowtie to remove residual rRNA reads from libraries. Non-aligning reads (mRNA) 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.

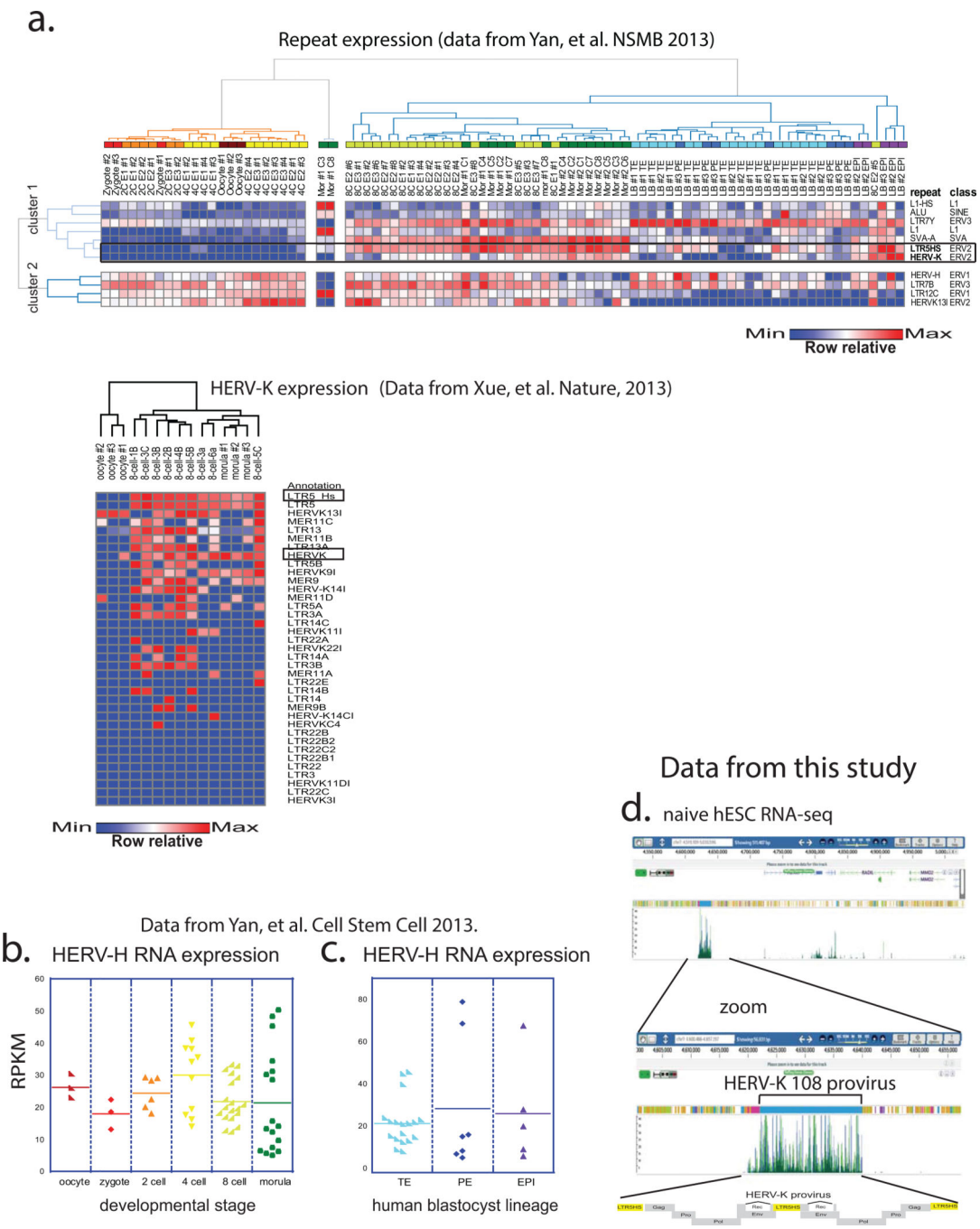
Influenza infection experiments

hECCs (NCCIT) were plated in duplicate (1.5×10^5 cells/well) on a 96-well flat-bottom plate in 100 μl Virus Diluent (DMEM, Gibco supplemented w/ 1% BSA, 1x antibiotics, and 20 mM HEPES). Cells were incubated at 37°C and 5% CO_2 for 1.5 hrs. WT-hECC and REC-hECC 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 hr. FBS (Hyclone) was added to the wells to a final concentration of 10% FBS. Cells were incubated at 37°C for 5 hrs. 20mM EDTA (20 μl) was added to all wells and mixed thoroughly to stop infection. Cells were washed with 200 μl 1x PBS (Hyclone), resuspended in 100 μl 1x 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. 100 μL FACS wash (1x HyClone DPBS with 2% FBS) was added to each well and plate was centrifuged. Cell pellets were resuspended in 200 μl BD FACS Permeabilizing Solution II (BD Biosciences). Cells were incubated at RT 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. Cells were incubated in the dark at RT for 30 min and washed twice. Cell pellets were resuspended in 2 μg/mL of secondary antibody (chicken anti-mouse Alexa647, Invitrogen) in 50 μl FACS wash and incubated in the dark at RT for 30 min. Cells were washed twice and cell pellets were

resuspended in 1% PFA (Electron Microscopy Sciences). Cells were analyzed 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 Ig, κ (BD) and negative control beads. Single stained cellular controls were run in parallel to infected and uninfected samples. Data was analyzed 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-only-hECC cells for each run.

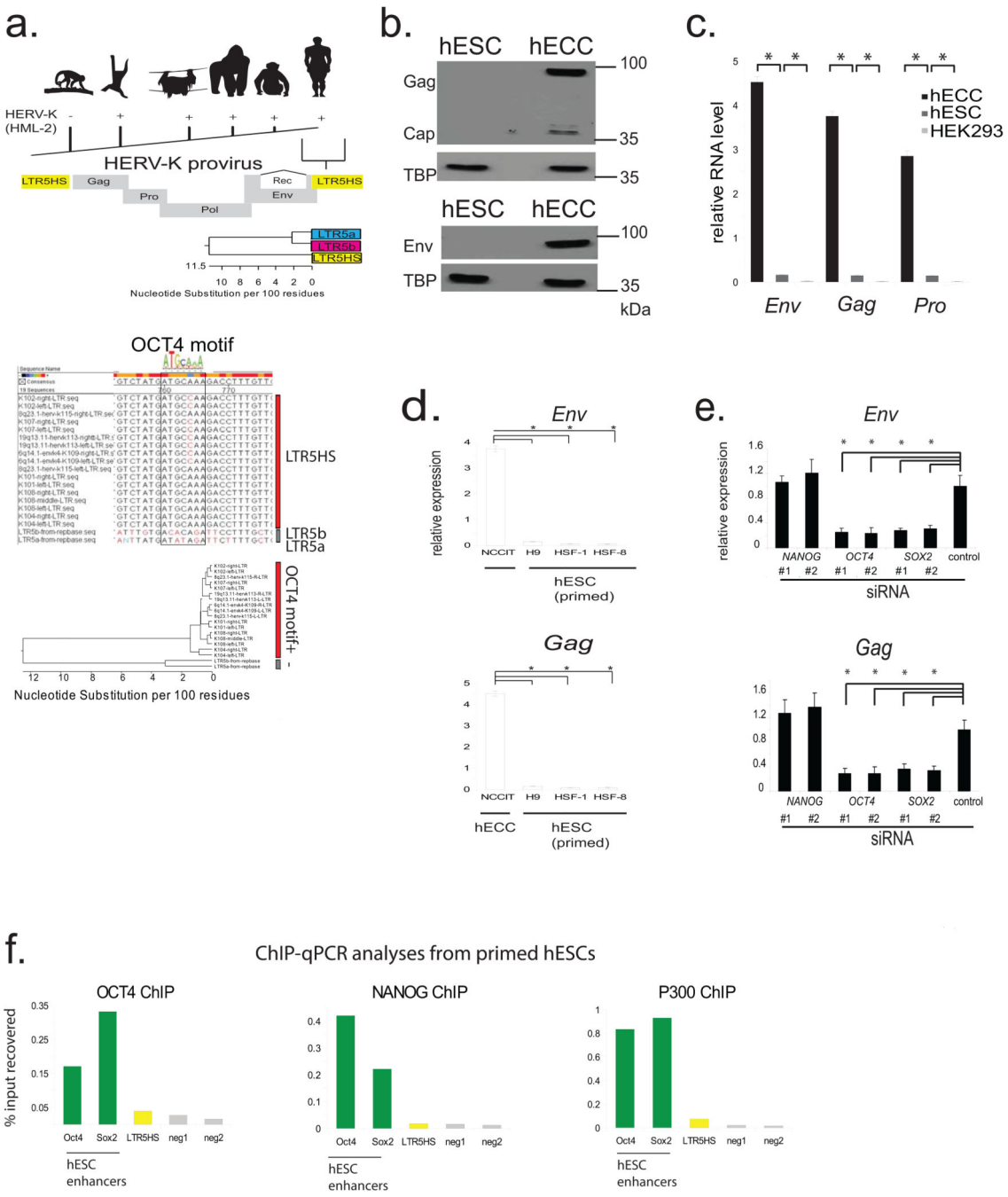
Extended Data



Extended Data Figure 1. Additional single-cell RNA-seq data analyses from pre-implantation 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 HERV-K expression (bottom) using indicated datasets.

b) HERV-H expression (RPKM) in single cells of human embryos at indicated preimplantation stages. Solid line = mean. RNA-seq from Yan, *et al*. 2013.
c) HERV-H expression (RPKM) in single cells of human blastocysts, grouped by lineage, solid line = mean. Oocyte (n=3), zygote (n=3), 2C (n=6), 4C (n=11), 8C (n=19), mor (n=16), TE (n=18), PE (n=7), EPI (n=5), p0 (n=8), p10 (n=26). RNA-seq dataset was from Yan, *et al*. 2013.
d) Genome browser snap-shot showing 100bp-PE-RNA-seq reads from ELF1 naïve hESC cells aligning at the HERV-K 108 provirus on chromsome 7.



Extended Data Figure 2. LTR5 alignments, HERV-K expression data in cell lines, and control ChIP-qPCR analyses in primed hESC (supporting Fig. 2)

a) Top: Presence of HERV-K (HML-2) sequences in Old World Primates, but absence in New World Primates. Middle: Schematic of HERV-K proviral genome; all human-specific insertions contain LTR5HS. Bottom: Phylogenetic relationship of HERV-K LTR sub-classes showing high degree of sequence similarity. Abbreviations: Gag = group specific antigen, Pro = protease, Pol= polymerase, Env= envelope, LTR= long terminal repeat, Rec = HERV-K accessory protein produced from a doubly-spliced subgenomic transcript. Bottom: ClustLW multiple sequence alignment of indicated HERV-K LTR sequences (top), region

around OCT4 motif is boxed, phylogenetic tree (bottom) indicating presence/absence of OCT4 motif.

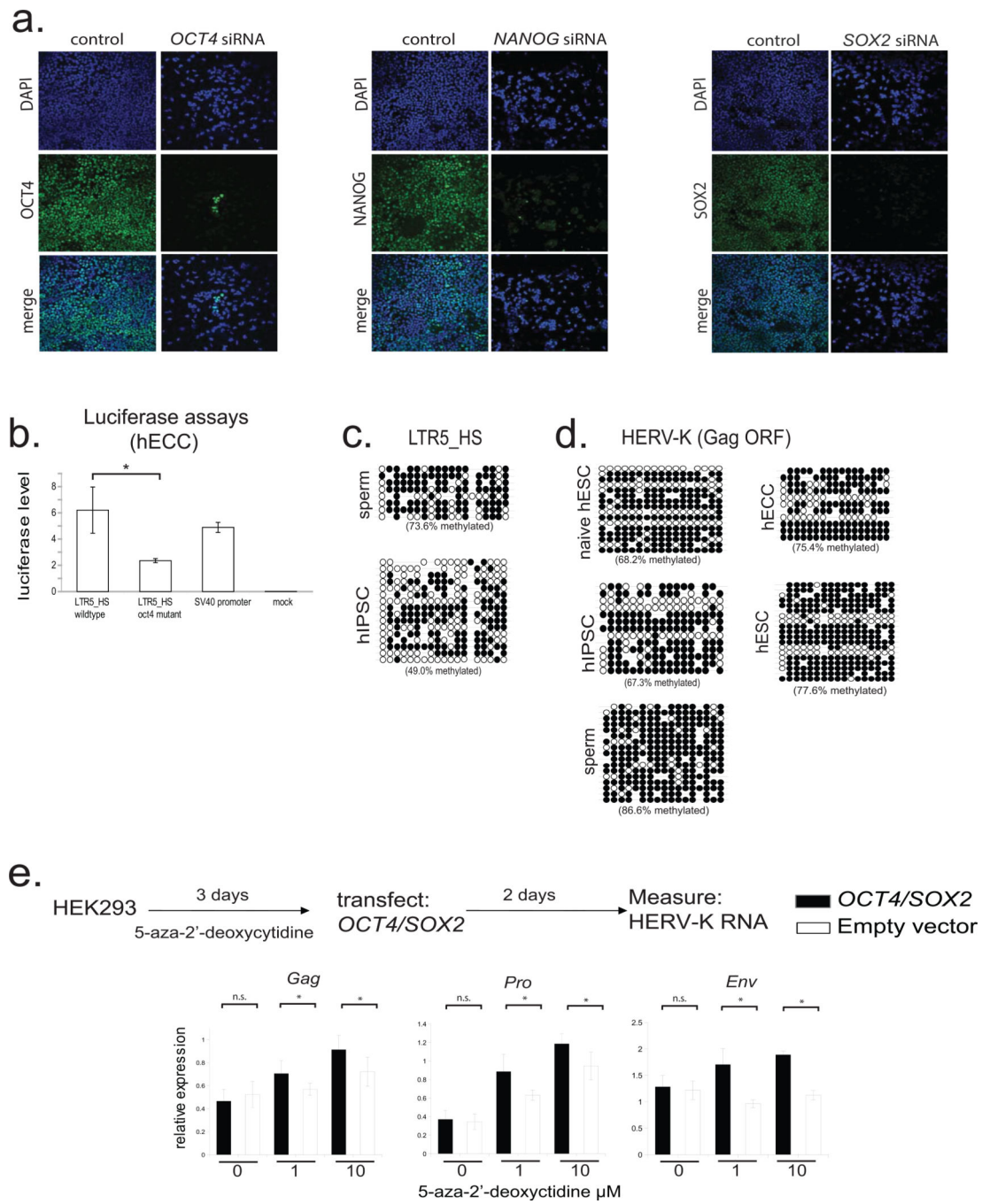
b) HERV-K protein expression in hECCs and hESCs. Protein extracts from hECCs (NCCIT) and hESC (H9) were analyzed by immunoblotting with an antibody detecting HERV-K Gag precursor and the processed Capsid (top), or glycosylated, unprocessed form of HERV-K envelope protein Env (bottom). Tata-binding protein (TBP) was used as a loading control. Shown is a representative result of three independent experiments.

c) RT-qPCR analysis of HERV-K RNA expression in hECC line NCCIT, hESC line H9, and HEK293 cells. Three distinct qPCR amplicons, corresponding to *Env*, *Gag* and *Pro* are shown. Samples were normalized to 18s rRNA levels. * denotes p-value <0.05, one-sided t-test, error bars= +/− 1 SD, n=3 biological replicates.

d) HERV-K *Gag* or *Env* expression in male hESC lines HSF-1, HSF-8, female hESC H9 and hECC line NCCIT.

e) RT-qPCR analysis of HERV-K transcripts after siRNA knockdown of NANOG, OCT4, or SOX2 in hECC (NCCIT). Signals were normalized to 18s rRNA. * denotes 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 hESCs (H9) with indicated antibodies. Signals were interrogated with primer sets for positive control regions (active hESC 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. HERV-K regulation by OCT4 and DNA methylation (supporting Fig. 2)

a) Transcription factor knockdown in hECCs (NCCIT). Cells were transfected with siRNA pools targeting indicated TFs and protein depletion was measured by immunofluoresence with respective antibodies in comparison to control, mock-transfected cells. DAPI (blue), OCT4 (green, left), NANOG (green, middle), SOX2 (green, right). Shown is one of three representative fields of view.

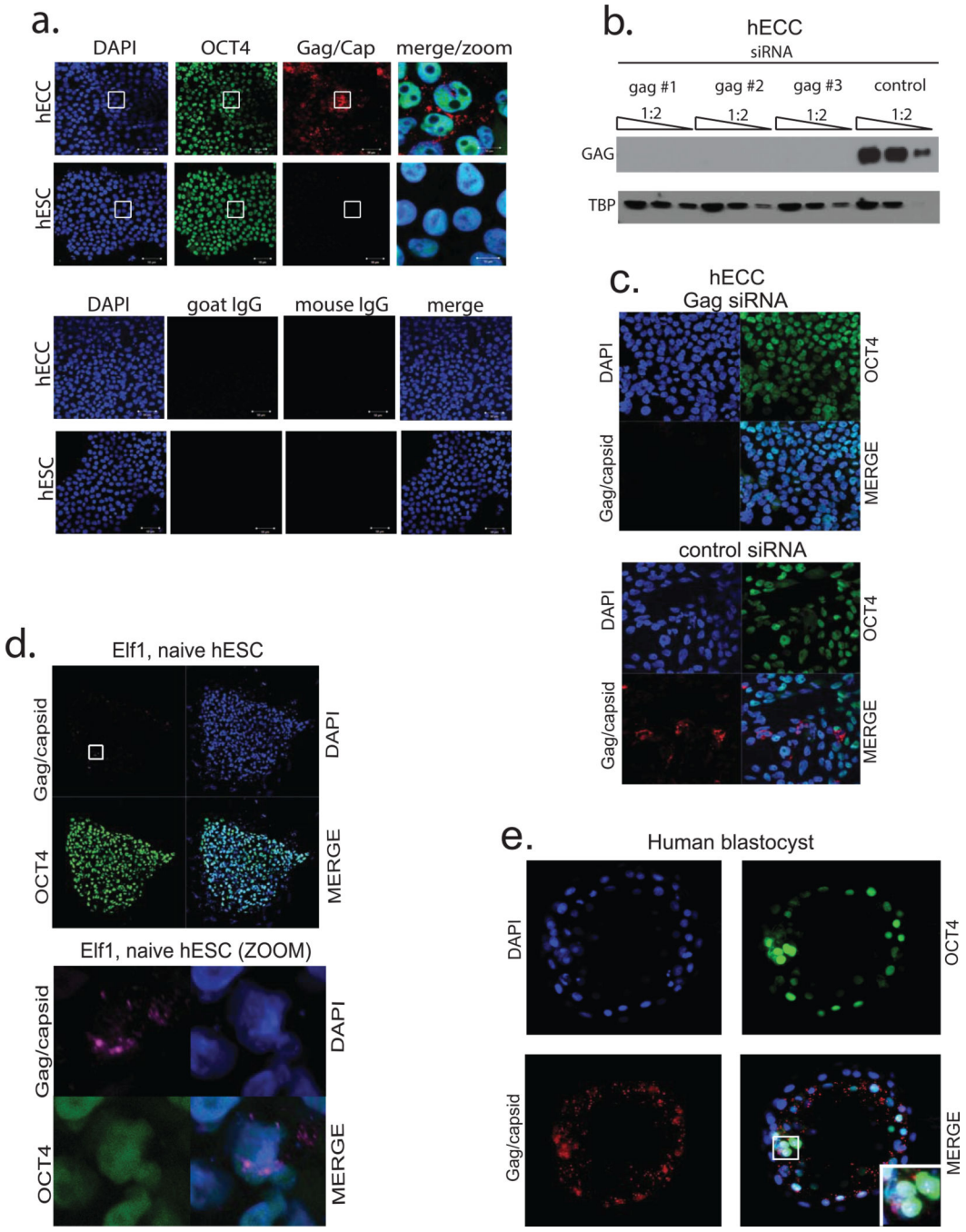
b) Dual luciferase assays with indicated reporter constructs in hECCs (NCCIT) showing that mutation of OCT4 site decreases reporter activity. N=3 biological replicates, error-bars = +/

− 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 hIPSC) analyzing consensus LTR5HS-specific amplicon as in Fig. 2f.

d) Bisulfite sequencing analysis of HERV-K 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 naïve, hESC, WT33 hIPSC, NCCIT hECC, or H9 hESC.

e) RT-qPCR analysis of HERV-K RNA levels in HEK293 cells treated with indicated concentrations of 5-aza-2'-deoxycytidine for three days, followed by transfection with OCT4/SOX2 expression constructs and RNA collection 48h after transfection. qPCR primer sets designed to 3 independent amplicons of HERV-K. *denotes p-value <0.05, one-sided t-test, n=4 biological replicates, error bars +/− 1 SD.

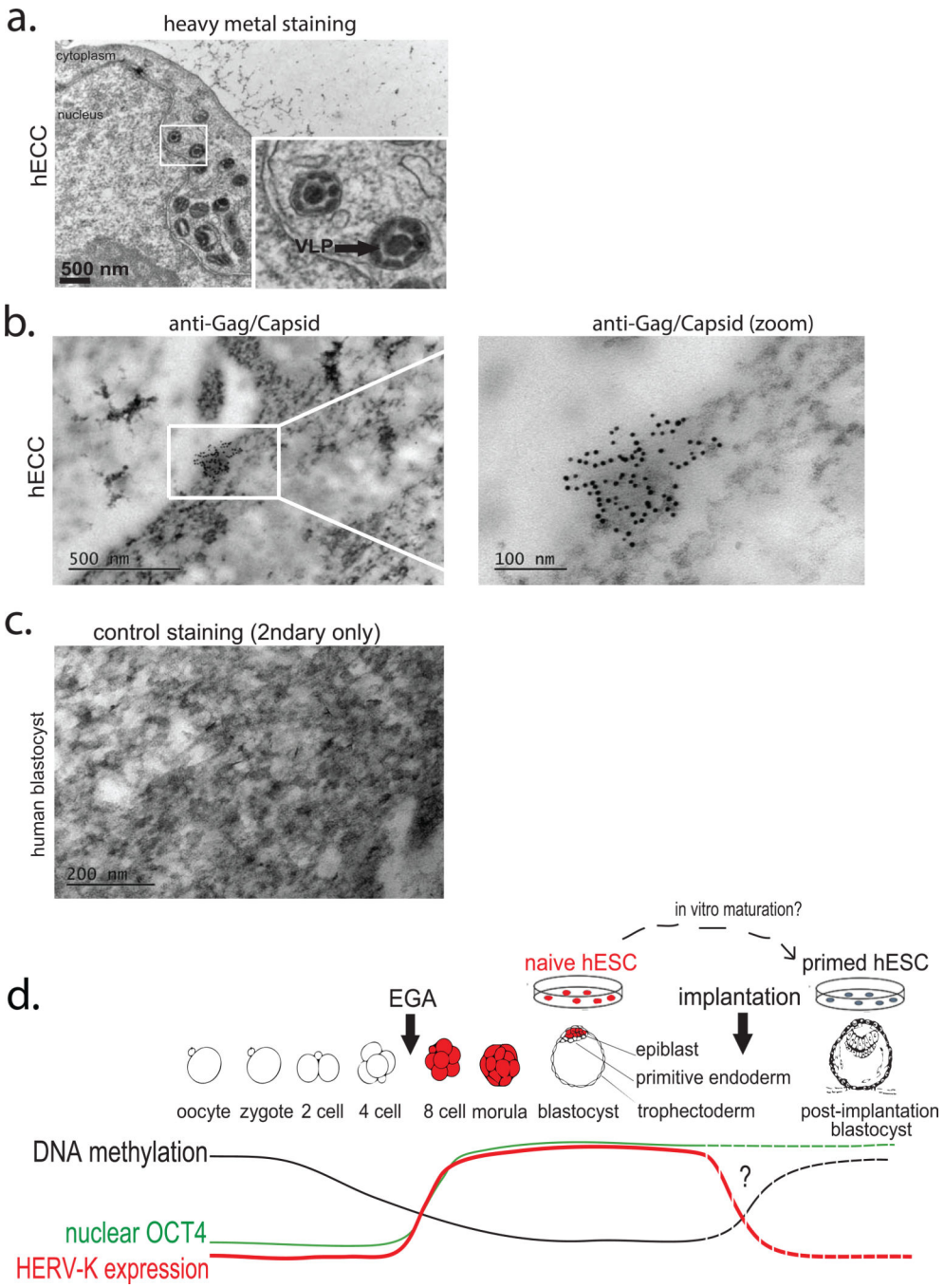


Extended Data Figure 4. HERV-K Gag/Capsid antibody validation and staining (supporting Fig. 3)

a) Immunofluorescence analysis of hECCs (NCCIT) and hESCs (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 HERV-K Gag/Capsid antibody immunoblot signal to HERV-K knockdown. hECCs were transfected with one of three independent siRNA pools targeting HERV-K Gag or with a control, non-targeting pool (synthesized against RFP) and total protein was

analyzed 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 HERV-K 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.
d) Immunoflourescence of naïve ELF1 hESC with antibodies against OCT4 (green), HERV-K Gag/Capsid (pink), DAPI in blue. Region marked with white box on left is shown with larger magnification (bottom).
e) Another representative example of immunoflourescence of human blastocyst with DAPI (blue), OCT4 (green), Gag/Capsid (red) shown (n=19 blastocysts), DPF=5–6.



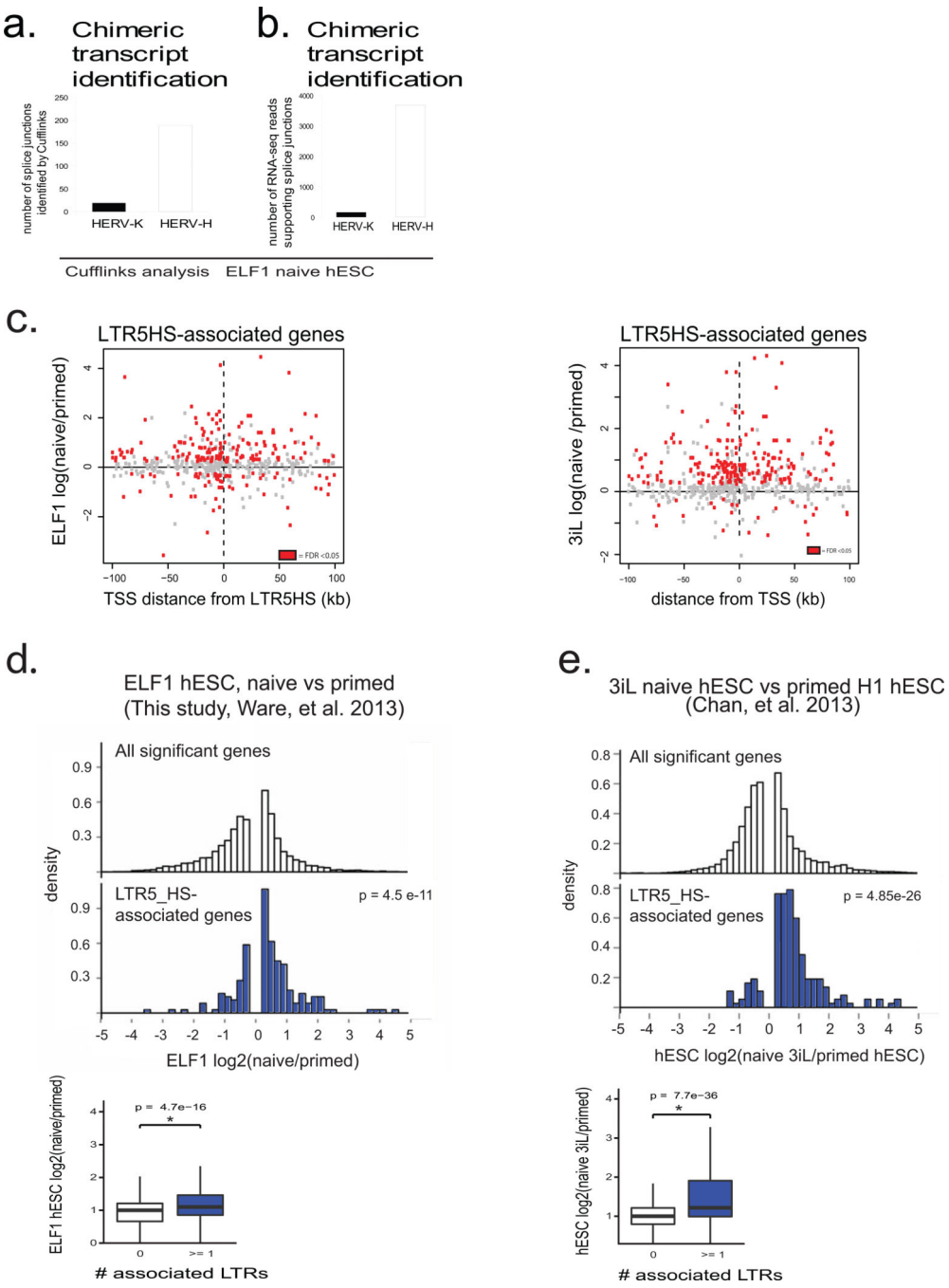
Extended Data Figure 5. TEM analyses of hECCs and control embryo staining (supporting Fig. 3)

a) TEM analysis of hECCs (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 labeling of hECC (NCCIT) with Gag/Capsid antibodies. Shown is a representative example from 2 independent experiments.

c) Secondary only control for Immuno-gold labeling of human blastocysts. Shown is a representative example from 8 fields of view.

d) Model figure summarizing HERV-K transcriptional regulation in human embryos and *in vitro* cultured pluripotent cells. Dashed lines indicate inference of OCT4, DNA methylation and HERV-K level changes at implantation from those observed between naïve and primed hESCs, in the absence of data from actual postimplantation human embryos.



Extended Data Figure 6. Correlation of HERV-K LTR5HS elements with gene expression (supporting Fig. 4)

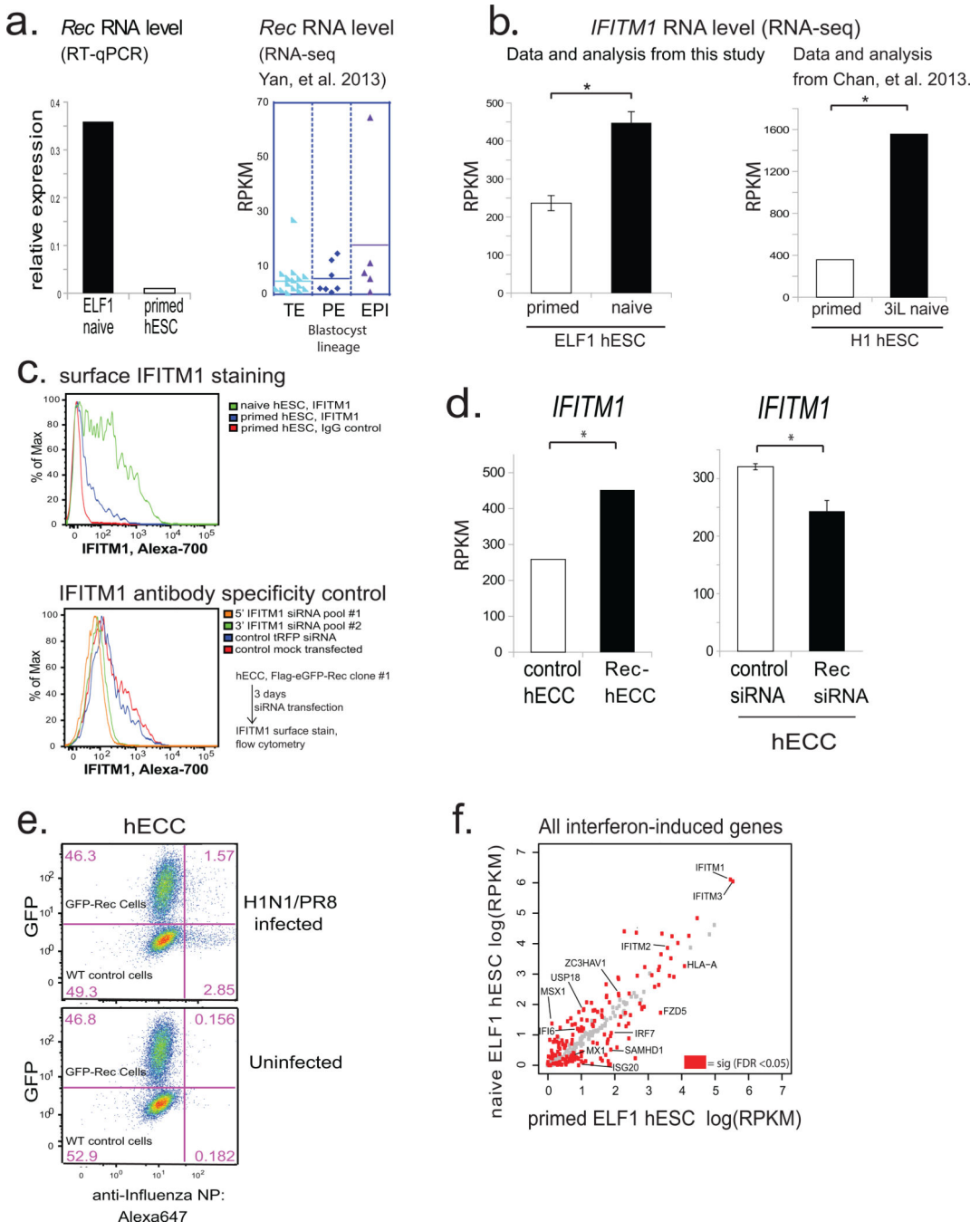
a) Number of splice junctions identified linking indicated HERV class to annotated ReqSeq genes. Analysis was done using RNA-seq dataset from ELF1 naïve hESC, n= 3 biological replicates.

b) Number of reads supporting chimeric transcripts from indicated HERV class in ELF1 naïve hESC, n =3 biological replicates.

c) Expression of LTR5_HS linked genes plotted as a function of distance to the gene's TSS. X axis: distance of TSS from the nearest LTR5_HS in kb; Y axis: fold change in expression in ELF1 naive vs primed hESC (this study, left) the 3iL versus primed H1 hESC (right, Chan *et al*. 2013).

d) Top panel: Histograms showing expression of all genes that significantly change in expression between naïve and primed ELF1 hESC (top histogram, white) or significantly changed genes that are LTR5_HS associated (bottom histogram, blue); expression values from naïve vs primed ELF1 hESC RNA-seq datasets (FDR <0.05 DESeq). Fischer's exact test gives indicated p-value indicating enrichment of LTR5HS linked genes in naïve upregulated category. Bottom panel: 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 panel: Histograms showing expression of all genes that significantly change in expression between 3iL and primed H1 hESC (top histogram, white) or significantly changed genes that are LTR5_HS associated (bottom histogram, blue); expression values from RNA-seq datasets reported by Chan, *et al*. 2013, FDR <0.05 DESeq. Fischer's exact test gives indicated p-value indicating enrichment of LTR5HS linked genes in naïve upregulated category. Bottom panel: 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 naïve hESC, and effect of Rec expression on H1N1(PR8) infection (supporting Fig. 4)

a) (left) RT-qPCR analysis of HERV-K *Rec* expression levels in ELF1 naïve hESC (n=3 biological replicates) or H9 primed hESC(one biological replicate). Normalized to 18s rRNA. Right, Rec RNA levels in indicated blastocyst lineages, solid line= mean; data from Yan, et al. 2013.

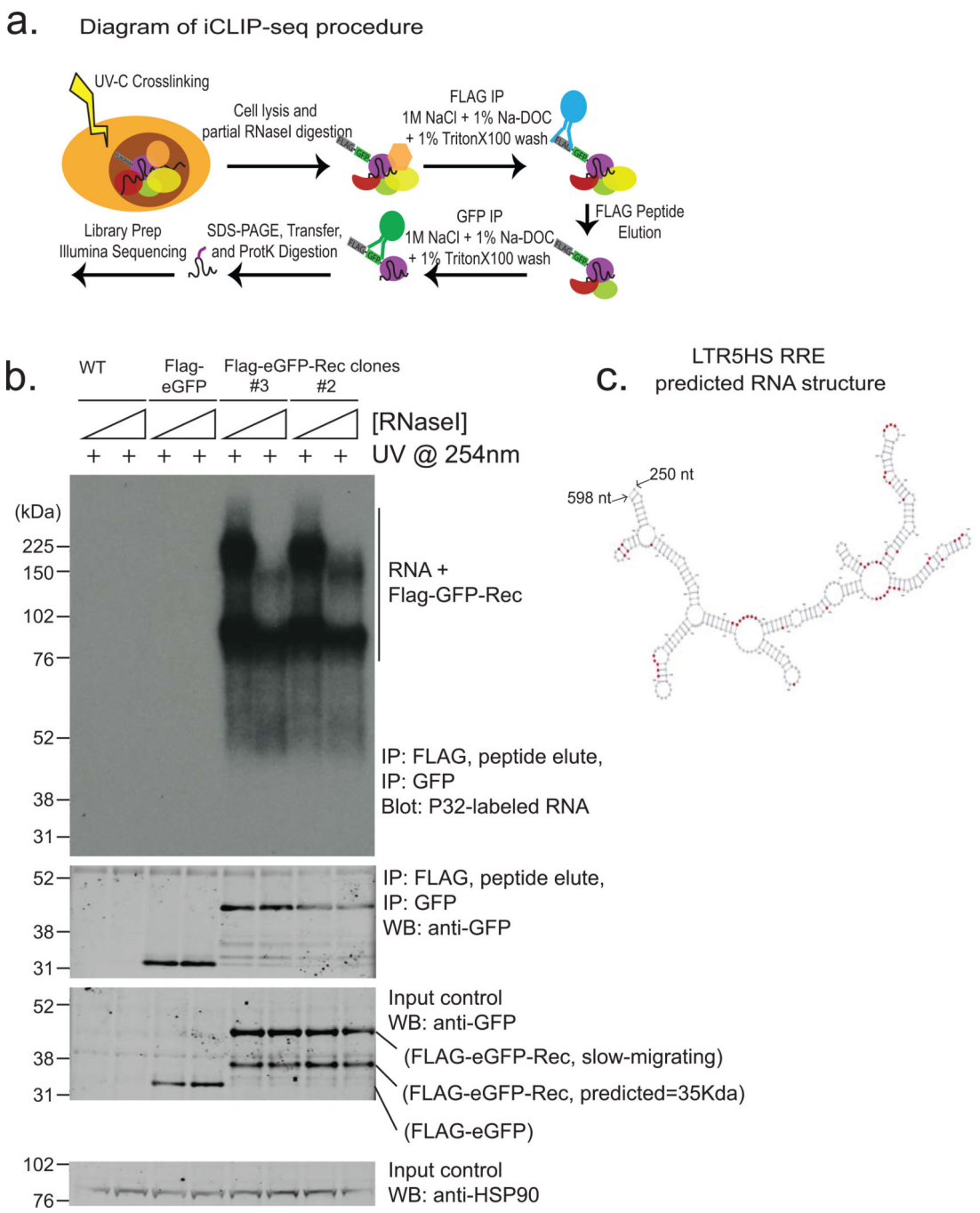
b) RNA-seq quantification of *IFITM1* RNA levels in naïve or primed ELF1 hESC (left) or 3iL hESC versus primed H1 hESC from Chan, *et al*. 2013 (right). N= 3 biological replicates for each condition, error-bars = +/− 1 S.D. * indicates significance at FDR<0.05, DESeq.

c) Flow-cytometry for surface-localized IFITM1 staining in the indicated H9 hESC or naïve ELF1 hESC (top panel) 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 panel).

d) Left: *IFITM1* expression in control hECC vs Rec-hECC (NCCIT) RNA-seq datasets. N = 2 biological replicates. Significance = FDR<0.05, DESeq. Right: *IFITM1* expression in control siRNA vs Rec siRNA-treated hECC (NCCIT) RNA-seq. N= 3 biological replicates, error-bars = +/− 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) wildtype (WT) control hECC or FLAG-GFP-Rec-hECC, clone #1. Shown is one representative example of 4 independent experiments showing a co-plating experiment in which GFP-Rec cells and wildtype 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 naïve vs primed hESC 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).



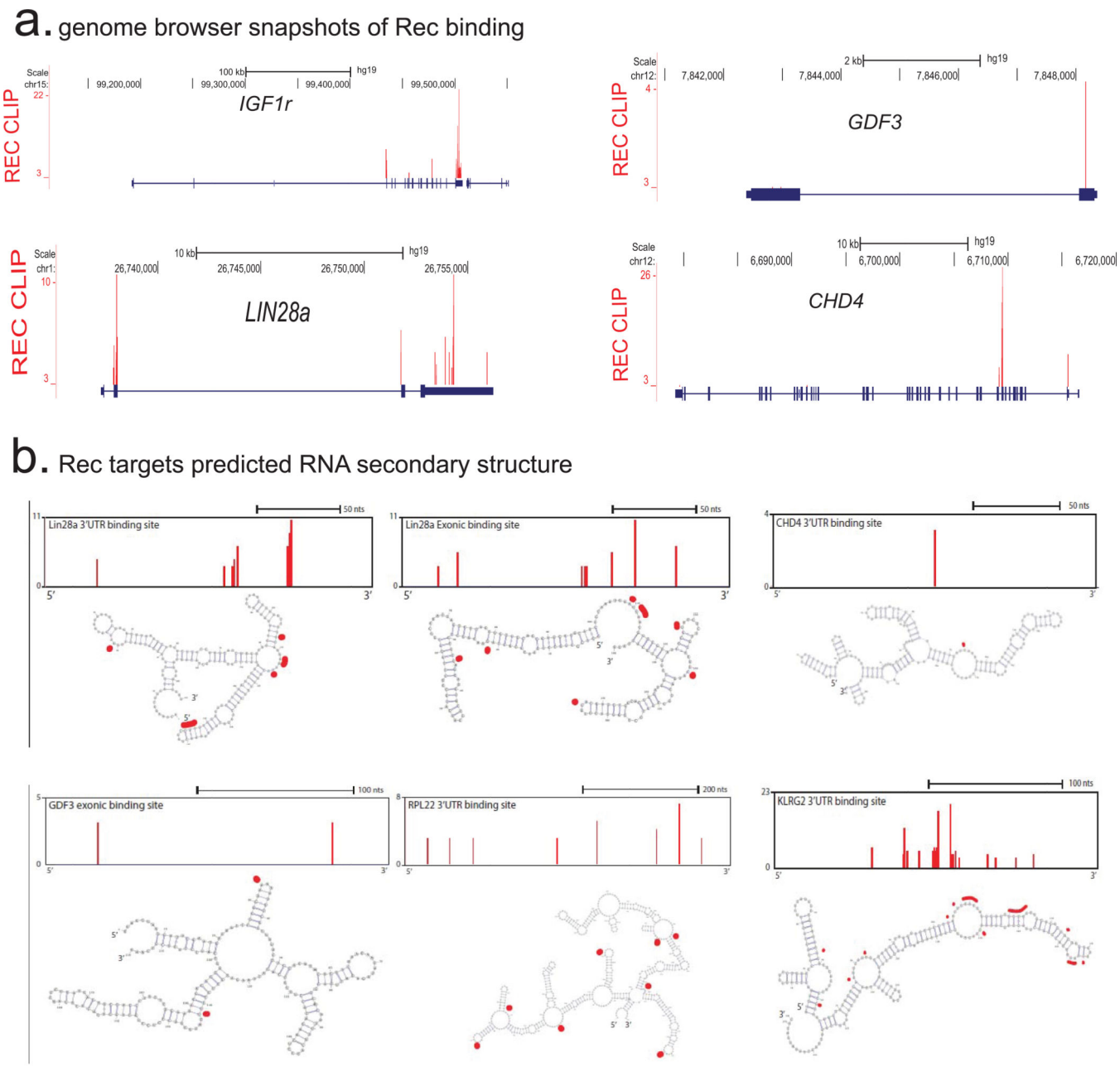
Extended Data Figure 8. iCLIP analysis of Rec-associated RNAs (supporting Fig. 4)

a) Diagram of iCLIP-seq procedure (see Methods for details). Briefly, cells are crosslinked using UV, lysed and digested with RNAse to trim RNAs. Sequential immunopurification is performed using FLAG M2, peptide elution, and GFP IP. After stringent washing, RNAs are recovered and either radiolabeled (shown in Extended Data Fig. 8b) or reverse transcribed and prepared for Illumina HTPS libraries.

b) Autoradiogram of labeled RNAs (top panel) recovered from UV-crosslinked cells using sequential Flag-eGFP IP from: wildtype hECC (lanes 1, 2), Flag-eGFP control hECC (lanes

3,4),or two independent Rec-hECC transgenic lines (lanes 5–8), separated on an 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: i) Rec-bound RNAs are resistant to even high concentrations of RNAseI, likely indicating extensive secondary RNA structures, and (ii) low/no background of contaminating RNAs in control IP from wildtype hECCs or Flag-eGFP control hECC. 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 IP fractions (middle panels). HSP90 was used as a loading control (bottom panel).

c) Computationally predicted (using mFold) secondary structure of LTR5HS sequence around the Rec-response element, (identified experimentally *in vitro* by Lower, *et al*.1997). Single nucleotide resolution Rec UV-crosslinking sites determined by iCLIP are shaded in red; (n=2 biological replicates).

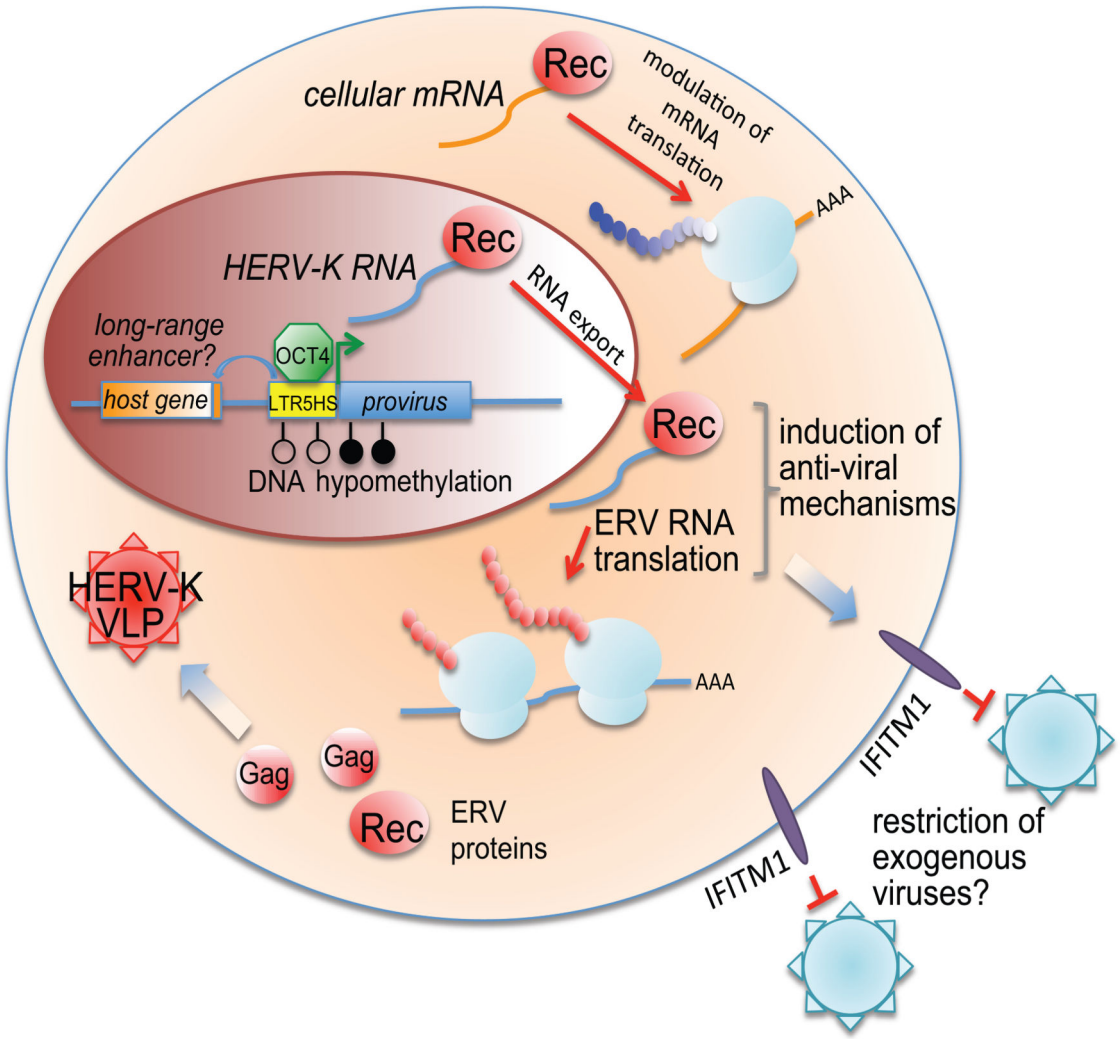


Extended Data Figure 9. Rec target mRNA analysis (supporting figure 4)

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 UV-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.

a.



Extended Data Figure 10. model figure (supporting figure 1–4)

a. Model figure summarizing HERV-K regulation and function.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank P. Bieniasz for the HERVK-con plasmid, P. Lovelace for assistance with FACS, M. Teruel for recombinant G. 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 lab for invaluable comments on the manuscript. This work was supported by equipment grant NIH S10 1S10RR02933801 and 1S10RR02678001; (NIH P01 GM099130, R01 GM112720 and CIRM RB3-05100 (J.W.), SGTP and NSF GRFP (E.J.G.), NIH DP2AI11219301 (C.B.), Smith Family Stanford Graduate Fellowship (N.L.B), CIRM RB4-05763 and NIH P50-HG007735 (H.Y.C.) and CIRM RB3-02209, March of Dimes 6-FY10-351 and U01 HL100397 (R.A.R.P.) grants.

References

1. Stoye JP. Studies of endogenous retroviruses reveal a continuing evolutionary saga. Nat. Rev. Microbiol. 2012; 10:395–406. [PubMed: 22565131]
2. Belshaw R, et al. Long-term reinfection of the human genome by endogenous retroviruses. Proc. Natl. Acad. Sci. U. S. A. 2004; 101:4894–4899. [PubMed: 15044706]
3. Barbulescu M, et al. Many human endogenous retrovirus K (HERV-K) proviruses are unique to humans. Curr. Biol. 1999; 9 861–S1.
4. Subramanian RP, Wildschutte JH, Russo C, Coffin JM. Identification, characterization, and comparative genomic distribution of the HERV-K (HML-2) group of human endogenous retroviruses. Retrovirology. 2011; 8:90. [PubMed: 22067224]
5. Herbst H, Sauter M, Mueller-Lantzsch N. Expression of human endogenous retrovirus K elements in germ cell and trophoblastic tumors. Am. J. Pathol. 1996; 149:1727–1735. [PubMed: 8909261]
6. Muster T, et al. An Endogenous Retrovirus Derived from Human Melanoma Cells. Cancer Res. 2003; 63:8735–8741. [PubMed: 14695188]
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. 2008; 82:9329–9336. [PubMed: 18632860]
8. Pace JK, Feschotte C. The evolutionary history of human DNA transposons: Evidence for intense activity in the primate lineage. Genome Res. 2007; 17:422–432. [PubMed: 17339369]
9. Kunarso G, et al. Transposable elements have rewired the core regulatory network of human embryonic stem cells. Nat. Genet. 2010; 42:631–634. [PubMed: 20526341]
10. Yan L, et al. Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. Nat. Struct. Mol. Biol. 2013; 20:1131–1139. [PubMed: 23934149]
11. Smith ZD, et al. DNA methylation dynamics of the human preimplantation embryo. Nature. 2014; 511:611–615. [PubMed: 25079558]
12. Chan Y-S, et al. Induction of a Human Pluripotent State with Distinct Regulatory Circuitry that Resembles Preimplantation Epiblast. Cell Stem Cell. 2013; 13:663–675. [PubMed: 24315441]
13. Gafni O, et al. Derivation of novel human ground state naive pluripotent stem cells. Nature. 2013; 504:282–286. [PubMed: 24172903]
14. Ware CB, et al. Derivation of naïve human embryonic stem cells. Proc. Natl. Acad. Sci. 2014; 111:4484–4489. [PubMed: 24623855]
15. Takashima Y, et al. Resetting Transcription Factor Control Circuitry toward Ground-State Pluripotency in Human. Cell. 2014; 158:1254–1269. [PubMed: 25215486]
16. Theunissen TW, et al. Systematic Identification of Culture Conditions for Induction and Maintenance of Naive Human Pluripotency. Cell Stem Cell. 2014; 15:471–487. [PubMed: 25090446]
17. Hohn O, Hanke K, Bannert N. HERV-K(HML-2), the Best Preserved Family of HERVs: Endogenization, Expression, and Implications in Health and Disease. Front. Oncol. 2013; 3
18. Shin W, et al. Human-Specific HERV-K Insertion Causes Genomic Variations in the Human Genome. PLoS ONE. 2013; 8:e60605. [PubMed: 23593260]
19. Boller K, et al. Evidence That HERV-K Is the Endogenous Retrovirus Sequence That Codes for the Human Teratocarcinoma-Derived Retrovirus HTDV. Virology. 1993; 196:349–353. [PubMed: 8356806]
20. Bieda K, Hoffmann A, Boller K. Phenotypic heterogeneity of human endogenous retrovirus particles produced by teratocarcinoma cell lines. J. Gen. Virol. 2001; 82:591–596. [PubMed: 11172100]
21. Dewannieux M, et al. Identification of an infectious progenitor for the multiple-copy HERV-K human endogenous retroelements. Genome Res. 2006; 16:1548–1556. [PubMed: 17077319]
22. Lee YN, Bieniasz PD. Reconstitution of an Infectious Human Endogenous Retrovirus. PLoS Pathog. 2007; 3:e10. [PubMed: 17257061]
23. Macfarlan TS, et al. Embryonic stem cell potency fluctuates with endogenous retrovirus activity. Nature. 2012; 487:57–63. [PubMed: 22722858]

24. Chuong EB, Rumi MAK, Soares MJ, Baker JC. Endogenous retroviruses function as species-specific enhancer elements in the placenta. Nat. Genet. 2013; 45:325–329. [PubMed: 23396136]
25. Löwer R, Tönjes RR, Korbmacher C, Kurth R, Löwer J. Identification of a Rev-related protein by analysis of spliced transcripts of the human endogenous retroviruses HTDV/HERV-K. J. Virol. 1995; 69:141–149. [PubMed: 7983704]
26. Brass AL, et al. The IFITM Proteins Mediate Cellular Resistance to Influenza A H1N1 Virus, West Nile Virus, and Dengue Virus. Cell. 2009; 139:1243–1254. [PubMed: 20064371]
27. Hanke K, et al. Staufen-1 Interacts with the Human Endogenous Retrovirus Family HERV-K(HML-2) Rec and Gag Proteins and Increases Virion Production. J. Virol. 2013; 87:11019–11030. [PubMed: 23926355]
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. 2001; 75:10359–10371. [PubMed: 11581404]
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. Nat. Cell Biol. 2013; 15:113–122. [PubMed: 23242216]
30. Lange UC, et al. Normal Germ Line Establishment in Mice Carrying a Deletion of the Ifitm/Fragilis Gene Family Cluster. Mol. Cell. Biol. 2008; 28:4688–4696. [PubMed: 18505827]

Additional references

31. Chavez SL, Meneses JJ, Nguyen HN, Kim SK, Pera RAR. 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. 2008; 17:535–546. [PubMed: 18513167]
32. Boyer LA, et al. Core Transcriptional Regulatory Circuitry in Human Embryonic Stem Cells. Cell. 2005; 122:947–956. [PubMed: 16153702]
33. Peng JC, et al. Jarid2/Jumonji Coordinates Control of PRC2 Enzymatic Activity and Target Gene Occupancy in Pluripotent Cells. Cell. 2009; 139:1290–1302. [PubMed: 20064375]
34. Myers, JWJEF, Jr. RNA Silencing. Carmichael, GG., editor. Humana Press; 2005. p. 93-196.at <http://link.springer.com/protocol/10.1385/1-59259-935-4%3A093>
35. Chavez SL, et al. Dynamic blastomere behaviour reflects human embryo ploidy by the four-cell stage. Nat. Commun. 2012; 3:1251. [PubMed: 23212380]
36. Wong C, Chen AA, Behr B, Shen S. Time-lapse microscopy and image analysis in basic and clinical embryo development research. Reprod. Biomed. Online. 2013; 26:120–129. [PubMed: 23273754]
37. Huppertz I, et al. iCLIP: protein-RNA interactions at nucleotide resolution. Methods San Diego Calif. 2014; 65:274–287.
38. Ingolia NT, Lareau LF, Weissman JS. Ribosome Profiling of Mouse Embryonic Stem Cells Reveals the Complexity and Dynamics of Mammalian Proteomes. Cell. 2011; 147:789–802. [PubMed: 22056041]
39. Flynn RA, et al. Dissecting noncoding and pathogen RNA-protein interactomes. RNA. 2015; 21:135–143. [PubMed: 25411354]
40. Rusinova I, Forster S, Yu S, Kannan A, Masse M, Cumming H, Chapman R, Hertzog PJ. INTERFEROME v2. 0: an updated database of annotated interferon-regulated genes. Nuc. Acids Res. 2013; 41

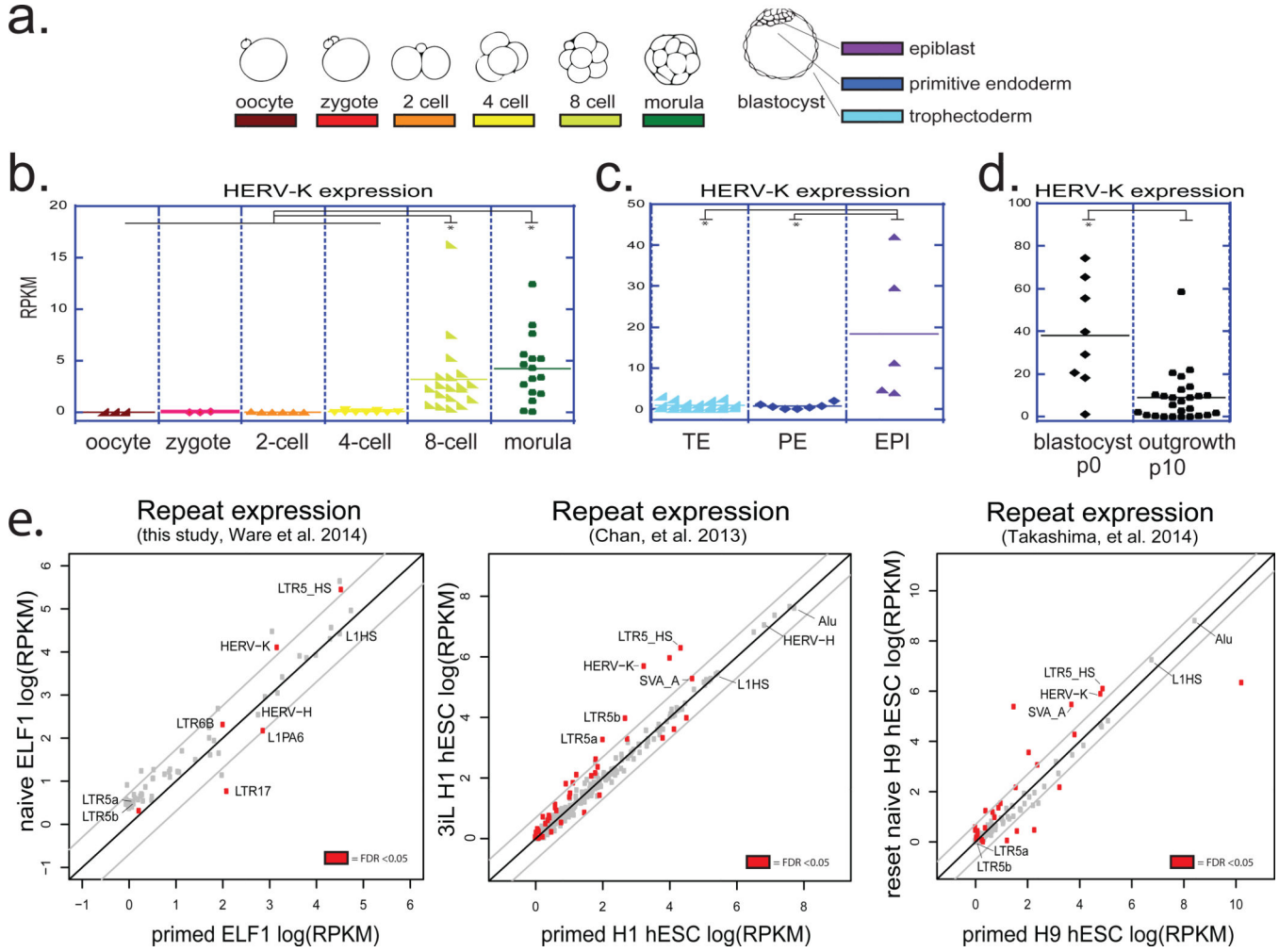


Figure 1. Transcriptional reactivation of HERV-K in human preimplantation embryos and naïve hESC

a) schematic of human preimplantation development.

b) HERV-K expression in single cells of human embryos at indicated stages. Solid line = mean. Oocyte (n=3), zygote (n=3), 2C (n=6), 4C (n=11), 8C (n=19), morula (n=16). (panels b,c,d; Yan *et al*., 2013). * denotes p-value <0.05, non-paired Wilcoxon test.

c) HERV-K expression in single cells of human blastocysts, grouped by lineage. Solid line = mean. TE (n=18), PE (n=7), EPI (n=5). Abbreviations: TE=trophectoderm, PE=primitive endoderm, EPI=epiblast.

d) HERV-K expression in single cells of blastocyst outgrowths (passage 0) or hESCs at passage 10. Solid line = mean. p0 (n=8), p10 (n=26).

e) Analysis of the repetitive transcriptomes of three, genetically matched naive/primed hESC pairs. Left: naïve/primed ELF1 hESC (this study; Ware, *et al*. 2014) (n= 3 biological replicates for both conditions). Middle: 3iL/primed H1 hESC (Chan, *et al*. 2013) (n=3 biological replicates for both conditions). Right: naïve/primed H9 hESC (Takashima, *et al*. 2014, right) (n=3 biological replicates for both conditions). Significant repeats indicated in red at FDR <0.05, DESeq.

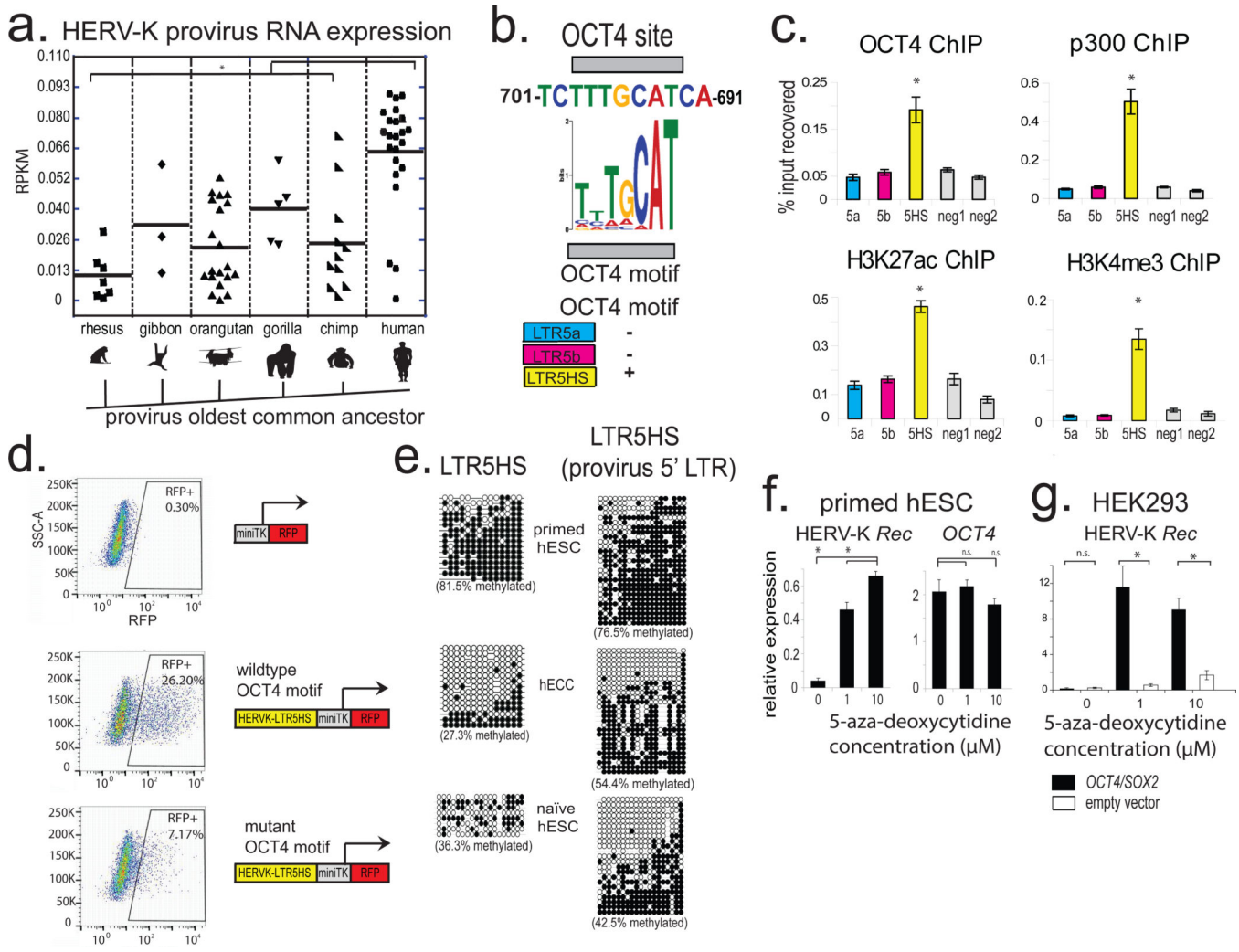


Figure 2. Transactivation by OCT4 and DNA hypomethylation of LTR5HS synergistically regulate HERV-K transcription

a) Expression of different HERV-K proviral sequences, grouped according to the oldest common ancestor, as defined by Subramanian *et al*. 2011. * denotes p-value <0.05, non-paired Wilcoxon test. Solid line = mean. RNA-seq dataset used for the analysis was from 3iL naïve H1 cells (Chan *et al*. 2013); 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 in Extended Data Fig. 2a.

c) ChIP-qPCR analyses from hECCs (NCCIT) using antibodies indicated on top of each graph. Signals were quantified using primer sets specific to LTR5HS, LTR5a, and LTR5b consensus sequences or two "negative" intergenic, non-repetitive regions. * denotes p-value <0.05 compared to negative control, one sided t-test, n=4 biological replicates, error bars are +/− 1 S.D.

d) Flow cytometry analysis of hECCs with integrated LTR5HS fluorescent reporters, either wild type (middle) or with OCT4 motif mutation (bottom). RFP positive population was

gated using side-scatter area (SSC-A) and cells with integrated negative control reporter (top). 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 hECCs (NCCIT) or hESCs (H9) or naïve hESC (ELF1). Filled circles depict modified cytosines, empty circles depict unmodified cytosines. hECC (NCCIT) and naïve hESC (ELF1) are less methylated than hESC (H9), $p < 0.05$, non-paired Wilcoxon test.

f) RT-qPCR analysis of hESC (H9) treated with indicated concentrations of 5-aza-2'-deoxycytidine for 24 hours. *denotes p-value <0.05, one-sided t-test, n=3 biological replicates, error bars +/− 1 SD.

g) RT-qPCR analysis of HERV-K Rec RNA levels in HEK293 cells treated with indicated concentrations of 5-aza-2'-deoxycytidine, followed by transfection with OCT4/SOX2 expression constructs. *denotes p-value <0.05, one-sided t-test, n=4 biological replicates, error bars +/− 1 S.D.

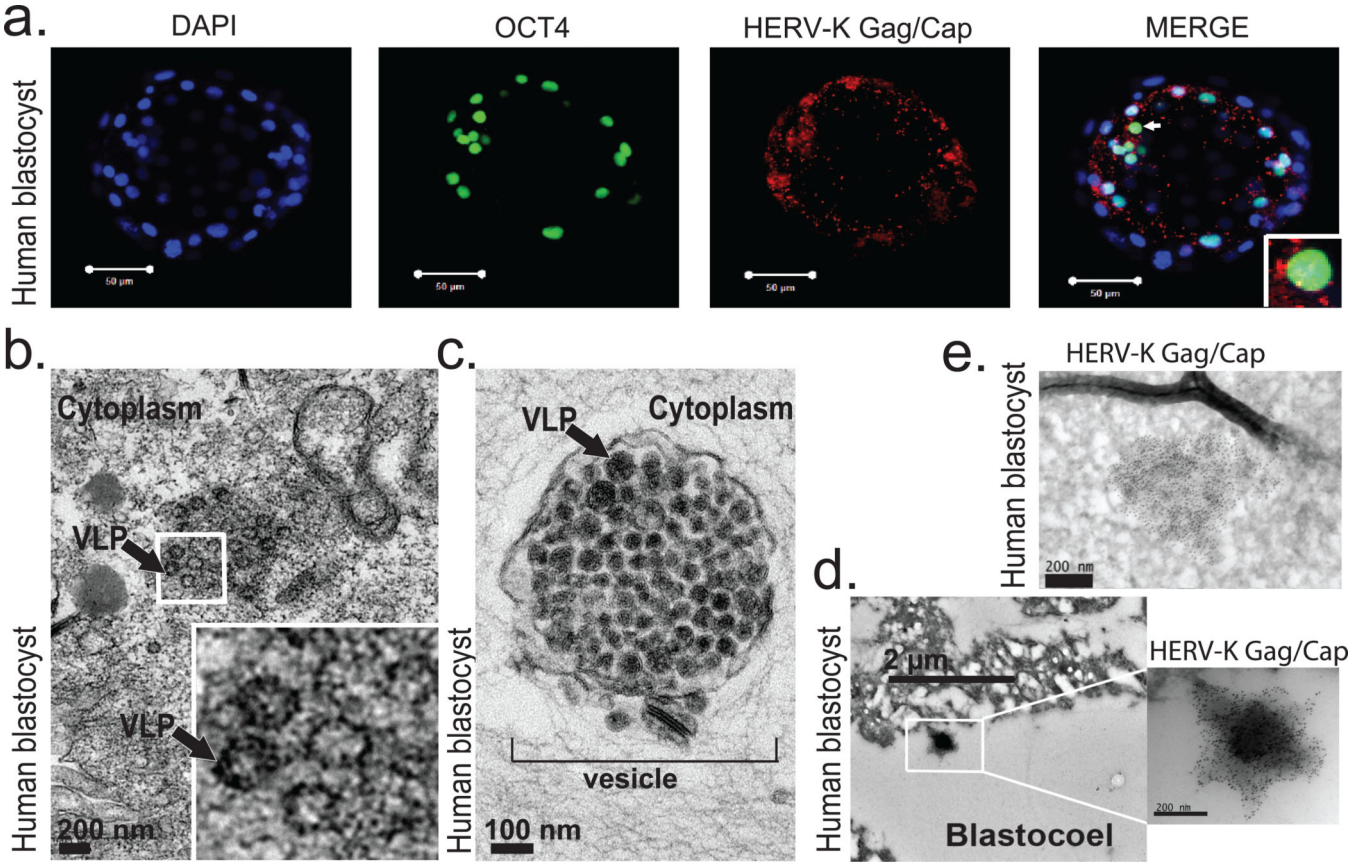


Figure 3. Human blastocysts contain HERV-K proteins and viral-like particles

a) Immunofluorescence of human blastocysts (days post fertilization, DPF =5–6) stained with DAPI (blue), OCT4 antibody (green), and HERV-K Gag/Capsid antibody (Red). Images show a representative example (n=19 embryos). Scale bar = 50 microns, 1 micron confocal z-slice. White arrow points to an OCT4+ cell, surrounded by cytoplasmic Gag/Capsid, which is shown with higher magnification in an inset.

b) Heavy metal staining transmission electron microscopy (TEM) of human blastocyst, arrow denotes putative VLP (found in n=2/3 blastocysts, DPF=5–6). Higher magnification of indicated region shown in inset. Scale bar = 200nm.

c) Heavy metal staining TEM of human blastocyst, arrow denotes 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 labeled particles in 2 embryos.

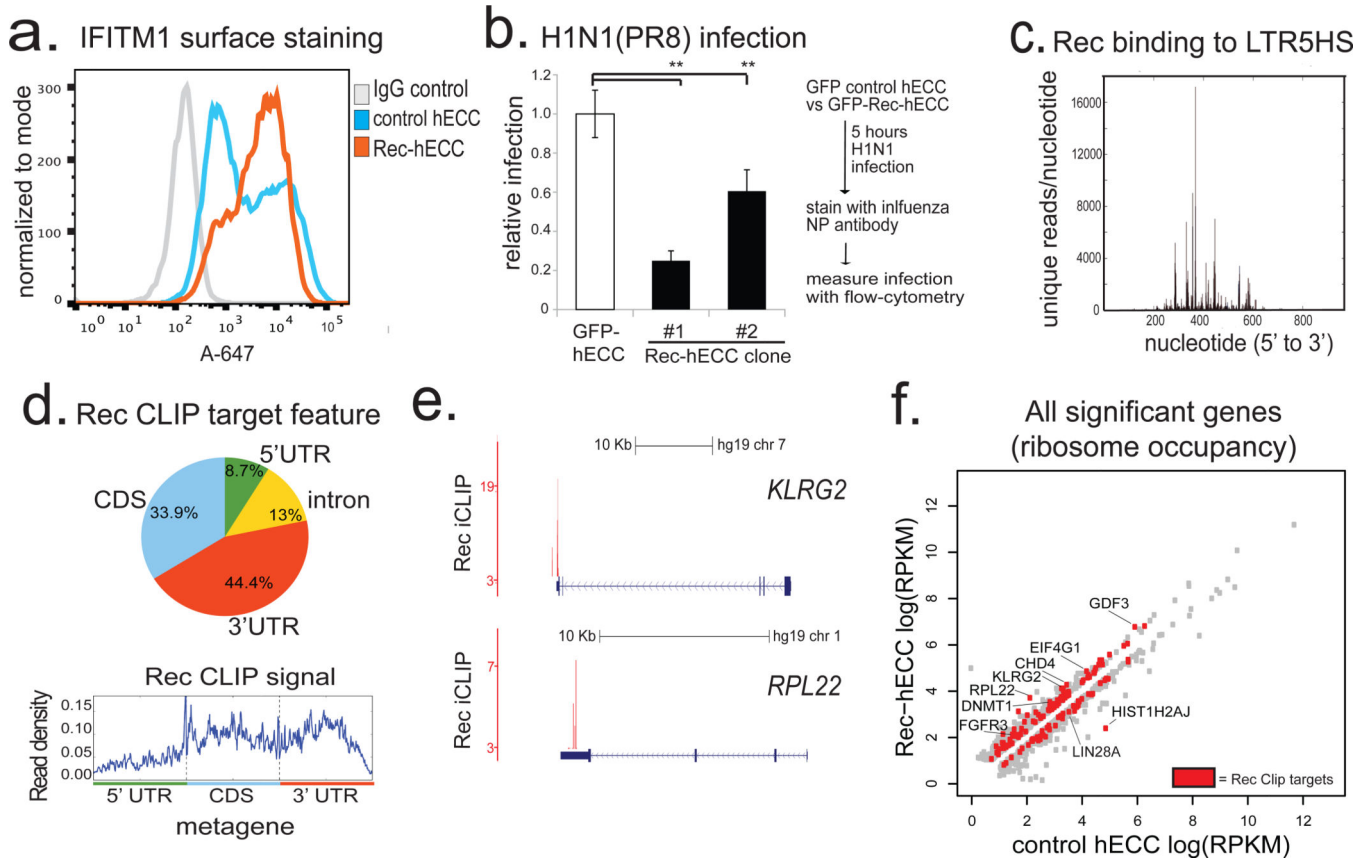


Figure 4. HERV-K accessory protein Rec upregulates viral restriction pathway and engages cellular mRNAs

a) Flow cytometry histograms of IFITM1 surface staining in control hECC or Rec-hECC (NCCIT) cells, histogram of negative control cells stained with isotype IgG+Alexa-647 secondary 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-hECC (NCCIT). Control cells were set as 100%, shown is aggregate results from 2 independent experiments, n=8 total biological replicates for each condition. Error bars are +/− 1 S.D. ** denotes 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.

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 wildtype hECC cells vs Rec-hECC (NCCIT), n=4 biological replicates. Rec iCLIP targets are colored in red