Antiviral activity of a human placental protein of retroviral origin

John A. Frank¹, Manvendra Singh¹, Harrison B. Cullen¹, Raphael A. Kirou¹, Carolyn B. Coyne², Cédric Feschotte¹

Affiliations
¹Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY, USA
²Department of Pediatrics, University of Pittsburgh, Pittsburgh, PA, USA

Summary
Viruses circulating in wild and domestic animals pose a constant threat to human health¹. Identifying human genetic factors that protect against zoonotic infections is a health priority. The RD-114 and Type-D retrovirus (RDR) interference group includes infectious viruses that circulate in domestic cats and various Old World monkeys (OWM), and utilize ASCT2 as a common target cell receptor². While human ASCT2 can mediate RDR infection in cell culture, it is unknown whether humans and other hominoids encode factors that restrict RDR infection in nature²,³. Here we test the hypothesis that Suppressyn, a truncated envelope protein that binds ASCT2 and is derived from a human endogenous retrovirus⁴,⁵, restricts RDR infection. Transcriptomics and regulatory genomics reveal that Suppressyn expression initiates in the preimplantation embryo. Loss and gain of function experiments in cell culture show Suppressyn expression is necessary and
sufficient to restrict RDR infection. Evolutionary analyses show Suppressyn was acquired in the genome of a common ancestor of hominoids and OWMs, but preserved by natural selection only in hominoids. Restriction assays using modern primate orthologs and reconstructed ancestral genes indicate that Suppressyn antiviral activity has been conserved in hominoids, but lost in most OWM. Thus in humans and other hominoids, Suppressyn acts as a restriction factor against retroviruses with zoonotic capacity. Transcriptomics data predict that other virus-derived proteins with potential antiviral activity lay hidden in the human genome.

Main

Viral zoonosis poses a constant threat to human health and has led to devastating epidemics such as those caused by Influenza\(^6\), HIV\(^7\), Ebola\(^8\), and SARS coronaviruses\(^9,10\). Some zoonotic viruses have gained access to new host species by acquiring envelope (env) glycoproteins that mediate target-cell entry by binding to host cell surface receptors\(^6,11\). Notably, beta- and gamma-retroviruses have captured the so-called RDR env, which has enabled them to infect and transfer across diverse mammalian hosts\(^11,12\). For instance, the feline leukemia virus RD-114, an infectious endogenous retrovirus from the domestic cat, emerged from the Felis catus endogenous virus by acquiring the RDR env from the Baboon endogenous virus\(^11\). Because all RDR env bind to the highly conserved and broadly expressed amino acid transporter ASCT2 (also known as SLC1A5) to mediate cell entry, RDR env-mediated infection poses a zoonotic threat to humans\(^2,13,14\). Thus, it is critical to assess whether humans are equipped with mechanisms to protect against RDR infection.
Previous reports have shown that endogenous retroviral env are capable of restricting retroviral infection by a mechanism of receptor interference in multiple vertebrates, including chicken, mouse, sheep, and cat\textsuperscript{15}. Some of these env-derived restriction factors acquired truncating mutations resulting in loss of their C-terminal membrane-anchoring transmembrane domain\textsuperscript{15}, but retention of receptor-binding activity. Suppressyn (SUPYN) is a protein that is derived from a human endogenous retroviral env, lacks a transmembrane domain, and is expressed throughout human placenta development\textsuperscript{4,5}. Previous \textit{in vitro} studies have shown that SUPYN, like Syncytin-1 (SYN1), binds ASCT2 and thereby modulates the fusogenic activity of SYN1 during placenta development\textsuperscript{4,5}. Here we investigate whether SUPYN confers resistance to RDR infection.

SUPYN embryonic expression is driven by pluripotency and placentation regulatory factors.

To obtain a detailed view of SUPYN expression and regulation during human embryonic development, we analyzed publicly available scRNA-seq, ATAC-seq, DNAse-seq and ChIP-seq datasets generated from preimplantation embryos and human embryonic stem cells (hESC) (Supplementary Table 1). SUPYN mRNA appears after the onset of embryonic genome activation at the eight-cell stage and peaks in morula (Fig 1a). By blastula formation, SUPYN expression persists in the inner cell mass, epiblast, ESCs, and in the trophectoderm, which gives rise to the placenta (Fig 1a, Extended Data 1a). Consistent with this expression pattern, the SUPYN locus is marked by open chromatin from 8-cell to blastocyst stages (Extended Data 1b). In hESCs, the SUPYN promoter
region is marked by H3K4me1 and H3K27ac histone modifications characteristic of
‘active’ chromatin, and bound by core pluripotency (OCT4, NANOG, KLF4, SMAD1) and
self-renewal (SRF, OTX2) transcription factors (Fig 1b). Together, these data indicate
SUPYN is robustly expressed throughout early embryonic development and likely under
the control of pluripotency factors. By contrast, we found little evidence for SYN1
expression in preimplantation embryos and hESCs (Fig 1a, Extended Data 1a).

To examine SUPYN expression throughout placentation, we interrogated RNA-seq and
ChIP-seq datasets generated from in vitro trophoblast (TB) differentiation models and
placenta explants isolated at multiple stages of pregnancy (Supplementary Table 1).
During hESC to TB differentiation, we observed that pluripotency factors NANOG and
OCT4 occupying the SUPYN promoter region are replaced by trophoblast-specific
transcription factors TFAP2A and GATA3 (Fig 1b). SUPYN expression likely persists
through the TB differentiation process because SUPYN transcripts and active chromatin
marks (H3K27ac, H3K4me3, H3K9ac) are maintained across all analyzed TB cell
lineages (Fig 1b). By contrast, expression of other envelope-derived genes SYN1, SYN2,
and ERVV1/V2 is only detectable in differentiated trophoblasts (Extended Data 1c). We
next mined scRNA-seq data generated from placenta at multiple developmental stages
to examine the cell-type specific expression of SUPYN (Supplementary Table 1). After
classifying cell clusters based on known markers (Fig 1c, d, Extended Data 2a, b, c),
we found SUPYN and ASCT2 expression specifically in the TB lineage (Fig. 1e, f;
Extended Data 2c, d). Consistent with previous reports5, SUPYN expression was
relatively high in cytotrophoblasts (CTB) and extra-villous trophoblasts (EVTB), but also
detectable in syncytiotrophoblasts (STB) (Fig 1e, Extended Data 2c, d). SUPYN expression in EVTB was maintained throughout placental development (Fig 1f). SYN1 expression appears restricted to CTB and STB lineages (Fig 1e, Extended Data 2c, d), as previously reported\textsuperscript{5,16-18}. To confirm these transcriptomic observations, we performed immunostaining of second (21w gestation) and third (31w gestation) trimester placenta with SUPYN antibody. These stains show SUPYN is widely expressed in STB, and perhaps CTB within the lumen of 2\textsuperscript{nd} trimester placental villi (Fig 1g, Extended Data 3). Together these analyses indicate SUPYN is expressed throughout human fetal development and has a broader expression pattern than SYN1.

**SUPYN confers resistance to RD114env-mediated infection**

SUPYN expression during early embryonic and placental development, which coincides with that of ASCT2 (Fig 1a), suggests SUPYN may interact with ASCT2 throughout fetal development and confer RDR resistance to the developing embryo. To begin testing this hypothesis, we first examined whether human placenta-derived cell lines Jar and JEG3, and hESC H1 cells are resistant to RDR env-mediated infection. We generated HIV-GFP viral particles pseudotyped with either the feline RD114env (HIV-RD114env) or the glycoprotein G of vesicular stomatitis virus (HIV-VSVg, which uses the LDL receptor\textsuperscript{19}) to monitor the level of infection in cell culture based on GFP expression (Fig 2a, Extended Data 4). These experiments revealed that Jar, JEG3, and H1 cells were susceptible to HIV-VSVg, as previously reported\textsuperscript{21-25}, but resistant to HIV-RD114env infection (Fig 2b, c). Concurrently infected 293T cells, which do not express SUPYN (Extended Data Fig 1a), were similarly susceptible to infection by HIV-RD114env and HIV-VSVg (Fig 2b, c).
To test whether SUPYN contributes to the HIV-RD114env resistance phenotype, we repeated these infection experiments in Jar cells engineered to stably express short hairpin RNAs depleting ~80% of SUPYN or SYN1 mRNAs (Extended Data Fig 5a). Depletion of SUPYN in Jar cells (shSUPYN) resulted in a significant increase in susceptibility to HIV-RD114env infection (Fig 2d), but did not affect infection by HIV-VSVg (Fig 2d). Importantly, SYN1 depletion from Jar cells did not increase susceptibility to HIV-RD114env infection (Fig 2d). These results suggest that SUPYN expression contributes to the RD114 resistance phenotype of Jar placental cells.

To account for possible off-target effects of SUPYN-targeting siRNAs, we transfected Jar-shSUPYN cells with two siRNA-resistant, HA-tagged SUPYN rescue constructs and examined their susceptibility to HIV-RD114env infection. Briefly, the siRNA-targeted SUPYN signal peptide sequence was replaced with a luciferase (SUPYN-lucSP) or modified signal peptide sequence (SUPYN-rescSP) that disrupts siRNA-binding but retains codon-identity. Transfection with either SUPYN-rescSP or SUPYN-lucSP restored a significant level of resistance to HIV-RD114env infection (Fig 2e). Western Blot analysis of transfected cell lysates showed SUPYN-rescSP was more abundantly expressed than SUPYN-lucSP (Fig 2f), which may account for the stronger HIV-RD114env resistance conferred by SUPYN-rescSP (Fig 2e). These results corroborate the notion that SUPYN restricts RD114env-mediated infection in Jar cells.
To test if SUPYN expression alone is sufficient to confer protection against RD114env-mediated infection, we transfected 293T cells with a SUPYN overexpression construct and subsequently infected with HIV-RD114env and HIV-VSVg respectively. As a positive control, we also transfected 293T cells with a RD114env overexpression construct, which is predicted to confer resistance to HIV-RD114env, but not to HIV-VSVg. Expression of either SUPYN or RD114env resulted in ~80% reduction in the level of HIV-RD114env infection (Fig 2g, Extended Data Fig 6a), but had no significant effect on HIV-VSVg infectivity (Extended Data Fig 6b). Taken together, our knockdown and overexpression experiments indicate SUPYN expression is both necessary and sufficient to confer resistance to RD114env-mediated infection.

SUPYN restricts RDR infection likely through receptor interference

Our RD114env-specific resistance phenotype (Extended Data Fig 6a, b) strongly suggests SUPYN functions by receptor interference. If so, this protective effect should extend to infection mediated by other RDR env$^{2,26,27}$ since they all use ASCT2 as receptor. To test this prediction, we generated HIV-GFP reporter virions pseudotyped with Squirrel Monkey Retrovirus (SMRV) env (HIV-SMRVenv)$^2$ (Fig 2a) and infected 293T cells previously transfected with SUPYN, SMRVenv (as a positive control) or an empty vector. Cells expressing SUPYN or SMRVenv showed an ~80% reduction of HIV-SMRVenv infected cells (Fig 2h). Thus, SUPYN can restrict infection mediated by multiple RDRenv.

Another prediction of RDR restriction via receptor interference is that it should be a property of env proteins recognizing ASCT2, such as SUPYN, but not those binding other
cellular receptors. Consistent with this prediction, expressing HA-tagged env from amphotrophic murine leukemia virus (aMLV) or human endogenous retrovirus H, neither of which are expected to interact with ASCT2, had no effect on HIV-RD114env nor HIV-VSVg infection in 293T cells, while HA-tagged SUPYN strongly restricted HIV-RD114env (Extended Data Fig 6a, b). Importantly, all tested env proteins were expressed at comparable levels (Extended Data Fig 6c). Furthermore, we observed that SUPYN overexpression did not significantly impact ASCT2 expression levels in 293T cells (Extended Data Fig 6c). This result suggests that if SUPYN acts by receptor interference, its interaction with ASCT2 does not result in receptor degradation, which is consistent with some instances of receptor interference. In agreement with previous observations, we noted that SUPYN knockdown in Jar cells resulted in the specific loss of a putative non-glycosylated isoform of ASCT2 (Extended Data 5b). We speculate that the presence of SUPYN-dependent non-glycosylated ASCT2 may be the result of SUPYN sterically interfering with the glycosylation machinery within the secretory pathway. It has not been reported whether ASCT2 glycosylation impacts RDR env-mediated infection in human cells, but it is known that receptor glycosylation may interfere with RDR infection in mouse and hamster cells. Collectively, these observations converge on the model that SUPYN restricts against RDR infection through receptor interference (Fig 2i).

SUPYN emerged in a catarrhine ancestor and evolved under functional constraint in hominoids

Little is known about the evolutionary origin of SUPYN. The gene was originally identified as derived from a copy of the HERV-Fb family of endogenous retroviruses (also known
as HERVH48\(^{36}\)) located on human chromosome 21q22.3 with an ortholog in chimpanzee\(^{4}\). Using comparative genomics (see Methods), we found that this HERVH48 element is shared at orthologous position across the genomes of all available hominoids (i.e. apes) and most Old World monkeys (OWM), but precisely lacking in New World monkeys and prosimians (Fig 3a, Extended Data 7, 8). Thus, the provirus copy that gave rise to SUPYN inserted in the common ancestor of catarrhine primates ~20-38 million years ago\(^{37}\) (Fig 3a).

All primates with HERVH48 orthologs also share a nonsense mutation which would have truncated the ancestral env protein at site 185 in the common ancestor of catarrhine primates (Fig 3a, Extended Data 8). Hominoids share an additional nonsense mutation further truncating the protein to the 160-aa SUPYN-encoding ORF annotated in the human reference genome (Fig 3a, Extended Data 8). The SUPYN ORF is almost perfectly conserved in length across hominoids, but not in OWM where some species display further frameshifting and truncating mutations (Extended Data 9), suggesting SUPYN may have evolved under different evolutionary regimes in hominoids and OWMs.

To test this idea, we analyzed the ratio \(\omega\) of nonsynonymous (dN) to synonymous (dS) substitution rates using codeml\(^{38}\), which provides a measure of selective constraint acting on codons. Log-likelihood ratio tests comparing models of neutral evolution with selection indicate SUPYN evolved under purifying selection in hominoids \((\omega = 0.38; p = 1.47E-02)\), but did not depart from neutral evolution in OWMs \((\omega = 1.44; p = 0.29)\) (Fig 3a). For comparison, we performed the same type of analysis for SYN1 and SYN2, two primate-specific env-derived genes thought to be involved in placentation\(^{17,39,40}\). Consistent with
previous reports\textsuperscript{41,42}, we found that both $\text{SYN1}$ ($\omega = 0.64; p = 1.80\text{E}-02$) and $\text{SYN2}$ ($\omega = 0.29; p = 3.22\text{E}-08$) evolved under purifying selection during hominoid evolution (Fig. 3a).

In OWMs, $\text{SYN2}$ also evolved under purifying selection ($\omega = 0.22, p = 2.78\text{E}-08$), while $\text{SYN1}$ was lost through an ancestral deletion\textsuperscript{14} (Fig. 3a). These results suggest that the level of functional constraint acting on $\text{SUPYN}$ during hominoid evolution is comparable to that seen on other $\text{env}$-derived genes with placental function.

**SUPYN antiviral activity is conserved across hominoid primates**

To assess whether primate $\text{SUPYN}$ orthologs have antiviral activity, we generated and transfected 293T cells with HA-tagged overexpression constructs for the orthologous $\text{SUPYN}$ sequences of chimpanzee, siamang, African green monkey, pigtailed macaque, crab-eating macaque, rhesus macaque, and olive baboon and challenged these cells with HIV-RD114$\text{env}$ virions. Both chimpanzee and siamang $\text{SUPYN}$ proteins displayed antiviral activity with potency comparable to or greater than human $\text{SUPYN}$, respectively (Fig 3b). By contrast, only one (African green monkey) of the five OWM orthologous $\text{SUPYN}$ proteins exhibited a modest but significant level of antiviral activity (Fig 3b, c).

The lack of restriction activity for some of the OWM proteins may be attributed to their relatively low expression level in these human cells (Fig 3b) and/or their inability to bind the human ASCT2 receptor due to $\text{SUPYN}$ sequence divergence (Extended Data 8, 9).

To gain further insight into the evolutionary origins of $\text{SUPYN}$ antiviral activity, we expressed $\text{SUPYN}$ sequences predicted for the common ancestor of hominoid and OWM (Extended Data 9) (see Methods) and assayed their antiviral activity in 293T cells. Both ancestral proteins were expressed at levels comparable to human and green monkey
SUPYN and exhibited strong restriction activity (Fig 3c). These data indicate that SUPYN antiviral activity against RDR env-mediated infection is an ancestral trait preserved over ~20 million years of hominoid evolution, but apparently lost in some OWM lineages.

Truncated HERV env-coding sequences are expressed in diverse cell types

The antiviral activity of SUPYN against RDR env-mediated infection raises the possibility that the human genome may harbor other env-derived protein-coding genes with antiviral function. To assess the potential pool of such sequences, we scanned the human genome for env-derived open reading frames (envORF) that minimally encode at least 70 aa predicted to include the receptor-binding surface domain (see Methods). This search identified a total of 1,507 unique envORFs, including ~20 env-derived sequences currently annotated as human genes such as SUPYN and SYN14,43,44. We then mined transcriptome datasets generated from human preimplantation embryos and various tissues (Supplementary Table 1), and observed that ~44% (668/1507) of envORFs showed evidence of RNA expression in at least one of the cell types surveyed (Fig 4).

These analyses revealed three general insights about expressed non-annotated envORFs: (1) like known env-derived genes, envORFs exhibit tissue-specific expression patterns (Fig 4; Extended Data 11, 12); (2) the majority of envORFs are expressed during human fetal development or in stem and progenitor cells (Fig 4; Extended Data 11); (3) envORFs are rarely expressed in differentiated tissues under normal conditions, with the exception of brain (Fig 4; Extended Data 11, 12). These analyses suggest that the human genome harbors a vast reservoir of env-derived sequences with coding and
receptor-binding potential, many of which are transcribed in a tissue-specific fashion and may encode receptor-binding antiviral peptides.

Discussion

Our expression and selection analyses firmly establish that SUPYN is a bona fide gene encoding a truncated envelope of retroviral origin that is expressed in the human preimplantation embryo and throughout placental development. Virologic assays in human cell culture show SUPYN is necessary and sufficient to confer resistance to RDR env-mediated infection, likely by competing for the receptor (ASCT2) utilized by this diverse set of retroviruses (see model, Fig 2i). The expression profile of SUPYN and the RD114 resistance phenotype of human ESCs and placental cells suggest SUPYN provides a layer of protection against RDR infection to the developing embryo, including the nascent germline. The observation that infectious and replication-competent RDRs are known to currently circulate in several mammals, both as exogenous or endogenous viruses, but not in hominoids\(^2\) lends further support to a model in which SUPYN has contributed to hominoid resistance to RDR infection and possibly endogenization.

Like SYN1, SUPYN emerged in the common ancestor of catarrhine primates and was preserved by natural selection in hominoids. The parallel evolutionary path of SYN1 and SUPYN and their pattern of expression in the placenta are compatible with a model in which SUPYN acts as a negative modulator of SYN1 fusogenic activity during STB development\(^4,5\). The developmental and antiviral functions of SUPYN are not mutually
exclusive, but may even be interlocked. Syncytins, including SYN1, are fully functional envelopes that can be incorporated into heterologous retroviral particles and exosomes originating from the placenta\textsuperscript{50-55,45}. Because ASCT2 is broadly expressed, SYN1-pseudotyped particles produced in the developing placenta have the potential to infiltrate a wide range of surrounding cell types. Thus, the physiological benefits afforded by Syncytins in promoting cell-cell fusion during STB development may have come with the cost of exposing the developing embryo to a wide variety of endogenous and exogenous invasive genetic elements, including but not limited to RDRs, that could be serendipitously enveloped by SYN1 throughout pregnancy. It is tempting to speculate that SUPYN has been maintained by natural selection to shield the developing embryo from the constant threat and adverse effects of SYN1-mediated infections. The conserved antiviral activity of ancestral hominoid and OWM SUPYN suggest resistance against RDR infection may have precipitated the initial retention of SUPYN in a catarrhine ancestor, and subsequently facilitated the domestication of SYN1 in hominoids.

More broadly, this study serves as a proof of principle that truncated envelope peptides expressed from relics of ancient retroviruses integrated in the human genome can exert and retain antiviral activities for millions of years. We identified hundreds of candidate env-derived genes in the human genome that may encode peptides with receptor-binding activity and antiviral function. Furthermore, Gag (capsid)-derived proteins encoded by endogenous retroviruses are also capable of retroviral restriction\textsuperscript{46}. Thus, it is possible that our genome holds a vast reservoir of retrovirus-derived proteins with protective activity against various zoonotic agents, including non-retroviral pathogens.
Main References


28. de Parseval, N., Casella, J.-F., Gressin, L. & Heidmann, T. Characterization of the Three HERV-H Proviruses with an Open Envelope Reading Frame


35. Marin, M., Lavillette, D., Kelly, S. M. & Kabat, D. N-linked glycosylation and sequence changes in a critical negative control region of the ASCT1 and ASCT2


**Figure 1:** Pluripotency and placentation regulatory factors drive SUPYN expression during fetal development.

(a) Violin plots summarizing SUPYN, SYN1, and ASCT2 expression in human preimplantation embryo and ESC single-cell RNA-seq data. (b) Genome browser view of the regulatory elements around the SUPYN locus in hESCs, cyto- (CTB), syncytio- (STB), BMP4 differentiated trophoblasts (TB), and 293T cells. ChIP-seq profiles are shown for indicated transcription factors and histone
modifications with shaded area highlighting regions of active chromatin. (c) UMAP visualization of TB cell clusters, shown in Extended Data 2. (d) Monocle2 pseudotime analysis of cell clusters in c illustrates the developmental trajectory of CTBs that give rise to STB and EVTB respectively. Color codes in c and d denote cell identity. (e) Heatmap represents the top 1000 differentially expressed genes (row) of single cells (column) sorted according to pseudotime analyzed in c and d. Cells are ordered according to the pseudotime progression of CTB (middle) to STB (left) and EVTB (right). SYN1, SUPYN, and ASCT2 were fetched from the heatmap below. (g) Confocal microscopy of placental villi explants stained for SUPYN (green), Actin (red) and DAPI. STBs are marked by arrowheads.
**Figure 2: SUPYN confers resistance to RDR env-mediated infection**

(a) Virus production and infection assay approach (see Methods). (b, d) Proportion of GFP+ 293T, JEG3, Jar, and shRNA-transduced Jar cells infected with HIV-RD114env or HIV-VSVg. (c) Relative infection rate of 293T and H1-ESCs normalized to mean proportion of HIV-VSVg-infected cells. (e, g, h) Relative infection rates of GFP+ 293T cells transfected with (e) wild-type (WT-SP), rescue
(Resc-SP), luciferase signal peptide (GPluc-SP), unmodified (e, g) SUPYN, (g) RD114env, or (h) SMRVenv overexpression constructs. Relative infection was determined by normalizing indicated constructs to empty vector (n ≥ 3 with ≥ 2 technical replicates; ***adj. p < 0.001; **adj. p < 0.01; *adj. p < 0.05; ANOVA with Tukey HSD). (f) Western Blot analysis (αHA, αGAPDH) of 293T cell lysates transfected with indicated constructs. (i) Model of SUPYN-dependent RDR infection restriction.
Figure 3: SUPYN emerged in a Catarrhine ancestor and has conserved antiviral activity in Hominoids.

(a) Consensus primate phylogeny with cartoon representation of SUPYN ORFs (blue box). Magenta boxes represent frame-shifts in SUPYN ORFs. Red dashed lines denote conserved premature stop codon positions. Grey bars represent degraded HERVH48env sequence. Labeled triangles denote ancestral lineage where HERVH48env was acquired. Colored circles indicate species used in b and c. SUPYN, SYN1 and SYN2 dN/dS values are shown in box (**p < 0.01; *p < 0.05; LRT). (b, c) Relative infection rates and Western Blot of 293T cells transfected with primate (b) or ancestral (c) SUPYN-HA constructs and infected with HIV-RD114env are shown. Relative infection rates were determined by normalizing GFP+ counts to empty vector. (n ≥ 3 with ≥ 2 technical replicates ***adj. p < 0.001; *adj. p < 0.05; ANOVA with Tukey HSD)
Figure 4: Expression profile of env-derived transcripts over a subset of human cell types.

Heatmap shows expression of 668 envORF loci in 66 distinct cell-types from 8 independent datasets shown in Extended Data 11. Rows represent individual envORF loci expressed (log2 CPM > 1) in at least one cell type (columns). Upper bars, located above the heatmap, denote the sequencing strategy (top) and tissue source (bottom) with the same color scheme shown in Extended Data 11. Rows are clustered into six distinct groups based on their expression. Significant envORF enrichment in cell-types was calculated using a hypergeometric test.
Methods

RNA-seq analyses

We mined published single cell transcriptome datasets (scRNA-seq) of human pre-implantation embryos isolated at developmental stages ranging from oocyte to blastocyst\textsuperscript{47} (GSE36552) and from human placenta\textsuperscript{48,49} (GSE89497, GSE87726). Reads were mapped to the human genome (hg19) with STAR\textsuperscript{50} using the following settings --

\texttt{alignIntronMin 20 --alignIntronMax 1000000 --chimSegmentMin 15 --}
\texttt{chimJunctionOverhangMin 15 --outFilterMultimapNmax 20}. Only uniquely mapped reads were considered for expression calculations. Gene level counts were obtained using \texttt{featureCounts}\textsuperscript{51} run with RefSeq annotations. Gene expression levels were calculated at \textit{Transcript Per Million} (TPM) from counts mapped over the entire gene (defined as any transcript located between \textit{Transcription Start Site} (TSS) and \textit{Transcription End Site} (TES)). Only genes and cells that met the following criteria were included in this analysis:

(1) each cell must express at least 5,000 genes; (2) each gene must be expressed in at least 1\% of cells; (3) each gene must be expressed with \texttt{log2 TPM} >1. We clustered cells meeting these criteria using the default parameters of the Seurat\textsuperscript{52} package (v3.1.1) implemented in R (v3.6.0). Seurat applies the most variable genes to get top principal components that are used to discriminate cell clusters in tSNE or UMAP plots. We set Seurat to use 10 principal components in this cluster analysis. For the placental scRNAseq data (Fig1, Extended Data 2), the 2000 most differentially expressed genes were used to define cell clusters. Major clusters corresponding to CTB, STB, EVTB, macrophages, and stromal cells were identified based on the expression of known marker
genes. Monocle2\textsuperscript{53} was used to perform single-cell trajectory analysis and cell ordering along an artificial temporal continuum. The top 500 differentially expressed genes were used to distinguish between CTB, STB and EVTB cell populations. The transcriptome from each single cell represents a pseudo-time point along an artificial time vector that denotes the progression of CTB to STB or EVTB respectively.

Data generated on the 10X Genomics scRNA-seq platforms were processed in the following way. The processed data matrix from Vento-Tormo et al.\textsuperscript{54} was retrieved from the E-MTAB-6701 entry. The normalized counts and cell-type annotations were used as provided by the original publications. Seurat was used for filtering, normalization and cell-type identification. The following data processing steps were performed: (1) Cells were filtered based on the criteria that individual cells must have between 1,000 and 5,000 expressed genes with a count \( \geq 1 \); (2) cells with more than 5% of counts mapping to mitochondrial genes were filtered out; (3) data was normalized by dividing uniquely mapping read counts (defined by Seurat as unique molecular identified (UMI)) for each gene by the total number of counts in each cell and multiplying by 10,000. These normalized values were then natural-log transformed. Cell types were defined by using the top 2000 variable features expressed across all samples. Clustering was performed using the “FindClusters” function with largely default parameters; except resolution was set to 0.1 and the first 20 PCA dimensions were used in the construction of the shared-nearest neighbor (SNN) graph and the generation of 2-dimensional embeddings for data visualization using UMAP. Cell types were assigned based on the annotations provided by the original publication.
Bulk RNAseq datasets generated from placenta\textsuperscript{55}, 293T\textsuperscript{56,57} and human immune cells\textsuperscript{58}, were processed as described above. Briefly, reads were mapped with STAR and uniquely mapped reads were counted with featureCounts.

ChIP-seq, DNase-seq and ATAC-seq data analysis

Various ChIP-seq datasets representing histone modifications and transcription factors in human embryonic stem cells and their differentiation were retrieved from\textsuperscript{59,60} (GSE61475, GSE99631). We obtained the H3K27Ac\textsuperscript{61} (GSE127288) for CTB to STB primary cultures, H3K4Me1 for trophoblasts\textsuperscript{62} (GSE118289), H3K4Me3, H3K27Me3 for differentiated trophoblasts\textsuperscript{63} (GSE105258), and GATA2/3, TFAP2A/C\textsuperscript{63} (GSE105081) ChIP-seq datasets in raw fastq format. DNase-seq and ATAC-seq datasets were retrieved from Gao et al.\textsuperscript{64} (GSA:CRA000297) and Wu et al.\textsuperscript{65} (GSE101571) respectively.

Reads from the above described datasets were aligned to the hg19 human reference genome using Bowtie2\textsuperscript{66} run in --very-sensitive-local mode. All reads with MAPQ < 10 and PCR duplicates were removed using Picard and samtools\textsuperscript{67}. All of the peaks were called by MACS2\textsuperscript{68} (https://github.com/macs3-project/MACS) with the parameters in narrow mode for TFs and broad mode for histone modifications keeping FDR < 1%. ENCODE-defined blacklisted regions\textsuperscript{69} were excluded from called peaks. We then intersected these peak sets with repeat elements from hg19 repeat-masked coordinates using bedtools \textit{intersectBed}\textsuperscript{70} with a 50\% overlap. To visualize over Refseq genes (hg19) using IGV\textsuperscript{71}, the raw signals of ChIP-seq were obtained from MACS2, using the parameters: -g hs -q 0.01 -B. The conservation track was visualized through UCSC
genome browser\textsuperscript{14} under net.chain alignment of given non-human primates (NHPs) and
merged beneath the IGV tracks.

Cell culture

293T cells were cultured in DMEM (GIBCO, 11995065) containing 10% Fetal Bovine Serum (FBS) (GIBCO, 10438026). Jar cells (provided by Carolyn Coyne) were cultured in RPMI (GIBCO, 11875093) containing 10% FBS. JEG3 cells were cultured in MEM (GIBCO, 11095080) containing 10% FBS. Culture medium for these cell lines was supplemented with sodium pyruvate (GIBCO, 11360070), glutamax (GIBCO, 35050061), and Penicillin Streptomycin (GIBCO, 15140122) according to manufacturer specifications. H1-ESCs (obtained from WiCell) were grown on Matrigel (Corning, 356277) coated plates in MTESR+ (Stemcell, 05825) growth-media and sub-cultured using Accutase (Innovative Cell Technologies, AT-104) and MTESR+ supplemented with CloneR (Stemcell, 05888). All cell lines were cultured at 37°C and 5% CO\textsubscript{2}.

Vector cloning

DHIV3-GFP, phCMV-RD114env, psi(-)-amphoMLV plasmids were provided by Vicente Planelles (University of Utah). pCGCG-SMRVenv plasmid was provided by Welkin Johnson (Boston University). psPAX2 and pVSVg plasmids were provided by John Lis (Cornell University). The following cloning approaches were performed using primers and constructs described in Supplementary Table 3. HERVH1env ORF was PCR-amplified using Q5 polymerase (NEB, M0491L) from HeLa and 293T genomic DNA respectively and cloned into a TOPO vector (ThermoFisher, 450245).
To generate stable SYN1 and SUPYN knock-down cell lines, pHIV lentiviral constructs containing shRNAs targeting SYN1 and SUPYN respectively were cloned using the following strategy. The shRNA encoded in pHIV7-U6-shW3, generously provided by Lars Aagaard (Aarhus University), targets SYN1. SUPYN-targeting shRNAs were designed using siRNA sequences employed by Jun Sugimoto as a template. pHIV7 lentiviral constructs were cloned using the pHIV7-U6-shW3 plasmid as a template. pHIV7-U6-shSup-cer, pHIV7-U6-shSup-puro, pHIV7-U6-shC-cer, pHIV7-U6-shC-puro, pHIV7-U6-shSyn1-cer, pHIV7-U6-shSyn1-puro were generated using a Gibson assembly approach. To replace the native GFP marker of pHIV7-U6-shW3 with a Cerulean reporter or puromycin resistance marker, we digested pHIV7-U6-shW3 with NheI (NEB, R3131S) and KpnI (NEB, R3142S). This digest resulted in the production of three DNA fragments: pHIV7 backbone, GFP-, and WPRE-containing fragments. We separately PCR amplified each selection marker and WPRE containing pHIV7 fragment. InFusion cloning was then used to ligate the digested pHIV7 backbone to the Cerulean or puromycin cassette and WPRE containing PCR product. shRNAs were cloned into the pHIV7-Cerulean/puromycin transfer construct previously digested with NotI (NEB, R0189S) and NheI. U6-promoter containing shRNA cassettes and the CMV promoter driving marker cassette expression were PCR amplified and subsequently InFusion cloned into the NotI/NheI digested pHIV7-cerulean/puromycin backbone. All pHCMVenv and SUPYN expression constructs, described in this study, were generated as follows: HA-tagged and untagged ORFs with pHCMV homologous overhanging sequence were either PCR amplified using Q5 polymerase (NEB, M0491S) or synthesized (IDT) (see Supplementary Table 2), and cloned into EcoRI (NEB,
R3101T) digested pHCMV backbones using the InFusion cloning kit (Takara Bio, 638920). To generate siRNA-resistant SUPYN rescue constructs, we replaced the native signal peptide sequence $^d$ (which is targeted by siRNAs used in this study) with (1) a *Gaussia princeps* luciferase SP (SUPYN-lucSP)$^{73,74}$ and (2) a shSUPYN resistant SUPYN rescue construct (SUPYN-rescSP) in which the codons were modified to retain the codon identity but disrupt siRNA binding.

**Antibodies**

All antibodies used in this study are commercially available. $\alpha$-GAPDH (D4C6R, D16H11), $\alpha$-βactin (D6A8), $\alpha$-HA (C29F4), $\alpha$-ASCT2 (V501) primary antibodies were purchased from Cell Signaling Technology. $\alpha$-Mouse (#7076) and $\alpha$-Rabbit (#7074) HRP conjugated secondary antibodies were purchased from Cell Signaling Technology. IRDye secondary antibodies were purchased from Licor (925-32211, 925-68072, 925-32210, 925-68073). $\alpha$-SUPYN primary antibody was purchased from Phoenix Pharma (H-059-052). Alexa-fluor conjugated secondary antibody was purchased from Invitrogen.

**Western Blot**

Whole cell extracts from cultured cell lines were prepared using 1x GLO lysis buffer (Promega, E266A). One third volume of 4x Laemli buffer was added to one volume whole cell extract samples, then incubated at 95°C for 5 minutes, and sonicated for 15 minutes at 4°C (amplitude 100; pulse interval 15 seconds on, 15 seconds off). Approximately 30 ug of protein were separated by SDS-PAGE (BioRad, 1610175), transferred to PVDF membrane (BioRad, 1620177), blocked according to antibody manufacturers
specification, and incubated overnight in appropriate primary antibody then incubated in
IRDye or peroxidase conjugated goat anti-mouse or anti-rabbit secondary antibodies for
1 hour at room temperature. Protein was then detected using ECL reagent (BioRad,
1705061) or the Licor Odyssey imaging system.

IF microscopy
Human second trimester placental tissue that resulted from elective terminations was obtained
from the University of Pittsburgh Health Sciences Tissue Bank through an honest broker system
after approval from the University of Pittsburgh Institutional Review Board (IRB) and in
accordance with the University of Pittsburgh’s tissue procurement guidelines. Tissue was
excluded in cases of fetal anomalies or aneuploidy. Third trimester placental tissue was obtained
through the Magee Obstetric Maternal & Infant (MOMI) Database and Biobank after approval from
the University of Pittsburgh IRB. Women who had previously consented for tissue donation and
underwent cesarean delivery were included. Placental tissues were fixed in 4% PFA (in 1x
PBS) for 30 minutes, permeabilized with 0.25% Triton X-100 for 30 minutes (on a rocker),
washed with 1x PBS and then incubated with primary anti-Suppressyn antibody at 1:200
in 1x PBS for 2-4 hours at room temperature. These samples were incubated with Alexa-
fluor conjugated secondary antibody (Invitrogen) diluted 1:1000 and counterstained with
actin. DAPI was included in our PBS and then mounted in Vectashield mounting medium
with DAPI (Vector Laboratories, H-1200).

Virus production
Low passage 293T cells were used to produce lentiviral particles. DHIV3-GFP and env-
expression plasmids were co-transfected at a mass ratio of 2:1 using lipofectamine 2000
shRNA encoding lentiviral particles were produced by co-transfecting pHIV7, psPAX2, pVSVg according to BROAD institute lentiviral production protocol (https://portals.broadinstitute.org/gpp/public/resources/protocols) using Lipofectamine 2000. Growth media was replaced on transfected cells after overnight incubation. At 72 hours post-transfection, virus containing supernatant was harvested, centrifuged to remove cell debris, filtered through a 0.45 um pore filter, and stored at -80°C.

Infection Assays

293T cells were transfected with env-overexpression constructs using Lipofectamine 2000 and incubated 24 hours. Transfected cells were infected with reporter virus by applying virus (HIV-RD114env, HIV-VSVg, HIV-SMRVenv) stocks in the presence of polybrene (Santa Cruz Bio, sc-134220) at a final concentration of 4 ug/mL. After 6-8 hours, virus stock was replaced with fresh growth media. Infected cells were maintained for 72 hours, replacing media when necessary, and harvested with trypsin. Detached cells were suspended in fresh growth media, strained and analyzed by flow cytometry. For the H1-ESC infection experiment, relative infection rates were calculated by normalizing the percent GFP+, HIV-RD114env infected cells to the percent GFP+ HIV-VSVg infected cells. For env/SUPYN overexpression experiments, relative infection rates were calculated by normalizing the percent GFP+ env/SUPYN-transfected cells to the percent GFP+ empty vector transfected cells. ANOVA with Tukey HSD tests were implemented in R (v3.6.3).
**Placental cell shRNA transduction**

Placenta-derived cell lines were treated with pHIV-shRNA-virus-containing supernatant and incubated for 72 hours as described in Infection Assays. Cerulean positive cells were sorted using the BD FACS Aria cytometer. Cells transduced with puroR cassette were treated with Puromycin (GIBCO, A1113802) at a final concentration of 3.5 μg/mL for 7 days, then cultured in regular growth media.

**RT-qPCR**

RNA was isolated from cultured cells using the RNeasy Mini Kit (Qiagen, 74104) and an on column dsDNAse digestion (Qiagen, 79254) was performed. 1-3 μg of total RNA were used to generate cDNA with the maxima cDNA synthesis with dsDNAse kit (ThermoFisher, K1681). qPCR reactions were performed using the LC480 Instrument with Sybr Green PCR master mix (Roche, 04707516001) according to manufacturer's protocol and using primers indicated in Supplementary Table 2. Gene expression was then quantified using the ΔΔCT method\(^{75}\). 18S expression was used as a reference housekeeping gene. Wilcoxon rank sum tests were performed using R (v3.6.3).

**Envelope evolutionary sequence analyses**

Orthologous \textit{SUPYN}, \textit{SYN1}, and \textit{SYN2} sequences were extracted from the 30-species MULTIZ alignment\(^{14}\) and formatted for sequence alignment using the phast package\(^{76}\). These and additional syntenic \textit{SUPYN} and \textit{SYN2} open reading frame sequences were validated/identified by BLASTn\(^{77}\) search with default settings of publicly available Catarrhine primate genomes (ncbi.nih.gov). Mariam Okhvat of the Carbone Lab (Oregon
Health and Science University) generously provided BAM files containing read alignment information for SUPYN, SYN1, and SYN2 generated from whole genome sequencing of Hoolock leuconedys (Hoolock Gibbon), Symphalangus syndactylus (Siamang), Hylobates muelleri (Müller's Gibbon), Hylobates lar (Lar Gibbon), Hylobates moloch (Silvery Gibbon), Hylobates pileatus (Pileated Gibbon), and Nomascus gabriellae (Yellow-cheeked Gibbon). Where multiple individuals were sequenced, a consensus sequence was generated using samtools\textsuperscript{67} and JalView\textsuperscript{78}.

To perform \(dN/dS\) analyses, orthologous env sequences (>90bp length) encoding the mature sequence downstream of the signal peptide cleavage site, were aligned using MEGA7\textsuperscript{79} and manually converted to PHYLIP format. A Newick tree was generated based on this alignment using the maximum likelihood algorithm implemented in MEGA7. *Codeml*, implemented in the PAML package, was run to calculate \(dN/dS\) values and log likelihood (LnL) scores generated under models M0, M1, M2, M7 and M8\textsuperscript{38}. Chi-square tests comparing LnL scores generated under models of neutral evolution and selection were performed.

We used two approaches to reconstruct ancestral hominoid and OWM SUPYN sequences. First, we reconstructed ancestral SUPYN sequences using the majority rule consensus sequence (calculated in JalView) of the hominoid and OWM clade respectively. At positions where nucleotide identity was ambiguous, the dominant nucleotide identity in the neighboring clade was used as a tiebreaker. These sequences were used for our infection assays shown in Fig 3c. We also employed a maximum likelihood approach using the *baseml* program, implemented in PAML\textsuperscript{38}. We reconstructed ancestral SUPYN sequences using the hominoid species, shown in Fig 3a,
and the 6 OWM monkeys with the most complete SUPYN-coding open reading frame (olive baboon, drill, crab-eating macaque, rhesus macaque, japanese macaque, green monkey) as our input sequences. Because PAML requires a Newick tree as an input, the MEGAX maximum likelihood algorithm was used to generate a Newick tree with the above described SUPYN sequences. Baseml was run using models 3-7 (F84, HKY85, T92, TN93, REV). As shown in Extended Data 10, both the consensus- and maximum likelihood reconstructions were identical for the OWM SUPYN sequences. The consensus-based hominoid sequence reconstruction differed from our maximum likelihood-based reconstruction by two amino acids. These two positions are unlikely to affect the function of the resulting protein because these sites are identical to siamang SUPYN, which restricts RD114env-mediated infection (Fig 3b, c).

**Genome-wide search for endogenous retrovirus derived envelope open reading frames**

Candidate envelope open reading frames (envORF) were identified by performing tBLASTn searches of the hg19 human genome assembly using envelope amino acid sequences, obtained from Repbase and published retroviral envelope sequences, as a query. Collected hits were used as a query to repeat a tBLASTn search, initially yielding 82715 candidate envORF sequences. This list of candidates was filtered using the following criteria. (1) EnvORFs must have a length of ≥70 aa. (2) Hits starting at position ≥300 aa were removed because such open reading frames are predicted to encode a portion of the envelope transmembrane domain, which is not expected to play a role in receptor binding. (3) After these processing steps, our list was further concatenated to
only include unique open reading frame sequences (n=1507) in the hg19 genome assembly.

To estimate envORF expression, we mined publicly available scRNA-seq datasets generated on the SMART-seq2 platform because this technology yields higher genomic coverage compared with methods employing poly(A) enrichment. We gathered scRNA-seq datasets in fastq format from preimplantation embryos47 (GSE36552), placenta48 (GSE89497), primordial germ cells85 (GSE63818), hematopoiesis86 (GSE75478), neuronal differentiation87 (GSE93593), prefrontal cortex88 (GSE67835), and pancreas89 (GSE81547) (Supplemental Table 1). We also included bulk RNA-seq of immune cells from six individuals comprising 25 subtypes58 (GSE118165) (Supplemental Table 1).

We curated our envORF sequences (in fasta format) and hg19 refseq gene models14 (gtf format) to calculate envORF expression with high confidence. First, we masked envORF DNA sequences from the hg19 genome. These sequences were added, as individual fasta contigs, to the hg19 reference assembly. We then modified the transcriptome model by appending the envORFs models (gtf format) to the hg19 refseq gene models. Next, we indexed the modified reference sequences using STAR default parameters, by providing the new transcriptome models that include the envORFs. This approach enabled us to simultaneously calculate envORF and protein-coding gene expression. After aligning and retaining uniquely mappable sequencing reads, we calculated envORF and gene expression (see Single cell RNA-seq analysis procedures above). We processed the individual datasets independently, then constructed the expression matrix using Seurat and checked if the obtained clusters were consistent with the original studies. Despite including the envORFs, we identified each cell type in similar proportions,
as shown in their corresponding studies (Extended Data 11). Then we merged all of the datasets into a single pool for our downstream analysis while maintaining cell type and tissue annotations. We scaled and normalized these datasets using Seurat within the R environment and checked the normalization status of merged datasets with a UMAP biplot. This was done by using the most variable genes to ensure that cells did not cluster based on platform or batch differences (Extended Data 11 and Supplemental Table 1). We calculated gene and envORF expression (TPM) using uniquely mapped reads. Any envORF with TPM < 1 in > 99% of cells across the data frame was discarded from further analysis. This strategy identified 668 unique envORFs with evidence for expression. Finally, expression of each envORF across all cell types was used to cluster envORF according to their expression profile, which were visualized in Fig. 4. To survey the expression of known envelope genes, we utilized Gtex\textsuperscript{90} bulk RNA-seq datasets (phs000424.v6.p1), which were generated from distinct post-mortem tissues.

**Methods References**


Acknowledgements

This work was funded by Cornell University and by the National Institutes of Health to CF (R01 GM112972, R35 GM122550). We thank Dr. Vicente Planelles for sharing reporter virus plasmids and training; Dr. Welkin Johnson for sharing SMRVenv expression plasmid; Dr. Lars Aagaard for sharing shRNA transduction constructs; Dr. John Lis for providing lentiviral packaging construct; Dr. Amnon Koren and Rita Rebello for providing embryonic stem cell cultures and technical support respectively; Dr Lucia Carbone and Dr. Mariam Okhvat for sharing unpublished gibbon genome sequences. We thank Ray Malfavon-Borja for producing an initial list of envelope open reading frames in the human genome. We thank Maia Clare for her contribution to evolutionary sequence analyses. We thank members of the Feschotte lab for helpful advice and discussion throughout the project.
Author contributions

CF conceived of this project. JAF and CF designed and developed this project. JAF designed and conducted all experiments, evolutionary sequence analyses, and analyzed all experimental data. MS performed all gene expression and regulation analyses. HBC and RAK helped perform infection assays and evolutionary sequence analyses. CBC performed placental stains. JAF, CF and MS wrote this manuscript.

Competing interest declaration

The authors declare no competing interests.

Additional Information
Extended Data

Extended Data Figure 1: SUPYN is expressed in human early embryo and placenta but not in 293T cells.

(a) Bar graphs show SUPYN and SYN1 expression in indicated cell types. (b) Genome browser view showing ATAC-seq and DNAse-seq signals at the SUPYN locus, including upstream and downstream sequences. Framed region highlights overlapping peaks at the SUPYN locus. (c) Line plot depicts HERVenv gene expression level during BMP4-mediated in vitro hESCs to trophoblast differentiation. Time points correspond to cells harvested 8hr, 24hr, 48hr, and 72hr post BMP4 treatment.
Extended Data Figure 2: Defining lineage-specific SYN1, SUPYN, and ASCT2 expression from placental single-cell transcriptomics

(a) UMAP plot generated from published scRNA-seq data generated from 1st trimester placental explants. Colors denote placental (pink, red, orange, and yellow) immune (blue and green) and maternal cell lineages (white and grey). (b) Feature plots visualize single-cell expression level of lineage-defining marker genes. (c) Simplified UMAP plot, shown in a, of scRNAseq data displaying trophoblast (yellow), decidual (green) and immune (purple) cell identity. Sub-panels display single-cell-level expression of indicated genes. (d) Violin plots denote single-cell SUPYN and ASCT2 expression in multiple placental-cell lineages.
Extended Data Figure 3: ASCT2 and SUPYN expression in 2nd and 3rd trimester human placenta.

Confocal microscopy of 2nd (week 21) and 3rd (week 31) trimester placental villi explants. Villi were stained for ASCT2 (green upper panels) or SUPYN (green lower panels) and Actin (red). Cell nuclei are marked with DAPI (blue).
Extended Data Figure 4: Flow Cytometry analysis scheme. Representative sequential gating scheme to assess reporter virus infection rate.
Extended Data Figure 5: Characterization of shRNA transduced Jar cells and validation of env overexpression constructs.

(a, b) SUPYN and SYN1 knock down was validated by qPCR. Bar plots represent mean relative gene expression normalized to shCON in cerulean-sorted and puromycin-selected cell lines respectively (n = 3). Error bars represent ± standard error mean (*p < 0.1; Wilcox rank sum test).

(b) Western Blot analysis (αGAPDH, αASCT2) of shRNA-transduced Jar cell lysates. Putatively glycosylated and unglycosylated ASCT2 are marked by filled and empty arrowheads respectively.
Extended Data Figure 6: SUPYN expression is sufficient to specifically restrict RDR env-mediated infection

(a, b) 293T cells, transfected with HA-tagged SUPYN and env constructs, were infected with HIV-RD114env (a) and -VSVg (b) respectively. Relative infection rates were determined by normalizing GFP+ counts to empty vector. (n ≥ 3 with ≥ 1 technical replicate; ***adj. $p < 0.001$; *adj. $p < 0.05$; Tukey HSD). (c) Western Blot analysis ($\alpha$HA, $\alpha$GAPDH, $\alpha$ASCT2) of 293T cell lysates following transfection with indicated constructs.
Extended Data Figure 7: SUPYN locus conservation in primates.

UCSC genome Browser snapshot of SUPYN-coding locus with surrounding sequence.

Modified NCBI RefSeq gene, simian whole genome alignment (from Multiz 30-species track), and RepeatMasker repetitive element tracks are shown. The SUPYN-coding HERVH48 provirus is highlighted by the yellow box.
Extended Data Figure 8: Nucleic acid sequence alignment of primate Suppressyn orthologs.

Suppressyn encoding nucleotide sequences are shaded blue based on a minimum sequence identity threshold of 45% (light), 75% (medium) and 80% (dark). Conserved ape-specific and ancestral stop codons are highlighted in red.
Extended Data Figure 9: Amino Acid sequence alignment of primate SUPYN orthologs.

Primate and ancestral SUPYN peptide sequences are shown. Sequences are shaded blue based on a minimum sequence identity threshold of 45% (light), 75% (medium) and 80% (dark).

Ancestral SUPYN sequences are based on our consensus-based sequence reconstruction (see Methods).
Extended Data Figure 10: Amino Acid sequence alignment of ancestral SUPYN sequences.

Peptide sequences of consensus and maximum likelihood-based SUPYN sequence reconstructions (see methods) are aligned.
Extended Data Figure 11: Analysis of envORF expression in single cell RNA-seq taken from various human tissue sources

(a) UMAP plots illustrate the Louvain clustering of seven independent sc-RNA-seq datasets corresponding to the development of human embryos and somatic tissues (see Methods). The large central UMAP plot represents our integrative analysis of seven scRNA-seq and one bulk...
RNA-seq datasets obtained from independent studies (Supplemental Table 1). Surrounding UMAP plots represent the combined expression of human RefSeq and envORF genes (supplemental Table 3) with cell identities labeled. (b) Multiple violin plots demonstrate the expression of annotated ERV envelopes in each dataset shown in a.

Extended Data 12: Annotated human endogenous retrovirus envelope expression in various human tissues.

(a) The boxplot shows the transcript expression distribution of annotated ERV envelopes in tissues assayed by the GTEx project (Supplementary Table 1). (b) The heatmap displays the
expression (log2 TPM) of individual ERV envelope genes (rows) in each tissue (columns). The color scheme ranges gradually from no expression (gold) to higher expression (midnight blue).

### Tables

#### Supplementary Table 1: External data sources

<table>
<thead>
<tr>
<th>Description</th>
<th>Author</th>
<th>Year</th>
<th>PMID</th>
<th>Methods Reference</th>
<th>Dataset</th>
<th>Tissue source</th>
<th>Figure</th>
</tr>
</thead>
<tbody>
<tr>
<td>scRNAseq</td>
<td>Yan et al.</td>
<td>2013</td>
<td>23934149</td>
<td>46</td>
<td>GSE36552</td>
<td>embryo</td>
<td>1a, 4d, Ext 11</td>
</tr>
<tr>
<td></td>
<td>Vento-Tormo et al.</td>
<td>2018</td>
<td>30429548</td>
<td>53</td>
<td>E-MTAB-6701</td>
<td>placenta</td>
<td>1d-f, Ext 2</td>
</tr>
<tr>
<td></td>
<td>Liu et al.</td>
<td>2018</td>
<td>30042384</td>
<td>47</td>
<td>GSE89497</td>
<td>placenta</td>
<td>1f, 4d, Ext 11</td>
</tr>
<tr>
<td></td>
<td>Pavlicev</td>
<td>2017</td>
<td>28174237</td>
<td>48</td>
<td>GSE87726</td>
<td>placenta</td>
<td>1f</td>
</tr>
<tr>
<td></td>
<td>Close</td>
<td>2017</td>
<td>28279351</td>
<td>85</td>
<td>GSE93593</td>
<td>Neuro_diff</td>
<td>4d, Ext 11</td>
</tr>
<tr>
<td></td>
<td>Guo</td>
<td>2015</td>
<td>26046443</td>
<td>83</td>
<td>GSE63818</td>
<td>PGC</td>
<td>4d, Ext 11</td>
</tr>
<tr>
<td></td>
<td>Velten</td>
<td>2017</td>
<td>28319093</td>
<td>84</td>
<td>GSE75478</td>
<td>HSC</td>
<td>4d, Ext 11</td>
</tr>
<tr>
<td></td>
<td>Enge</td>
<td>2017</td>
<td>28965763</td>
<td>87</td>
<td>GSE81547</td>
<td>Pancreas</td>
<td>4d, Ext 11</td>
</tr>
<tr>
<td></td>
<td>Darmanis</td>
<td>2015</td>
<td>26060301</td>
<td>86</td>
<td>GSE67835</td>
<td>PFC</td>
<td>4d, Ext 11</td>
</tr>
<tr>
<td>Bulk RNAseq</td>
<td>Zadora</td>
<td>2017</td>
<td>28904069</td>
<td>available via approval</td>
<td>placenta</td>
<td>1f</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GTEx Consortium</td>
<td>2017</td>
<td>29022597</td>
<td>88</td>
<td>phs000424.v6.p1</td>
<td>Human tissues</td>
<td>Ext 12</td>
</tr>
<tr>
<td></td>
<td>D Antonio</td>
<td>2017</td>
<td>28874753</td>
<td>55</td>
<td>E-MTAB-5714</td>
<td>293T</td>
<td>Ext 1a</td>
</tr>
<tr>
<td></td>
<td>Chung</td>
<td>2018</td>
<td>29395325</td>
<td>56</td>
<td>GSE99249</td>
<td>293T</td>
<td>Ext 1a</td>
</tr>
<tr>
<td></td>
<td>Calderon</td>
<td>2019</td>
<td>31570894</td>
<td>57</td>
<td>GSE118165</td>
<td>Immune cells</td>
<td>4d, Ext 11</td>
</tr>
<tr>
<td>Method</td>
<td>Author</td>
<td>Year</td>
<td>PMID</td>
<td>GSE/GSA</td>
<td>Tissue</td>
<td>Ext</td>
<td></td>
</tr>
<tr>
<td>-------------------</td>
<td>-----------------</td>
<td>-------</td>
<td>-----------------------</td>
<td>---------</td>
<td>-----------------</td>
<td>------</td>
<td></td>
</tr>
<tr>
<td>ChIPseq</td>
<td>Barakat</td>
<td>2018</td>
<td>PMID: 30033119</td>
<td>GSE99631</td>
<td>hESC</td>
<td>1b</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Krendl</td>
<td>2017</td>
<td>PMID: 29078328</td>
<td>GSE105258</td>
<td>hESC</td>
<td>1c, Ext 1a-b</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Krendl</td>
<td>2017</td>
<td>PMID: 29078328</td>
<td>GSE105081</td>
<td>hESC</td>
<td>1c, Ext 1a-b</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Tsankov</td>
<td>2015</td>
<td>PMID: 25693565</td>
<td>GSE61475</td>
<td>hESC</td>
<td>1b, Ext 1d</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Dunn-Fletcher</td>
<td>2018</td>
<td>PMID: 30231016</td>
<td>GSE118289</td>
<td>placenta</td>
<td>1c</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Kwak</td>
<td>2019</td>
<td>PMID: 31294776</td>
<td>GSE127288</td>
<td>placenta</td>
<td>1c</td>
<td></td>
</tr>
<tr>
<td>DNAse-seq</td>
<td>Gao</td>
<td>2018</td>
<td>PMID: 29526463</td>
<td>GSA:CRA000297</td>
<td>embryo</td>
<td>Ext 1c</td>
<td></td>
</tr>
<tr>
<td>ATAC-seq</td>
<td>Wu</td>
<td>2018</td>
<td>PMID: 29720659</td>
<td>GSE101571</td>
<td>embryo</td>
<td>Ext 1c</td>
<td></td>
</tr>
</tbody>
</table>
### Supplemental table 2: Primer list

<table>
<thead>
<tr>
<th>ID</th>
<th>Sequence</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>JF023</td>
<td>TTTTGGCAAAGAATTCATGGCCTGTATCTACCCA</td>
<td>pHCMV human suppressyn fwd</td>
</tr>
<tr>
<td>JF024</td>
<td>CCTAGGAGTGAAATTCTTATAAGATTTTGTATAAA GGATGG</td>
<td>pHCMV human suppressyn rev</td>
</tr>
<tr>
<td>JF025</td>
<td>TTTTGGCAAAGAATTCATGGCCTGTATCTACCCA ACC</td>
<td>pHCMV human suppressyn-HA fwd</td>
</tr>
<tr>
<td>JF026</td>
<td>CCTAGGAGTGAAATTCTTATAAGATTTTGTATAAA GGATGG ATCG</td>
<td>pHCMV human suppressyn-HA rev</td>
</tr>
<tr>
<td>JF046</td>
<td>TTTTGGCAAAGAATTCATGGCCTGTATCTACCCA T</td>
<td>pHCMV amphiMLVenv fwd</td>
</tr>
<tr>
<td>JF047</td>
<td>CCTAGGAGTGAAATTCTTATAAGATTTTGTATAAA GGATGG GGGT</td>
<td>pHCMV amphiMLVenv rev</td>
</tr>
<tr>
<td>JF081</td>
<td>GCATAGTATGCATCTGACTACCCA</td>
<td>HERVH-1 (H62) provirus fwd</td>
</tr>
<tr>
<td>JF082</td>
<td>CATGACTCGGATGAGGGGAC</td>
<td>HERVH-1 (H62) provirus rev</td>
</tr>
<tr>
<td>JF091</td>
<td>ATCATTTTGGCAAAGAATTCATGATCTTTTGTGAAGCACC</td>
<td>pHCMV HERVH-1env fwd</td>
</tr>
<tr>
<td>JF093</td>
<td>CAGCCTGCACCTGAGGAGTGAattCtaagctatctgtatgttgtaAGCTGAAGGGAGGGATTAAGTGAAG</td>
<td>pHCMV HERVH-1env-HA rev</td>
</tr>
<tr>
<td>JF130</td>
<td>GGCGCTATAGACTGCTTTAATTACTATGCTTACC</td>
<td>pHIV7 eGFP downstream sequence fwd</td>
</tr>
<tr>
<td>JF131</td>
<td>GCGAGAGGCAGCCAGGCCAGCCAGCCAGCCAGGGAGAGGGTTTACTTGTAAG</td>
<td>pHIV7 eGFP downstream sequence rev</td>
</tr>
<tr>
<td>JF135b</td>
<td>tttgggcctactgacgacagtcaaggtcgg gcagga</td>
<td>U6 shRNA cloning fwd (pHIV-shSup cloning)</td>
</tr>
<tr>
<td>JF136</td>
<td>ccaaccacttctatactacacaataatagaaa gtgtgggtagatacatcttttgcctcttcacaa gatatataaagccgaaga</td>
<td>U6-shSup rev (pHIV-shSup cloning)</td>
</tr>
<tr>
<td></td>
<td>Sequence</td>
<td>Description</td>
</tr>
<tr>
<td>------</td>
<td>--------------------------------------------------------------------------</td>
<td>--------------------------------------------------</td>
</tr>
<tr>
<td>JF138</td>
<td>CTACCTCCTTTTGCATACCTACAAATATCGAAAG GAAGTTAAGGTAGGCTCGGGGCTTTTGTCTCTTTCCACAA GATATATAAGCCGAAAG</td>
<td>U6-shC rev (pHIV-shC cloning)</td>
</tr>
<tr>
<td>JF145</td>
<td>gtagtatagaaagttggtggtagatacatctttgtaccggagtcggtatccactagagatgg</td>
<td>shSup-CMV promoter fwd (pHIV7-shSup cloning)</td>
</tr>
<tr>
<td>JF147</td>
<td>GTAGTATCGAGGAAGTAAAGTGCTCCGGGCTtttttgatccgagctgatccactagagatgg</td>
<td>shC-CMV promoter fwd (pHIV7-shC cloning)</td>
</tr>
<tr>
<td>JF152b</td>
<td>TGAACCGTCAGATCCGCTAGCATGGCAGCGCCACCGAGTACAAG</td>
<td>cerulean fwd (pHIV7-shRNA-cerulean cloning)</td>
</tr>
<tr>
<td>JF153b</td>
<td>AGAATACCACTTTACTTTGACGCTGGACATGCATCCATGCC</td>
<td>cerulean rev (pHIV7-shRNA-cerulean cloning)</td>
</tr>
<tr>
<td>JF154b</td>
<td>CTGTACAAGTAAACTGGTATTCTTAAGATGGTGTC</td>
<td>cerulean-WPRE fwd (pHIV7-shRNA-cerulean cloning)</td>
</tr>
<tr>
<td>JF155</td>
<td>TGAACCGTCAGATCCGCTAGCATGGCACCGAGTACAAG</td>
<td>puroR fwd (pHIV7-shRNA-puroR cloning)</td>
</tr>
<tr>
<td>JF156</td>
<td>AGAATACCACTTTATGTTGAACCATTATAAGCCTGAC</td>
<td>puroR rev (pHIV7-shRNA-puroR cloning)</td>
</tr>
<tr>
<td>JF157</td>
<td>TTACAAATAACTTGGTATCTTAAGATGGTGCTTC</td>
<td>puroR-WPRE fwd (pHIV7-shRNA-puroR cloning)</td>
</tr>
<tr>
<td>JF155</td>
<td>CCGTGcccctgtagcggatcttgacggttc</td>
<td>pHIV7-CMV promoter rev (pHIV-shRNA puromycin cloning)</td>
</tr>
<tr>
<td>JF166</td>
<td>TGTCACCACTGCTAGCGGATGCTGTCGCTCCCTC</td>
<td>pHCMV SMRVenv cloning primer fwd</td>
</tr>
<tr>
<td>RAK010</td>
<td>TTTTGGCAAGAGATTACATGCTCTGACTCCTCATCTTTC</td>
<td>pHCMV SMRVenv cloning primer rev</td>
</tr>
<tr>
<td>RAK011</td>
<td>CCTGGAGGATGTTAATTCTACAGTCTTGCTGACTATTCAGAGGATGTCACGA</td>
<td>qPCR primers</td>
</tr>
<tr>
<td>JF042</td>
<td>GCAATTATTCCCCCATGAACG</td>
<td>18S fwd</td>
</tr>
<tr>
<td>JF043</td>
<td>GGCCTCCTAAACCATGCCA</td>
<td>18S rev</td>
</tr>
<tr>
<td>JF108</td>
<td>CTGTGCATGCACATCGGCTACTG</td>
<td>Suppressyn fwd</td>
</tr>
<tr>
<td>JF109</td>
<td>GACACATGGGCCGACACAAACT</td>
<td>Suppressyn rev</td>
</tr>
<tr>
<td>JF112</td>
<td>ACCAGGAGGAGCACATCGCAA</td>
<td>Syncytin-1 fwd</td>
</tr>
<tr>
<td>JF113</td>
<td>GCCACTTTAACCGACCTGG</td>
<td>Syncytin-1 rev</td>
</tr>
<tr>
<td>Act-F</td>
<td>CGACAGATGCAAGAGG</td>
<td>Actin fwd</td>
</tr>
<tr>
<td>Act-R</td>
<td>GTACTTGCGCTAGGAGG</td>
<td>Actin rev</td>
</tr>
<tr>
<td>GUSB-F</td>
<td>AGAGTGGGTGCTGAGGATG</td>
<td>GUSB fwd</td>
</tr>
<tr>
<td>GUSB-R</td>
<td>CCCTCATGCTCTAGCGTGTC</td>
<td>GUSB rev</td>
</tr>
<tr>
<td>---------</td>
<td>----------------------</td>
<td>----------</td>
</tr>
<tr>
<td>1065</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>