ANTIVIRAL ACTIVITY AND EVOLUTION OF SUPPRESSYN, A HUMAN PLACENTAL PROTEIN OF RETROVIRAL ORIGIN

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

John Anthony Frank

May 2020

ANTIVIRAL ACTIVITY AND EVOLUTION OF SUPPRESSYN, A HUMAN PLACENTAL PROTEIN OF RETROVIRAL ORIGIN

John Anthony Frank, Ph. D.
Cornell University 2020

Viruses circulating in non-human populations have the potential to infect humans in a process defined as zoonosis. Zoonotic infections can dramatically impact human health and the evolution of human immune factors. Endogenous retroviruses (ERV), which are remnants of ancestral germline insertions, provide a reservoir of protein-coding material with the potential to be domesticated for host cellular functions. ERV-derived envelope (env) proteins have been reported to confer resistance to exogenous retroviral infection in several vertebrates. While previous studies have shown ERV*env* restrict exogenous retroviral infection in non-human organisms, there is no direct evidence of human ERV *env* conferring resistance to extant retroviruses. We hypothesize that a subset of HERV *env* may function as antiviral factors against potentially zoonotic retroviruses. To address this hypothesis, we investigated a truncated and placentally expressed human ERV*env*, *Suppressyn* (*SUPYN*). *SUPYN* binds the cell surface amino acid transporter ASCT2, which is the target receptor for diverse mammalian retroviruses dubbed the RD114 and Type-D retrovirus (RDR) interference group. RDRs are known to circulate in Old World monkeys as well as domestic cats and can infect human cells. Here we report *SUPYN* expression initiates in the human preimplantation embryo and persists through human placental development. We show SUPYN is necessary and sufficient to restrict RDRenv mediated cell entry. Our evolutionary sequence analyses indicate *SUPYN* was acquired in the common ancestor of Catarrhine primates and preserved by

natural selection in Apes, where its antiviral activity is conserved. Our data suggest *SUPYN* can protect the developing fetus from zoonotic retroviral infection and potential invasion of the nascent germline. SUPYN represents the first example of a human virus-derived protein with antiviral activity against extant exogenous viruses and implies that our genomes may harbor further virus-derived genes with antiviral activity.

BIOGRAPHICAL SKETCH

John Frank obtained his B.S. Degree in Biology with Minors in Chemistry and Philosophy from Linfield College, located in McMinnville Oregon. During this time, John conducted undergraduate research with Anne Kruchten, PhD studying how cortactin phosphorylation state affects cytoskeleton remodeling within the context of Ewing Sarcoma cell motility. After completing his degree, John entered the Molecular Biology program at the University of Utah to begin his PhD. There he joined the Feschotte Lab studying how human endogenous retrovirus derived envelopes contribute to host immunity. John transferred to Cornell with the Feschotte Lab after his sixth year to complete his PhD. Following graduation, John will begin postdoctoral research in the Iwasaki Lab at Yale University studying the evolution and function of host receptors and antiviral factors encoded in mammals.

To my wife, Rosika Frank, and all those who accompanied me on this journey

ACKNOWLEDGMENTS

"It takes a village to raise a child."

- African proverb -

This phrase could equally be applied to the process of training a Doctoral student. There are many people and institutions that helped make this work and my success possible. First, I want to thank Ellen Pritham for the opportunity to rotate in her lab and gain exposure to transposable elements and the Feschotte Lab. If not for her kindness and training, I never would have had the opportunity to join the Feschotte lab. This brings me to Cedric Feschotte. I cannot thank Cedric enough for taking me on despite my failures in my previous lab. Cedric provided immeasurable science and communication training, career guidance and empowered me to play to my strengths while helping me work on my weaknesses. I also want to thank the members of the Feschotte and Pritham labs who welcomed me with open arms, helped guide me through this process, and set a scientific standard I continue to aspire to. I especially want to acknowledge colleagues Ray Malfavon-Borja, Manvendra Singh who contributed key data that helped to make this work possible. I also want to thank my incredibly hard-working, enthusiastic undergraduate and rotation students Harrison Cullen, Raphael Kirou, Maia Clare, Mriganka Nerkar and Luc Truong for helping to generate much of the data contained in this thesis.

I also want to thank the University of Utah, including Molecular Biology PhD training program, Human Genetics department, and T32 Microbial Pathogenesis training grant for the training and funding these programs provided. I also owe thanks to the Cornell University Genetic Genomics and Development program for making my transition to this new environment seamless and providing a productive, welcoming environment that helped me to push my work forward. My dissertation committees at both the University of Utah and Cornell provided invaluable feedback that helped improve this work. I also want to thank the National Institutes of Health for funding my work and the numerous excellent collaborators, including, Vicente Planelles, Carolyn Coyne, Welkin Johnson, Laars Aagaard, Hector Aguilar Carreno, Juana Zamora, Jose Garcia-Perez, and John Lis for providing expertise, reagents, experimental support and data.

Beyond the professional contributions described above, I also owe a tremendous debt of gratitude to my lab mates, friends and family for keeping my spirits up and supporting me through the inevitable challenges of graduate school. I really appreciate my past and present colleagues Aurelie Kapusta, Claudia Marquez, John McCormick, Xiaoyu Zhuo, Mitch Lokey, Jainy Thomas, Thomas Carter, Clement Goubert, Julius Judd, Michelle Steitzer, and Manvendra (Manu) Singh. You make every workday a pleasure. Sylvia Chang, Jonathan Well, and Rachel and Kevin Cosby have been a constant source of comfort and support throughout this PhD process. I truly appreciate your friendship and kindness be it during game nights or trips together. My parents Christine and Gregory Frank, and my brother Robert Frank have always supported me and encouraged me to persevere through setbacks and pursue my interests. Finally, I am most thankful to my wife Rosika Frank, for your comfort, support, encouragement and patience. You've made tremendous sacrifices to allow me to pursue my PhD and accompanied me from Linfield College in Oregon, to Salt Lake City, Ithaca and now New Haven. I appreciate all that you do and am blessed to share our life together.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

EVE	Endogenous viral element
ERV	Endogenous retrovirus
Env	Envelope
HERV	Human endogenous retrovirus
MLV	murine leukemia virus
LTR	Long terminal repeat
HIV	Human immunodeficiency virus
EBLN	endogenous bornavirus-like nucleoprotein
DNA	Deoxyribonucleic acid
mRNA	Messenger ribonucleic acid
ALV	Avian leukosis virus
TF	Transcription factor
lncRNA	Long noncoding ribonucleic acid
ESC	Embryonic stem cell
RDR	RD114 and Type-D retrovirus
OWM	Old World monkey
SUPYN	Suppressyn
SARS	Severe acute Respiratory Syndrome
CTB	Cytotrophoblast
STB	Syncytiotrophoblast
EVT	Extravillous trophoblast
SYN	Syncytin
scRNAseq	Single cell RNA sequencing

ATAC-seq	**A**ssay for **T**ransposase-**A**ccessible **C**hromatin using **seq**uencing
DNAse-seq	DNase I hypersensitive sites sequencing
ChIP-seq	Chromatin immunoprecipitation with sequencing
hESC	Human embryonic stem cell
ICM	Inner cell mass
EPI	Epiblast
TE	Trophectoderm
VSVg	Vesicular stomatitis virus glycoprotein
GFP	Green fluorescent protein
SP	Signal peptide
SMRV	Squirrel monkey retrovirus
ECL2	Extracellular loop 2
BLAST	Basic local alignment search tool
ORF	Open reading frame
GTEx	Genotype-tissue expression

CHAPTER 1
CO-OPTION OF ENDOGENOUS VIRAL SEQUENCES FOR HOST CELL FUNCTION[1]

1.1 ABSTRACT

Eukaryotic genomes are littered with sequences of diverse viral origins, termed endogenous viral elements (EVEs). Here we used examples primarily drawn from mammalian endogenous retroviruses to document how the influx of EVEs has provided a source of prefabricated coding and regulatory sequences that were formerly utilized for viral infection and replication, but have been occasionally repurposed for cellular function. While EVE co-option has benefited a variety of host biological functions, there appears to be a disproportionate contribution to immunity and antiviral defense. The mammalian embryo and placenta offer opportunistic routes of viral transmission to the next host generation and as such they represent hotbeds for EVE cooption. Based on these observations, we propose that EVE cooption is initially driven as a mean to mitigate conflicts between host and viruses, which in turn acts as a stepping-stone toward the evolution of cellular innovations serving host physiology and development.

[1] This work is published as "John A. Frank and Cédric Feschotte (2017) Co-option of endogenous viral sequences for host cell function. Current Opinion in Virology" and is reprinted here with permission. The author contributions are as follows: Frank JA and Feschotte C chose the topic of the review, Frank JA conducted the literature research, generated figures, and wrote the manuscript. Feschotte C assisted in writing and manuscript preparation.

1.2 INTRODUCTION

Endogenous viral elements (EVE) are sequences of viral origin that have integrated into the host germline genome and, as a result, become vertically inherited in the host population. Viral endogenization is pervasive across all branches of cellular life resulting in the accumulation of EVEs of diverse origins and varying ages within the genomes of infected species[1-4]. As such, EVEs represent a fossil record of past viral infections that can be harnessed to trace the deep origins of viruses and decipher their intricate co-evolution with their hosts[1-12]. As a source of genetic material added to the host genome, EVEs provide a rich compendium of sequences previously serving viral replication that natural selection can act upon at the level of the host organism to foster the emergence of novel cellular function. Here we review a variety of molecular processes, cellular mechanisms, and biological pathways that appear to have repeatedly benefited from such viral co-option events. We place emphasis on recently described examples involving mammalian EVEs, but certainly the phenomenon of EVE cooption is not restricted to mammals[13-15]. While it is now clear that virtually any major type of virus can be endogenized, most coopted mammalian EVEs derive from endogenous retroviruses (ERVs)[4,16]. This bias reflects in part the fact that ERVs are the most common EVEs in mammals, where they account for ~5-15% of nuclear genome content[2,17,18].

1.3 EVE AS RESTRICTION FACTORS: FIGHTING FIRE WITH FIRE

Antiviral function is a recurrent theme of EVE cooption. When expressed, EVE products can in principle interfere with any step of viral infection, thereby acting as restriction factors. The most direct mechanisms of restriction are those involving direct interactions between EVE-derived peptides with viral or cellular proteins that control virus replication (Figure 1). In multiple vertebrates, ERV-encoded envelope (Env) proteins protect host cells from viral entry by competing with exogenous Env for cell surface receptors, a phenomenon analogous to superinfection resistance[19] (Figure 1A and Figure 2). To date, no human ERV (HERV) Env have been reported to restrict modern exogenous retroviruses. However, a recent ‘paleovirological’ study revealed that a primate-specific env derived from a copy of the HERV-T gammaretrovirus family is capable of restricting an experimentally reconstituted HERV-T Env-mediated infection[20]. These data suggest that the acquisition of this endogenous HERV-T Env gene, which has evolved under functional constraint in the human lineage, may have led to the extinction of the cognate retrovirus infecting our ancestors[20,21]. It cannot be excluded, however, that this HERV-T Env locus has been evolutionary preserved to serve another cellular function distinct from viral restriction[20].

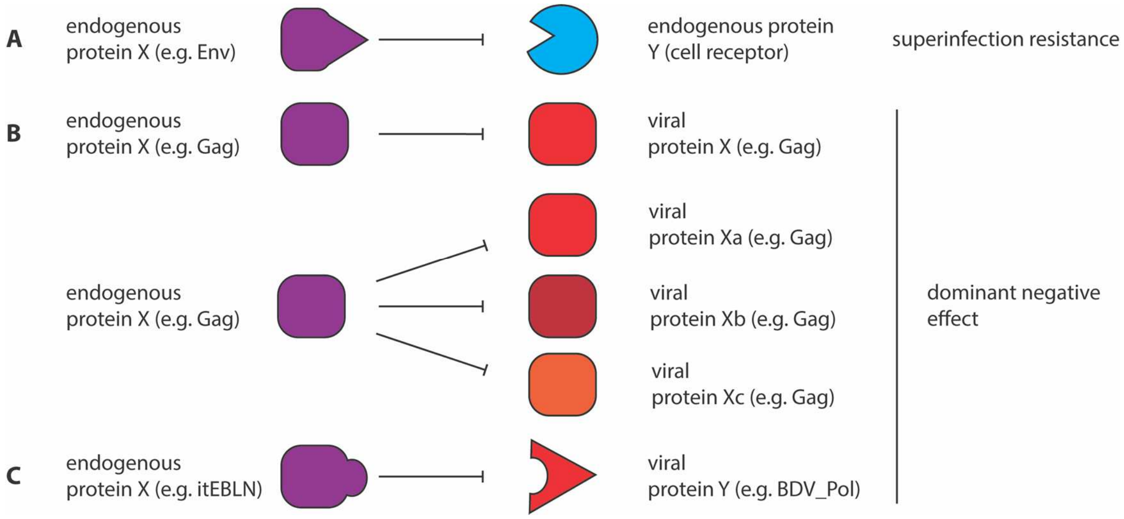
A
endogenous protein X (e.g. Env)
endogenous protein Y (cell receptor)
superinfection resistance
B
endogenous protein X (e.g. Gag)
viral protein X (e.g. Gag)
viral protein Xa (e.g. Gag)
endogenous protein X (e.g. Gag)
viral protein Xb (e.g. Gag)
dominant negative effect
viral protein Xc (e.g. Gag)
C
endogenous protein X (e.g. itEBLN)
viral protein Y (e.g. BDV_Pol)

Figure 1.1: Direct interference of EVE proteins with exogenous viral replication
Coopted EVE proteins can compete with virus replication by binding cellular proteins otherwise bound by exogenous virus (A). Physical interactions between coopted EVE proteins and homologous (B) or non-homologous (C) proteins encoded by exogenous viruses can result in dominant-negative effects on virus replication.

Several ERV-derived Gag proteins are known to interfere with post-entry steps of the infection cycle of exogenous retroviruses. For example, the mouse Fv1 protein restricts murine leukemia virus (MLV) prior to chromosomal integration (Figure 2), by restricting capsid disassembly through direct binding to MLV capsid proteins[22,23]. As Gag proteins accumulate mutations, while remaining expressed, endogenous Gags may also interfere with their exogenous counterparts by exerting trans-dominant negative effects on virus particle assembly or release[24-26] (Figure 1B). This restriction mechanism has been documented for the sheep enJSRV[24] and a similar mechanism involving the production of truncated Gag isoforms is used by the yeast Ty1 long terminal repeat (LTR) retrotransposon, a retroviral-like element, as a form of copy number control[25,26] (Figure 2).

Such direct, conflicting interactions between EVE- and viral-encoded proteins are likely to drive rapid adaptive evolution of both viral and coopted endogenous genes. The resulting allelic diversification of EVE-derived genes may lead to the selection of alleles that expand the range of viruses restricted by this mechanism (Figure 1B). This scenario would explain why *Fv1*, which exhibits a strong signature of diversifying selection in mouse populations, presently restricts murine leukemia virus (MLV), despite being derived from an evolutionary distant lineage of retroviruses (ERV-L)[27,28]. Human-specific HERV-K Gag, which interferes with HIV-1 capsid assembly and release, may currently be serving such a restricting activity[29,30]. These observations indicate that co-option of ERV-derived proteins for viral defense is a common, dynamic, and ongoing evolutionary process.

It is also conceivable that EVE-derived proteins could interfere with exogenous viral replication by interacting with non-homologous viral proteins (Figure 1C). This model is supported by a recent study of endogenous bornavirus-like nucleoproteins (EBLN) encoded in the ground squirrel genome (itEBLN). Cell culture experiments showed that itEBLN expression conferred resistance to human Borna Disease Virus infection by inhibiting viral polymerase activity[31]. These observations may point to a more common theme of EVE cooption for viral defense that merits further investigation.

A recent study of the Mavirus virophage, a small DNA virus that parasitizes the machinery of the giant DNA virus *Cafeteria roenbergensis* virus (CroV) suggests a path through which EVE-mediated antiviral immunity may be established[32]. The authors show that Mavirus integrates within the genome of its marine host protozoan, but lays dormant until transcriptionally activated in response to CroV superinfection. Lysis of cells containing Mavirus particles inhibits CroV replication in neighboring cells thereby enhancing host survival while permitting Mavirus replication[32]. This study illustrates how mutualistic interactions between a virus capable of endogenization and its host may pave the way towards cooption.

1.4 IMMUNE SYSTEMS UNDER EVE INFLUENCE

There is growing evidence that the acquisition of EVEs can shape host immune systems in various ways. Notably EVE-derived noncoding sequences may act as cis-regulatory DNA enhancers of antiviral or pro-inflammatory genes (Figure 3A). The LTRs of mammalian ERVs frequently contain interferon-inducible enhancers that in some instances have been coopted to regulate adjacent host genes encoding critical innate immune factors[33,34]. A need for more efficient immune induction may have provided the selective pressure on ERV LTR sequences, which initially controlled proviral genes, to be maintained in the host population. Over the course of evolution, recombination between proviral LTRs, which results in the loss of internal ERV genes, would have eliminated the potential fitness cost of expressing ERV sequences while still providing the beneficial enhancer effects of the LTR.

EVE-encoded proteins may also regulate the expression of innate immune factors (Figure 2). For instance, the HERV-K-encoded *Rec* protein is expressed in preimplantation embryos where it apparently modulates the translation of many cellular mRNAs (Figure 3B), which may have wide-ranging effects on embryonic function, including antiviral defenses[35]. Consistent with this idea, *Rec* overexpression in embryonic carcinoma cells confers resistance to H1N1 influenza virus infection[35]. Together these observations suggest that the expression of *Rec* during early development may prime embryonic cells for a rapid response to viral infection. In addition to their regulatory effects on immune gene expression, EVE-encoded proteins may also modulate host immunity more directly through processes linked to their viral origins. For instance, ERV-derived Env peptides can be recognized as antigens that effectively shape T cell repertoires and the humoral response[36,37]. In extreme cases, some endogenous Env can even behave as 'superantigens' eliciting non-specific T cell activation[38]. Yet other Env proteins can exert immunosuppressive effects that dampen the immune response[37,39]. While these various immune-modulatory properties have been investigated primarily in the context of ERV overexpression in certain disease states, it is tempting to speculate that some of these activities have coevolved with and become integral components of the host immune response. In all the cases described above, ancestral properties of ERV-encoded proteins appear to have been preserved to varying degrees for the benefit host immunity.

Other potentially protective effects of EVEs include the production of noncoding RNAs that act as adjuvants in antiviral systems (see Figure 2). For example, some EBLNs in rodents and primates appear to have inserted into piwi-interacting RNA (piRNA) genomic clusters and as a result produce piRNA-like RNAs in the testis[40]. Similarly, chickens also exhibit testis-specific piRNA expression, which appear to mostly map to young ALV derived ERV insertions, some of which are known to produce infectious viral particles[41]. It has been proposed that these small RNAs offer some protection to the host by silencing exogenous viral mRNAs[16,40,41]. It has also been reported that elevated levels of ERV-derived RNAs leads to the accumulation of cytosolic nucleic acids, including double-stranded RNAs and complementary DNAs, which are recognized by nucleic acid sensors that direct cells to mount an antiviral and inflammatory response[42-44]. These studies highlight how EVE-derived noncoding RNAs can directly or indirectly enhance antiviral immunity.

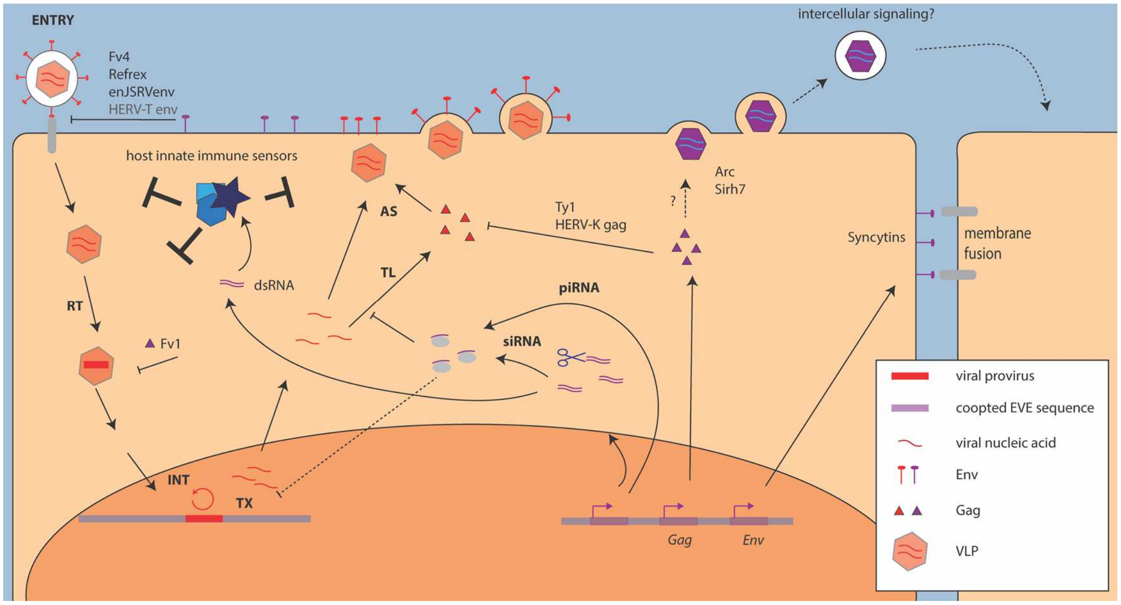

Figure 1.2: Mechanisms of EVE co-option for antiviral immunity and cell physiology

A prototypical retroviral life cycle (shown in red) proceeds through cell entry (ENTRY), reverse transcription (RT), chromosomal integration (INT), proviral transcription (TX), translation (TL) and particle assembly (AS). EVE-encoded proteins and RNAs (shown in purple) can interfere with many steps of virus replication. EVE-encoded proteins may block virus entry (Env), provirus release (Gag), virus genome replication, and capsid assembly (Gag). Small RNAs (piRNAs, siRNAs) derived from EVE loci may also repress virus expression transcriptionally or post-transcriptionally. EVEs can also mediate cell fusion (Env) and may be involved in intercellular signaling (Gag). Viral proteins and nucleic acids can be recognized by host innate immune sensors (shown in blue) resulting in stimulation of the innate immune response.

1.5 ERV CHOREOGRAPHY IN EARLY EMBRYONIC DEVELOPMENT

The early embryo represents a logical battleground for selfish genetic elements, including viruses, as it opens vulnerable routes for vertical and horizontal transmission[45]. In line with this paradigm, many genomics studies have revealed a complex interplay between ERV expression and early embryonic development[46-50]. For example, totipotent 2-cell (2C) mouse embryos are characterized by massive transcriptional activation of MERV-L loci[46,48]. Notably, a trio of recent studies showed that MERV-L activation is driven by the host transcription factor mouse Dux[51-53]. Past the 2C stage, mouse ESCs exhibit markedly reduced MERV-L transcription along with a subsequent peak in ERVK and MaLR expression[54] driven by binding of pluripotency-associated TFs like Nanog and Oct4[54]. This choreography of ERV expression is likely to reflect regulatory pathways hijacked by different ERVs to take advantage of developmental niches that favor their own transcription and propagation[45]. But it raised the possibility that a subset of these elements has been coopted into the regulatory network orchestrating early mouse development. Consistent with this hypothesis, transient siRNA-mediated depletion of a subset of ERVK- and MaLR-derived long noncoding RNAs (lncRNA) highly expressed in mouse ESCs leads to reduced expression of cellular pluripotency markers, suggesting that these lncRNAs exert some form of control over the maintenance of a pluripotent state[54]. Similarly, a recent biochemical study showed that a lncRNA derived from a MERV-L locus, called *LincGET*, is required for *in vitro* embryonic development to proceed beyond the 2C stage[49]. Biochemical experiments and reporter assays suggest that *LincGET* functions as a scaffold for the recruitment of TFs and splicing factors (Figure 3C), some of which are known to be important for embryonic development[49,55].

A strikingly convergent pattern is emerging in human embryonic development involving primate-specific ERVs. Deep RNA sequencing has revealed that the expression of individual HERV families is precisely regulated during early embryonic development[35,51,56]. Notably, DUX4, a human homolog of mouse Dux, appears to be a crucial regulator of HERV-L LTR transcription in 4-cell-stage embryos[51,53]. Hundreds of ape-specific HERV-H elements are also transcriptionally activated by pluripotency TFs in human ESCs[57-60]. Knockdown experiments indicate that HERV-H transcript levels positively correlate with the expression of pluripotency factors and the 'stemness' of certain embryonic cell subpopulationson[57,61,62]. Recent studies of the HERV-H-derived lncRNA *lnc-RoR*[63,64] and of another lncRNA called *HPAT5*[65] derived from a distinct HERV family revealed that both lncRNAs, despite their distinct evolutionary origins, act as miRNA sponges (Figure 3D) to dampen miRNA-mediated translation repression of Nanog and other TFs. These results establish a mechanistic framework to understand how the levels of HERV-derived lncRNAs modulate the pluripotency of ESCs.

The data summarized above suggest that the finely tuned, stage-specific transcriptional activities of human and mouse ERVs may have been co-opted to orchestrate early embryonic development through cis- and trans-regulatory mechanisms. However, more work is needed to test whether these regulatory activities have become truly indispensable for proper embryonic development or are merely relics of selfish manipulations that facilitated ERV propagation.

1.6 THE PLACENTA AS A HOTSPOT OF EVE

At the interface between maternal and fetal tissues, the placenta must mediate nutrient exchange between mother and fetus, protect the fetus from infection by maternally carried pathogens, while avoiding stimulation of the maternal immune system. The trophoblast layer of the placenta exhibits globally elevated EVE expression, which is potentiated by a seemingly general hypomethylation of repetitive DNA[66-68]. In addition, the LTRs of several ERV families exhibit placenta-specific enhancer activity[69,70] (Figure 3A). Together these properties open the door for the cooption of certain LTRs to drive novel adaptive pattern of host gene expression. A recently described example is a primate-specific HERVP71A-LTR that functions as an enhancer for *HLA-G* expression in human extravillous trophoblasts, which confers maternal immune tolerance to the developing placenta by inhibiting natural killer cell-mediated cytotoxicity[70,71].

The frequent transcriptional activity of EVEs in the placenta may also facilitate the cooption of some of their gene products to foster the remarkable anatomical diversification of this organ. A classic example is provided by the *s*yncytins, which are endogenous retroviral Env genes highly expressed in the placenta that have been coopted in diverse mammals[72,73]. Syncytins typically preserve the fusogenic activity of the ancestral Env, and genetic studies of mouse syncytins have established that this activity is essential for the formation of the bi-layered syncytiotrophoblast characteristic of the murid placenta[72,74,75] (Figure 2). Interestingly, multiple syncytins have been independently acquired from various ERVs in several mammalian lineages, suggesting Env co-option as a recurrent force driving the evolution of placentation[72,75]. Interestingly, the fusogenic properties of syncytins also appear to have been harnessed to support sex-specific muscle development because knockout of syncytin B in mouse results in reduced myoblast fusion and muscle mass in males[76] (Figure 2). These data illustrate how the biochemical properties of viral envelopes have been recycled multiple times during evolution to serve mammalian development.

Gag proteins from ancient LTR retrotransposons have also been repurposed for placenta biology in both marsupial and eutherian mammals[77]. Mouse knockout studies indicate that at least three ancient Gag genes derived from distinct retrotransposon families, *Peg10*, *Peg11*, and *Sirh7*, are required for successful completion of pregnancy[78-82]. Though biochemical studies of these Gag-derived proteins are sparse, current evidence suggests that they have distinct, non-redundant cellular functions[83-87]. This is not unexpected because retroviral and retrotransposon Gag proteins exert a variety of biochemical functions, including complex nucleic acid-, protein-, and lipid-binding activities[88-90]. It is therefore possible that the sole common factor driving co-option of these ancient Gags in placenta may have been placenta-specific expression of these genes. Interestingly, two of these genes (*Peg10*, *Peg11*) are only expressed from the paternal allele, yet reside in different regions of the genome – suggesting a predisposition for genomic imprinting and/or that their cooption was driven by parental conflict[91,92].

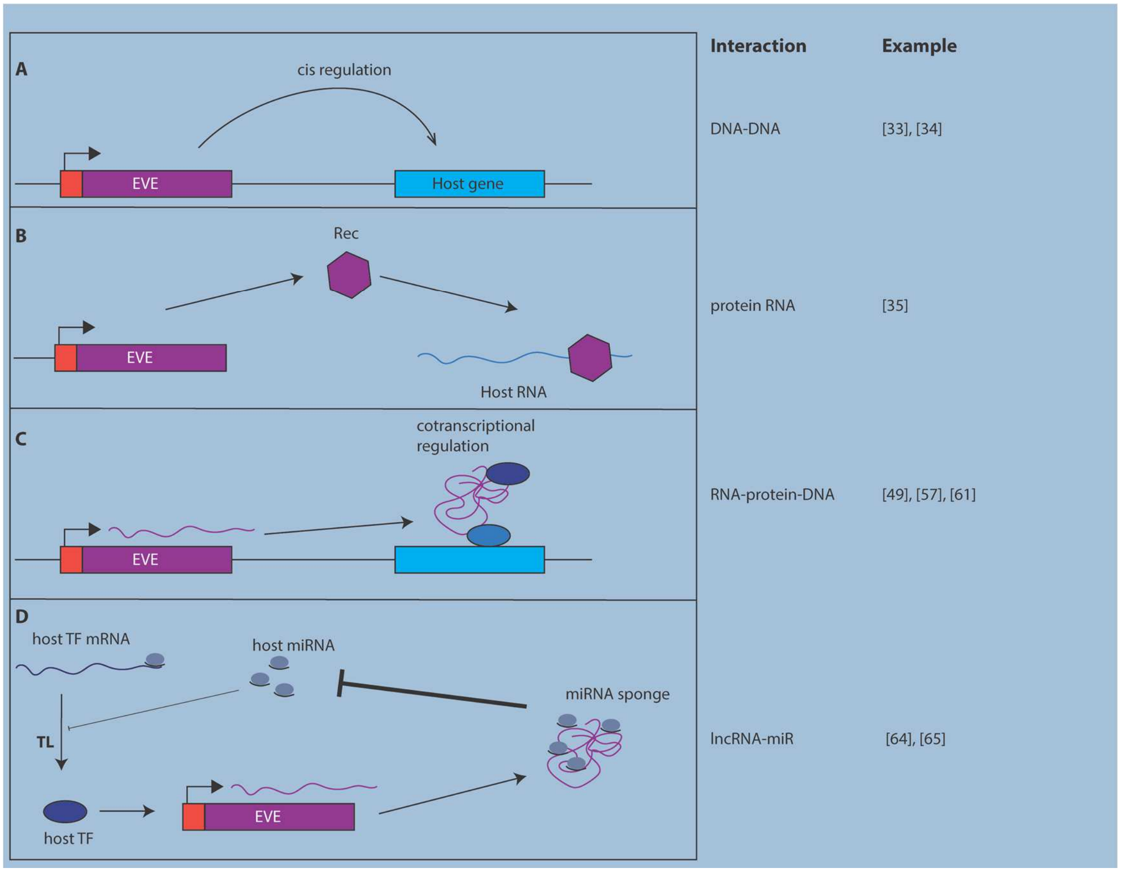

Figure 1.3: Coopted EVEs affect host gene expression by diverse mechanisms
(A) EVE sequences may function as cis-regulatory DNA elements such as enhancers or promoters. (B) EVE-derived lncRNAs can also affect gene expression by acting as co-transcriptional regulators (C) or miRNA sponges (D). EVE-encoded proteins may also regulate gene expression. For instance, Rec and Gag proteins may bind to and modulate host mRNA stability, localization, or translation.

1.7 EVE COOPTED FOR BRAIN FUNCTION

Whereas most coopted EVEs tend to be derived from younger elements, several ancient retrotransposon-derived Gag proteins appear to have contributed to the evolution of the mammalian brain[93-95]. In particular, *Arc* has emerged as a significant player in memory formation and brain development[96,97]. Molecular studies indicate Arc regulates glutamate receptor turnover, a process key to the regulation of synaptic plasticity[94,98]. Additionally, Arc plays a role in synapse pruning during brain development[97]. Far less is known about *Sirh11*, another Gag-derived gene that is strongly conserved across eutherians and highly expressed in the brain[95,99]. Knockout of *Sirh11* in mice has revealed behavioral alterations that may be explained by reduced extracellular noradrenaline levels in the prefrontal cortex[95]. Thus, like *Arc, Sirh11* appears to play a role in neuronal signal transmission. While it is unclear what property these Gag-derived proteins share, it is likely that ancestral activities typical of Gag proteins, such as membrane binding or capsid assembly, may have been repurposed for cellular processes serving brain function.

1.8 OUTLOOK

The viral life cycle is intimately intertwined with cell physiology because virus replication is inherently dependent on the cell's machinery and function. Consequently, viruses have established complex interactions with host cellular factors, often involving direct physical interactions. The endogenization of viral sequences offers an opportunity for these activities to be deployed in a different cellular context, which may occasionally benefit host fitness leading to their fixation and cooption. Indeed, mechanistic studies of coopted EVEs have revealed that their functional activities are often directly descended from their ancestral viral sequences. For instance, the physical binding of cellular factors by coopted EVE-encoded proteins, such as *Env*[72,76] and *Gag*[25,26] can frequently be traced to ancestral protein interaction domains pre-existing in the viral proteins. Likewise, coopted EVE-encoded regulatory sequences are typically derived from ancestral TF binding sites that were presumably used formerly by the virus to promote expression of their own genes[33,69,70,100]. This model does not preclude that some host-EVE adaptive interfaces evolve *de novo* through sequence modification and fortuitous interactions. The pairing of EVE-derived lncRNA with a host-encoded miRNA might represent such a fortuitous interaction that could have provided an initial selective advantage to the host, and possibly also to the virus, as a mechanism to dampen viral expression. Regardless of their origins, any emerging host-EVE interaction that mitigates the conflict between cell and virus is predicted to promote the fixation, retention, and diversification of an EVE[32]. In turn, this cascade might facilitate the emergence of novel adaptive contributions from the coopted EVE sequence. Such a steppingstone model might explain why some transitions from viral to cellular functions (e.g. Syncytins, LTRs, Fv1)[28,33,72,101] have occurred repeatedly during evolution to establish seemingly redundant or convergent organismal function.

REFERENCES

1. Feschotte, C. & Gilbert, C. Endogenous viruses: insights into viral evolution and impact on host biology. *Nat. Rev. Genet.* **13,** 283–296 (2012).
2. Johnson, W. E. Endogenous Retroviruses in the Genomics Era. *Annu Rev Virol* **2,** 135–159 (2015).
3. Metegnier, G. *et al.* Comparative paleovirological analysis of crustaceans identifies multiple widespread viral groups. *Mob DNA* **6,** 16 (2015).
4. Aiewsakun, P. & Katzourakis, A. Endogenous viruses: Connecting recent and ancient viral evolution. *Virology* **479-480,** 26–37 (2015).
5. Gilbert, C. & Feschotte, C. Genomic fossils calibrate the long-term evolution of hepadnaviruses. *PLoS Biol.* **8,** e1000495 (2010).
6. Niewiadomska, A. M. & Gifford, R. J. The extraordinary evolutionary history of the reticuloendotheliosis viruses. *PLoS Biol.* **11,** e1001642 (2013).
7. Aswad, A. & Katzourakis, A. Convergent capture of retroviral superantigens by mammalian herpesviruses. *Nat Commun* **6,** 8299 (2015).
8. Blanc, G., Gallot-Lavallée, L. & Maumus, F. Provirophages in the Bigelowiella genome bear testimony to past encounters with giant viruses. *Proc. Natl. Acad. Sci. U.S.A.* **112,** E5318–26 (2015).
9. Hayward, A., Cornwallis, C. K. & Jern, P. Pan-vertebrate comparative genomics unmasks retrovirus macroevolution. *Proc. Natl. Acad. Sci. U.S.A.* **112,** 464–469 (2015).
10. Han, G.-Z. & Worobey, M. A primitive endogenous lentivirus in a colugo: insights into the early evolution of lentiviruses. *Molecular Biology and Evolution* **32,** 211–215 (2015).
11. Diehl, W. E., Patel, N., Halm, K. & Johnson, W. E. Tracking interspecies transmission and long-term evolution of an ancient retrovirus using the genomes of modern mammals. *Elife* **5,** e12704 (2016).
12. Aiewsakun, P. & Katzourakis, A. Marine origin of retroviruses in the early Palaeozoic Era. *Nat Commun* **8,** 13954 (2017).

13. Malik, H. S. & Henikoff, S. Positive selection of Iris, a retroviral envelope-derived host gene in Drosophila melanogaster. *PLoS Genet* **1,** e44 (2005).

14. Sinzelle, L., Carradec, Q., Paillard, E., Bronchain, O. J. & Pollet, N. Characterization of a Xenopus tropicalis endogenous retrovirus with developmental and stress-dependent expression. *J. Virol.* **85,** 2167–2179 (2011).

15. Henzy, J. E., Gifford, R. J., Kenaley, C. P. & Johnson, W. E. An Intact Retroviral Gene Conserved in Spiny-Rayed Fishes for over 100 My. *Molecular Biology and Evolution* **34,** 634–639 (2017).

16. Honda, T. & Tomonaga, K. Endogenous non-retroviral RNA virus elements evidence a novel type of antiviral immunity. *Mob Genet Elements* **6,** e1165785 (2016).

17. Dewannieux, M. & Heidmann, T. Endogenous retroviruses: acquisition, amplification and taming of genome invaders. *Current Opinion in Virology* **3,** 646–656 (2013).

18. Zhuo, X., Rho, M. & Feschotte, C. Genome-Wide Characterization of Endogenous Retroviruses in the Bat Myotis lucifugus Reveals Recent and Diverse Infections. *J. Virol.* **87,** 8493–8501 (2013).

19. Malfavon-Borja, R. & Feschotte, C. Fighting Fire with Fire: Endogenous Retrovirus Envelopes as Restriction Factors. *J. Virol.* **89,** 4047–4050 (2015).

20. Blanco-Melo, D., Gifford, R. J. & Bieniasz, P. D. Co-option of an endogenous retrovirus envelope for host defense in hominid ancestors. *Elife* **6,** 11 (2017).

21. de Parseval, N., Lazar, V., Casella, J.-F., Bénit, L. & Heidmann, T. Survey of human genes of retroviral origin: identification and transcriptome of the genes with coding capacity for complete envelope proteins. *J. Virol.* **77,** 10414–10422 (2003).

22. Hilditch, L. *et al.* Ordered assembly of murine leukemia virus capsid protein on lipid nanotubes directs specific binding by the restriction factor, Fv1. *Proc. Natl. Acad. Sci. U.S.A.* **108,** 5771–5776 (2011).

23. Goldstone, D. C. *et al.* Structural studies of postentry restriction factors reveal antiparallel dimers that enable avid binding to the HIV-1 capsid lattice. *Proc. Natl. Acad. Sci. U.S.A.* **111,** 9609–9614 (2014).

24. Mura, M. *et al.* Late viral interference induced by transdominant Gag of an endogenous retrovirus. *Proc Natl Acad Sci USA* **101,** 11117–11122 (2004).

25. Saha, A. *et al.* A trans-dominant form of Gag restricts Ty1 retrotransposition and mediates copy number control. *J. Virol.* **89,** 3922–3938 (2015).

26. Nishida, Y. *et al.* Ty1 retrovirus-like element Gag contains overlapping restriction factor and nucleic acid chaperone functions. *Nucleic Acids Res.* **43,** 7414–7431 (2015).

27. Yan, Y., Buckler-White, A., Wollenberg, K. & Kozak, C. A. Origin, antiviral function and evidence for positive selection of the gammaretrovirus restriction gene Fv1 in the genus Mus. *Proc. Natl. Acad. Sci. U.S.A.* **106,** 3259–3263 (2009).

28. Yap, M. W., Colbeck, E., Ellis, S. A. & Stoye, J. P. Evolution of the retroviral restriction gene Fv1: inhibition of non-MLV retroviruses. *PLoS Pathog* **10,** e1003968 (2014).

29. Monde, K. *et al.* Molecular mechanisms by which HERV-K Gag interferes with HIV-1 Gag assembly and particle infectivity. *Retrovirology* **14,** 27 (2017).

30. Monde, K., Contreras-Galindo, R., Kaplan, M. H., Markovitz, D. M. & Ono, A. Human endogenous retrovirus K Gag coassembles with HIV-1 Gag and reduces the release efficiency and infectivity of HIV-1. *J. Virol.* **86,** 11194–11208 (2012).

31. Fujino, K., Horie, M., Honda, T., Merriman, D. K. & Tomonaga, K. Inhibition of Borna disease virus replication by an endogenous bornavirus-like element in the ground squirrel genome. *Proc. Natl. Acad. Sci. U.S.A.* **111,** 13175–13180 (2014).

32. Fischer, M. G. & Hackl, T. Host genome integration and giant virus-induced reactivation of the virophage mavirus. *Nature* **540,** 288–291 (2016).

33. Chuong, E. B., Elde, N. C. & Feschotte, C. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. *Science* **351,** 1083–1087 (2016).

34. Manghera, M., Ferguson-Parry, J., Lin, R. & Douville, R. N. NF-κB and IRF1 Induce Endogenous Retrovirus K Expression via Interferon-Stimulated Response Elements in Its 5' Long Terminal Repeat. *J. Virol.* **90,** 9338–9349 (2016).

35. Grow, E. J. *et al.* Intrinsic retroviral reactivation in human preimplantation embryos and pluripotent cells. *Nature* **522,** 221–225 (2015).

36. Ebert, P. J. R., Jiang, S., Xie, J., Li, Q.-J. & Davis, M. M. An endogenous positively selecting peptide enhances mature T cell responses and becomes an autoantigen in the absence of microRNA miR-181a. *Nat. Immunol.* **10,** 1162–1169 (2009).

37. Kassiotis, G. & Stoye, J. P. Immune responses to endogenous retroelements: taking the bad with the good. *Nature Reviews Immunology* **16,** 207–219 (2016).

38. Tai, A. K. *et al.* Murine Vbeta3+ and Vbeta7+ T cell subsets are specific targets for the HERV-K18 Env superantigen. *J. Immunol.* **177,** 3178–3184 (2006).

39. Schlecht-Louf, G. *et al.* Retroviral infection in vivo requires an immune escape virulence factor encrypted in the envelope protein of oncoretroviruses. *Proc Natl Acad Sci USA* **107,** 3782–3787 (2010).

40. Parrish, N. F. *et al.* piRNAs derived from ancient viral processed pseudogenes as transgenerational sequence-specific immune memory in mammals. *RNA* **21,** 1691–1703 (2015).

41. Sun, Y. H. *et al.* Domestic chickens activate a piRNA defense against avian leukosis virus. *Elife* **6,** 8634 (2017).

42. Zeng, M. *et al.* MAVS, cGAS, and endogenous retroviruses in T-independent B cell responses. *Science* **346,** 1486–1492 (2014).

43. Chiappinelli, K. B. *et al.* Inhibiting DNA Methylation Causes an Interferon Response in Cancer via dsRNA Including Endogenous Retroviruses. *Cell* **162,** 974–986 (2015).

44. Roulois, D. *et al.* DNA-Demethylating Agents Target Colorectal Cancer Cells by Inducing Viral Mimicry by Endogenous Transcripts. *Cell* **162,** 961–973 (2015).

45. Haig, D. Transposable elements: Self-seekers of the germline, team-players of the soma. *Bioessays* **38,** 1158–1166 (2016).

46. Macfarlan, T. S. *et al.* Embryonic stem cell potency fluctuates with endogenous retrovirus activity. *Nature* **487,** 57–63 (2012).

47. Maksakova, I. A. *et al.* Distinct roles of KAP1, HP1 and G9a/GLP in silencing of the two-cell-specific retrotransposon MERVL in mouse ES cells. *Epigenetics Chromatin* **6,** 15–16 (2013).

48. Ishiuchi, T. *et al.* Early embryonic-like cells are induced by downregulating replication-dependent chromatin assembly. *Nat. Struct. Mol. Biol.* **22,** 662–671 (2015).

49. Wang, J. *et al.* A novel long intergenic noncoding RNA indispensable for the cleavage of mouse two-cell embryos. *EMBO Rep.* **17,** 1452–1470 (2016).

50. Choi, Y. J. *et al.* Deficiency of microRNA miR-34a expands cell fate potential in pluripotent stem cells. *Science* **355,** eaag1927 (2017).

51. Hendrickson, P. G. *et al.* Conserved roles of mouse DUX and human DUX4 in activating cleavage-stage genes and MERVL/HERVL retrotransposons. *Nature Publishing Group* **49,** 925–934 (2017).

52. Whiddon, J. L., Langford, A. T., Wong, C.-J., Zhong, J. W. & Tapscott, S. J. Conservation and innovation in the DUX4-family gene network. *Nature Publishing Group* **49,** 935–940 (2017).

53. De Iaco, A. *et al.* DUX-family transcription factors regulate zygotic genome activation in placental mammals. *Nature Publishing Group* **49,** 941–945 (2017).

54. Fort, A. *et al.* Deep transcriptome profiling of mammalian stem cells supports a regulatory role for retrotransposons in pluripotency maintenance. *Nature Publishing Group* **46,** 558–566 (2014).

55. Zhou, W. *et al.* Far Upstream Element Binding Protein Plays a Crucial Role in Embryonic Development, Hematopoiesis, and Stabilizing Myc Expression Levels. *Am. J. Pathol.* **186,** 701–715 (2016).

56. Göke, J. *et al.* Dynamic transcription of distinct classes of endogenous retroviral elements marks specific populations of early human embryonic cells. *Cell Stem Cell* **16,** 135–141 (2015).

57. Ng, S.-Y., Johnson, R. & Stanton, L. W. Human long non-coding RNAs promote pluripotency and neuronal differentiation by association with chromatin modifiers and transcription factors. *EMBO J.* **31,** 522–533 (2012).

58. Santoni, F. A., Guerra, J. & Luban, J. HERV-H RNA is abundant in human embryonic stem cells and a precise marker for pluripotency. *Retrovirology* **9,** 111 (2012).

59. Wang, J. *et al.* Primate-specific endogenous retrovirus-driven transcription defines naive-like stem cells. *Nature* **516,** 405–409 (2014).

60. Izsvák, Z., Wang, J., Singh, M., Mager, D. L. & Hurst, L. D. Pluripotency and the endogenous retrovirus HERVH: Conflict or serendipity? *Bioessays* **38,** 109–117 (2016).

61. Lu, X. *et al.* The retrovirus HERVH is a long noncoding RNA required for human embryonic stem cell identity. *Nat. Struct. Mol. Biol.* **21,** 423–425 (2014).

62. Ohnuki, M. *et al.* Dynamic regulation of human endogenous retroviruses mediates factor-induced reprogramming and differentiation potential. *Proc. Natl. Acad. Sci. U.S.A.* **111,** 12426–12431 (2014).

63. Loewer, S. *et al.* Large intergenic non-coding RNA-RoR modulates reprogramming of human induced pluripotent stem cells. *Nature Publishing Group* **42,** 1113–1117 (2010).

64. Wang, Y. *et al.* Endogenous miRNA sponge lincRNA-RoR regulates Oct4, Nanog, and Sox2 in human embryonic stem cell self-renewal. *Dev. Cell* **25,** 69–80 (2013).

65. Durruthy-Durruthy, J. *et al.* The primate-specific noncoding RNA HPAT5 regulates pluripotency during human preimplantation development and nuclear reprogramming. *Nature Publishing Group* **48,** 44–52 (2016).

66. Chapman, V., Forrester, L., Sanford, J., Hastie, N. & Rossant, J. Cell lineage-specific undermethylation of mouse repetitive DNA. *Nature* **307,** 284–286 (1984).

67. Sanford, J. P., Chapman, V. M. & Rossant, J. DNA methylation in extraembryonic lineages of mammals. *Trends Genet.* **1,** 89–93 (1985).

68. Chuong, E. B. Retroviruses facilitate the rapid evolution of the mammalian placenta. *Bioessays* **35,** 853–861 (2013).

69. Chuong, E. B., Rumi, M. A. K., Soares, M. J. & Baker, J. C. Endogenous retroviruses function as species-specific enhancer elements in the placenta. *Nature Publishing Group* **45,** 325–329 (2013).

70. Ferreira, L. M. R. *et al.* A distant trophoblast-specific enhancer controls HLA-G expression at the maternal–fetal interface. *Proc Natl Acad Sci USA* **113,** 5364–5369 (2016).

71. Rouas-Freiss, N., Marchal, R. E., Kirszenbaum, M., Dausset, J. & Carosella, E. D. The α1 domain of HLA-G1 and HLA-G2 inhibits cytotoxicity induced by natural killer cells: Is HLA-G the public ligand for natural killer cell inhibitory receptors? *Proc Natl Acad Sci USA* **94,** 5249–5254 (1997).

72. Lavialle, C. *et al.* Paleovirology of 'syncytins', retroviral env genes exapted for a role in placentation. *Philosophical Transactions of the Royal Society B: Biological Sciences* **368,** 20120507–20120507 (2013).

73. Cornelis, G. *et al.* Retroviral envelope gene captures and syncytin exaptation for placentation in marsupials. *Proc. Natl. Acad. Sci. U.S.A.* **112,** E487–96 (2015).

74. Heidmann, A. D. C. L. T., Lavialle, C. & Heidmann, T. From ancestral infectious retroviruses to bona fide cellular genes: Role of the captured syncytins in placentation. *Placenta* **33,** 663–671 (2012).

75. Vernochet, C. *et al.* The captured retroviral envelope syncytin-A and syncytin-B genes are conserved in the Spalacidae together with hemotrichorial placentation. *Biology of Reproduction* **91,** 148 (2014).

76. Redelsperger, F. *et al.* Genetic Evidence That Captured Retroviral Envelope syncytins Contribute to Myoblast Fusion and Muscle Sexual Dimorphism in Mice. *PLoS Genet* **12,** e1006289 (2016).

77. Kaneko-Ishino, T. & Ishino, F. The role of genes domesticated from LTR retrotransposons and retroviruses in mammals. *Front Microbiol* **3,** 262 (2012).

78. Sekita, Y. *et al.* Role of retrotransposon-derived imprinted gene, Rtl1, in the feto-maternal interface of mouse placenta. *Nature Publishing Group* **40,** 243–248 (2008).

79. Ono, R. *et al.* Deletion of Peg10, an imprinted gene acquired from a retrotransposon, causes early embryonic lethality. *Nat. Genet.* **38,** 101–106 (2006).

80. Naruse, M. *et al.* Sirh7/Ldoc1 knockout mice exhibit placental P4 overproduction and delayed parturition. *Development* **141,** 4763–4771 (2014).

81. Ito, M. *et al.* A trans-homologue interaction between reciprocally imprinted miR-127 and Rtl1 regulates placenta development. *Development* **142,** 2425–2430 (2015).

82. Koppes, E., Himes, K. P. & Chaillet, J. R. Partial Loss of Genomic Imprinting Reveals Important Roles for Kcnq1 and Peg10 Imprinted Domains in Placental Development. *PLoS ONE* **10,** e0135202 (2015).

83. Okabe, H. *et al.* Involvement of PEG10 in human hepatocellular carcinogenesis through interaction with SIAH1. *Cancer Res.* **63,** 3043–3048 (2003).

84. Lux, A. *et al.* Human retroviral gag- and gag-pol-like proteins interact with the transforming growth factor-beta receptor activin receptor-like kinase 1. *J. Biol. Chem.* **280,** 8482–8493 (2005).

85. Inoue, M. *et al.* LDOC1, a novel MZF-1-interacting protein, induces apoptosis. *FEBS Lett.* **579,** 604–608 (2005).

86. Chen, H., Sun, M., Liu, J., Tong, C. & Meng, T. Silencing of Paternally Expressed Gene 10 Inhibits Trophoblast Proliferation and Invasion. *PLoS ONE* **10,** e0144845 (2015).

87. Li, X. *et al.* PEG10 promotes human breast cancer cell proliferation, migration and invasion. *Int. J. Oncol.* **48,** 1933–1942 (2016).

88. Campillos, M., Doerks, T., Shah, P. K. & Bork, P. Computational characterization of multiple Gag-like human proteins. *Trends Genet.* **22,** 585–589 (2006).

89. Matreyek, K. A. & Engelman, A. Viral and cellular requirements for the nuclear entry of retroviral preintegration nucleoprotein complexes. *Viruses* **5,** 2483–2511 (2013).

90. Freed, E. O. HIV-1 assembly, release and maturation. *Nat. Rev. Microbiol.* **13,** 484–496 (2015).

91. Haig, D. Retroviruses and the placenta. *Curr. Biol.* **22,** R609–13 (2012).

92. Peters, J. The role of genomic imprinting in biology and disease: an expanding view. *Nat. Rev. Genet.* **15,** 517–530 (2014).

93. Mortelmans, K., Wang-Johanning, F. & Johanning, G. L. The role of human endogenous retroviruses in brain development and function. *APMIS* **124,** 105–115 (2016).

94. Zhang, W. *et al.* Structural basis of arc binding to synaptic proteins: implications for cognitive disease. *Neuron* **86,** 490–500 (2015).

95. Irie, M. *et al.* Cognitive Function Related to the Sirh11/Zcchc16 Gene Acquired from an LTR Retrotransposon in Eutherians. *PLoS Genet* **11,** e1005521 (2015).

96. Plath, N. *et al.* Arc/Arg3.1 Is Essential for the Consolidation of Synaptic Plasticity and Memories. *Neuron* **52,** 437–444 (2006).

97. Mikuni, T. *et al.* Arc/Arg3.1 Is a Postsynaptic Mediator of Activity-Dependent Synapse Elimination in the Developing Cerebellum. *Neuron* **78,** 1024–1035 (2013).

98. Shepherd, J. D. & Bear, M. F. New views of Arc, a master regulator of synaptic plasticity. *Nat Neurosci* **14,** 279–284 (2011).

99. Irie, M., Koga, A., Kaneko-Ishino, T. & Ishino, F. An LTR Retrotransposon-Derived Gene Displays Lineage-Specific Structural and Putative Species-Specific Functional Variations in Eutherians. *Front Chem* **4,** 26 (2016).

100. Flemr, M. *et al.* A retrotransposon-driven dicer isoform directs endogenous small interfering RNA production in mouse oocytes. *Cell* **155,** 807–816 (2013).

101. Esnault, C., Cornelis, G., Heidmann, O. & Heidmann, T. Differential evolutionary fate of an ancestral primate endogenous retrovirus envelope gene, the EnvV syncytin, captured for a function in placentation. *PLoS Genet* **9,** e1003400–12 (2013).

CHAPTER 2

ANTIVIRAL ACTIVITY OF SUPPRESSYN, A HUMAN PLACENTAL PROTEIN COOPTED FROM A RETROVIRUS[2]

2.1 SUMMARY

Viruses circulating in non-human populations have recurrently infected humans in a process known as zoonosis[1]. The human genome may harbor undiscovered genetic factors that restrict zoonoses. Some endogenous retroviruses, which are remnants of ancestral germline infections, can confer protection against viruses circulating in host populations[2-13]. The RD114 and Type-D retrovirus (RDR) interference group includes infectious viruses known to circulate in domestic cats and various Old World monkeys (OWM), but not healthy hominoids[14,15]. However, RDRs can infect humans and a wide range of vertebrate cells in culture, by utilizing the conserved cell surface amino acid transporter ASCT2 as a target receptor[15-20]. Suppressyn (*SUPYN*) is a truncated envelope protein derived from an endogenous retrovirus previously reported to be expressed in the human placenta and binds ASCT2 to modulate placental cell fusion[16,17]. Here we report that *SUPYN* expression initiates in the human preimplantation embryo and is necessary and sufficient to protect human cells against RDR infection. We found that *SUPYN* was acquired in the common ancestor of hominoids and OWM, but preserved by natural selection only in hominoids where its antiviral activity is conserved. Our data suggest *SUPYN* can protect the developing fetus from zoonotic infection and retroviral infiltration of the nascent germline and imply further endogenous virus-derived genes with antiviral properties lay hidden in the human genome.

[2] This chapter is currently under preparation for publication, and will be available on bioRxiv after submission (John A. Frank, Manvendra Singh, Harrison B. Cullen, Raphael A. Kirou, Maia G. Clare,

Carolyn B. Coyne, Cedric Feschotte). JAF developed this project, designed and conducted all experiments, validated evolutionary sequence analyses, analyzed all experimental data. MS performed all gene expression and regulation analyses. HBC and RAK helped perform infection assays and evolutionary sequence analyses. MGC performed evolutionary sequence analyses. CF aided in project definition and manuscript preparation.

2.2 MAIN

Viral zoonosis poses a constant threat to human health and has led to devastating epidemics such as those caused by Influenza[18], HIV[19], Ebola[20], and SARS Coronaviruses[21,22]. Some zoonotic viruses have gained access to new host species by recurrently capturing heterologous glycoproteins that mediate target-cell entry by binding to host cell surface receptors[12,18,22,23]. Over the course of mammalian history, capture of gammaretroviral env, including RDRenv, has led to the emergence of novel viruses capable of jumping between species[12,23]. In fact, the endogenous feline leukemia virus RD114 emerged as result of Baboon endogenous virus *env* (an RDRenv) capture by the Felis catus endogenous retrovirus[24]. RDRenv-mediated infection could pose a serious threat to humans because RDRenv utilize the highly conserved and broadly expressed amino acid transporter ASCT2 (also known as SLC1A5)[15,25,26]. Thus, it is critical to assess whether humans are equipped with mechanisms to protect against RDR zoonosis.

During pregnancy, viral infections can severely impact the developing fetus and potentially result in miscarriage[27,28]. The placenta is a critical barrier to fetal infection and frequently challenged by various pathogens including zoonotic viruses[27,29]. However we still know little about the mechanisms that prevent pathogenic infiltration of the placenta and restrict viral replication throughout pregnancy[30].

Syncytins are endogenous retrovirus *env*-derived genes that were independently co-opted during primate evolution[16,17]. Syncytins are thought to play an essential role in placental development by mediating cytotrophoblast (CTB) cell fusion events required for syncytiotrophoblast (STB) formation, a multinucleated structure that serves as a physical barrier at the fetal-maternal interface[31]. *SUPYN* is another protein derived from an endogenous retroviral *env* reported to be expressed in 1st-3rd trimester placenta predominantly in CTB and extravillous cytotrophoblasts (EVT)[16,17], which mediate invasion of and anchoring to the maternal decidua[30]. SUPYN lacks a transmembrane domain and therefore cannot act as a fusogenic protein. However, previous *in vitro* studies have shown that SUPYN, like SYN1, binds ASCT2 and thereby modulates the fusogenic activity of SYN1[16,17]. Given that endogenous retroviral env are capable of conferring resistance to retroviral infection by a mechanism of receptor interference[3,32,33], we hypothesized that SUPYN confers resistance to RDR infection during human fetal development.

2.2.1 SUPYN EMBRYONIC EXPRESSION IS DRIVEN BY PLURPOTENTY AND PLACENTATION REGULATORY FACTORS

Table 2.1 External data sources

Description	Author	Year	Publicaition	GEO	Seq Platform
scRNAseq	Yan et al.	2013	PMID: 23934149	GSE36552	Illumina HiSeq 2000
	Liu et al.	2018	PMID: 30042384	GSE89497	Illumina HiSeq 4000
	Vento-Tormo et al.	2018	PMID: 30429548	E-MTAB-6701 (see methods)	10X Genomics
ChIPseq	Tsankov	2015	PMID: 25693565	GSE61475	Illumina HiSeq 2000
	Kwak	2019	PMID: 31294776	GSE127288	Illumina HiSeq 2500
	Dunn-Fletcher	2018	PMID: 30231016	GSE118289	Illumina HiSeq 3000
	Krendl	2017	PMID: 29078328	GSE105258	Illumina HiSeq 2500
	Krendl	2017	PMID: 29078328	GSE105081	Illumina NextSeq 500

To characterize when and in which cell types *SUPYN* is expressed during human development, we analyzed publicly available scRNA-seq, ATAC-seq, DNAse-seq and ChIP-seq datasets generated from human preimplantation embryos and human embryonic stem cells (hESC) **(Table 2.1).** We observed *SUPYN* mRNA appears after the onset of embryonic genome activation at the eight-cell stage and peaks in morula **(Fig 2.1a)**[34-36]**.** By blastula formation, *SUPYN* expression persists in the inner cell mass (ICM), epiblast (EPI), ESCs, and in the trophectoderm (TE) which will give rise to the placenta **(Fig 2.1a)**[34-36]. Consistent with this expression pattern, we found that in hESCs the *SUPYN* promoter region is marked by H3K4Me1 and H3K27Ac modified histones, and bound by core pluripotency (Oct4, Nanog, KLF4, SMAD1) and self-renewal (SRF, OTX2) transcription factors **(Fig 2.1b)**[37]. Analyses of ATAC-seq and DNAse-seq datasets generated from human preimplantation embryos indicate the *SUPYN* locus is marked by open chromatin from 2-cell to blastocyst stages[38,39] **(Fig 2.2a)**. Together these data indicate *SUPYN* is robustly expressed throughout early embryonic development and likely activated by pluripotency factors. By contrast, we found no evidence for *SYN1* expression in preimplantation embryos and hESCs **(Fig 2.1a)**.

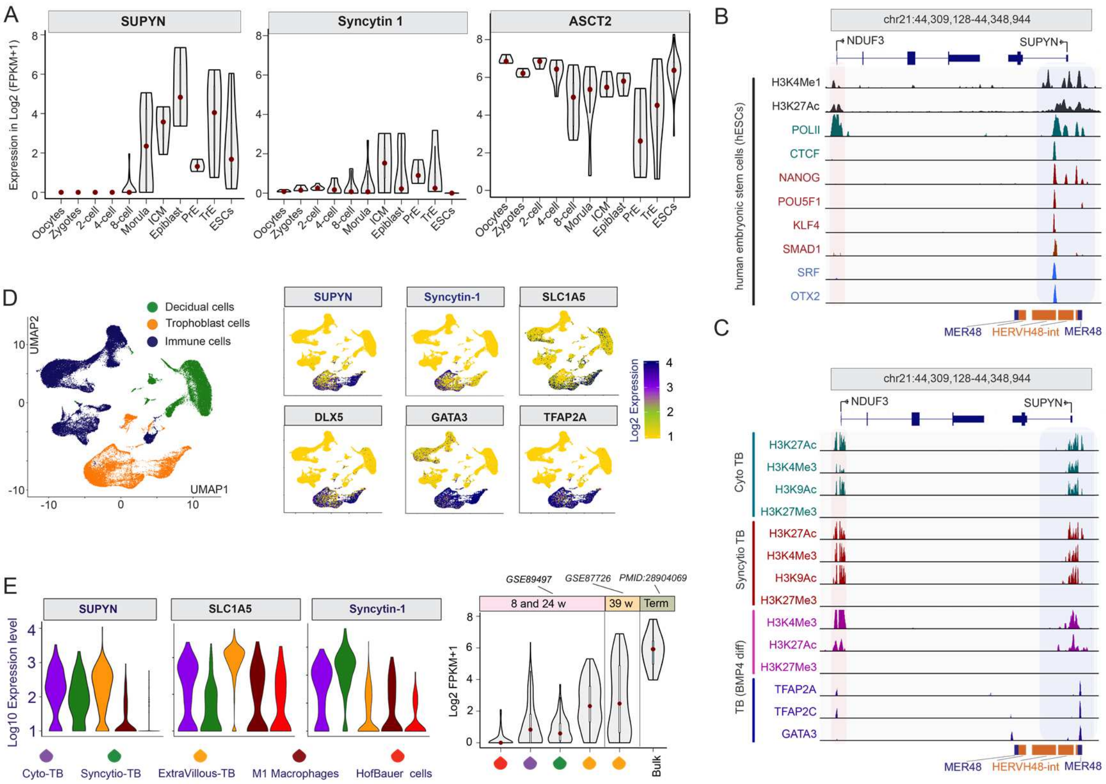

Figure 2.1: Pluripotency and placentation regulatory factor driven SUPYN expression during fetal development.

(a) Violin plots summarizing SUPYN, SYN1 and ASCT2 expression in human preimplantation embryos and ESCs single-cell RNA-seq data. **(b, c)** Genome browser view of the SUPYN locus in hESCs **(b)** and TBs **(c)**. ChIP-seq profiles for H3K27Ac **(b, c),** H3K4Me1, POLII, NANOG, OCT4, KLF4, SMAD1, SRF **(b)**, H3K4Me3, H3K9Ac, H3K27Me3, GATA3, TFAP2A, and TFAP2C **(c)** are shown. Shaded area represents regions of active chromatin. **(d)** UMAP plot of scRNAseq data displaying trophoblast (yellow), decidual (green) and immune (purple) cell identity. Sub-panels display single-cell-level *SUPYN*, *SYN1*, *ASCT2*, *GATA3*, *TFAP2A*, *DLX5* and GATA2 expression at the maternal-fetal interface. **(e, f)** Violin plots denoting single-cell *SUPYN* and *ASCT2* expression in multiple placental-cell lineages **(e)** and at distinct placental development stages **(f)**.

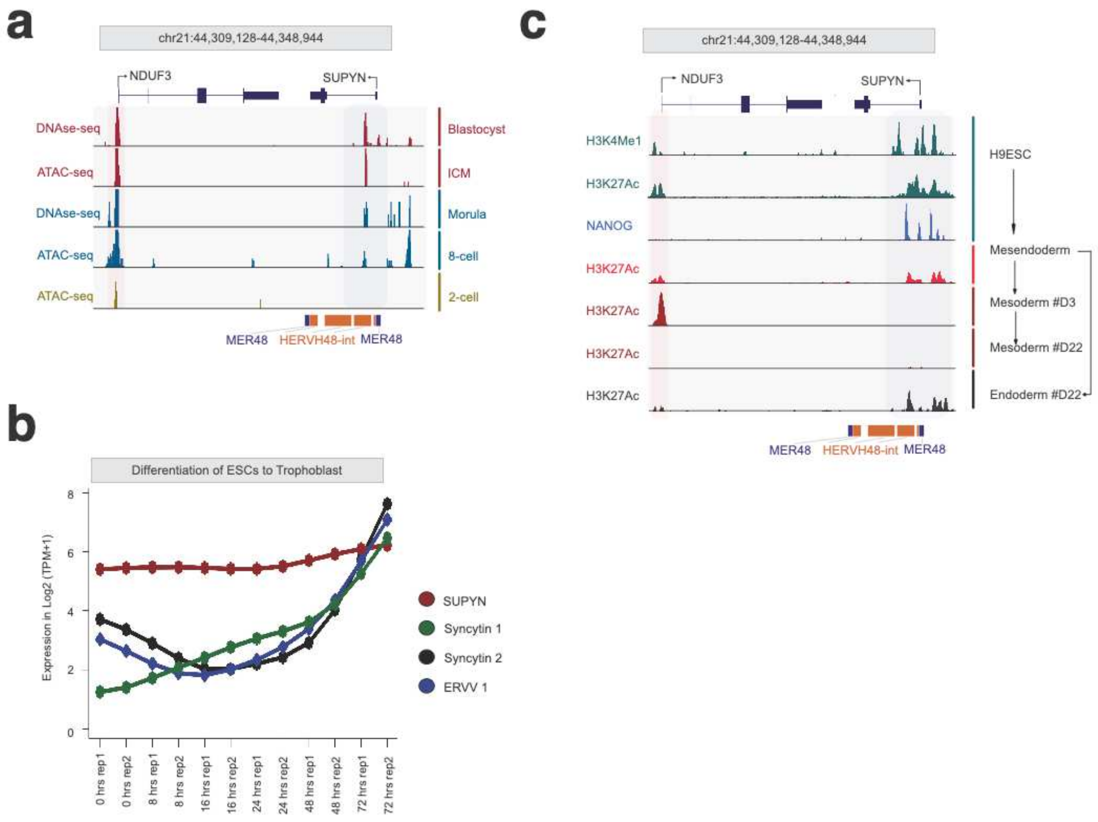

Figure 2.2: SUPYN is constitutively expressed throughout human pluripotency and placentation.
(a) Genome browser view showing ATAC-seq signals and DNAse-seq at the SUPYN locus, including upstream and downstream sequences. Framed region highlights the overlapping peaks at the SUPYN locus. **(b)** Line plot depicts HERVenv gene expression level during BMP4-mediated *in vitro* hESCs to TB differentiation. Time points correspond to cells harvested 8hr, 24hr, 48hr, and 72hr post BMP4 treatment. **(c)** Genome browser view of the SUPYN locus. ChIPseq profiles for NANOG, H3K4me1, and H3K27Ac in ESCs as well as H3K27Ac marks during human ESC to mesoderm and mesendoderm to endoderm differentiation are shown.

To examine *SUPYN* expression throughout placentation, we interrogated publicly available RNA- and ChIP-seq datasets generated from in vitro TB differentiation models[40,41] and placenta explants isolated at multiple developmental stages[42-45]. 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[40] **(Fig 2.1c).** *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 2.1c; Fig 2.2c)**[40,41]. By contrast, expression of other envelope-derived genes *SYN1*, *SYN2*, and *ERVV1*/*V2* is only detectable in differentiated trophoblasts **(Fig 2.2b)**[40]**.** We next mined publicly available scRNA-seq data generated from placenta at multiple developmental stages to examine the cell-type specificity of *SUPYN* expression **(Table 2.1)**[42,43]. After classifying cell clusters based on expression of known markers **(Fig 2.1d, e; Fig 2.3a, b, c)**, we found *SUPYN* expression specifically in the TB lineage **(Fig 2.1d, e, f; Fig 2.3d)**. TB-specific *SUPYN* expression was corroborated by active chromatin marks and binding of TB-specific transcription factors[40,44,45] to the *SUPYN* promoter region **(Fig 2.1c)**. Consistent with previous reports[16,17], *SUPYN* expression was relatively high in CTB and EVT, but also detectable in STB **(Fig 2.1e)**. SUPYN expression in EVT was maintained throughout placental development **(Fig 2.1f)**. Consistent with previous reports[17,46-48], SYN1 expression appears restricted to CTB to STB lineages **(Fig 2.1d, e; Fig 2.3 c, d).** To confirm these transcriptomic observations, we performed immunostaining of 2nd and 3rd trimester placenta with SUPYN antibody. The results indicate SUPYN is widely expressed in STB, and likely cytotrophoblasts within the lumen of 2nd trimester placental villi **(Fig 2.4).** Together these analyses indicate SUPYN is expressed throughout human fetal development and shows only partial overlap with SYN1 expression, which hint at an additional function independent of its proposed role in modulating SYN1-mediated cell fusion during STB development.

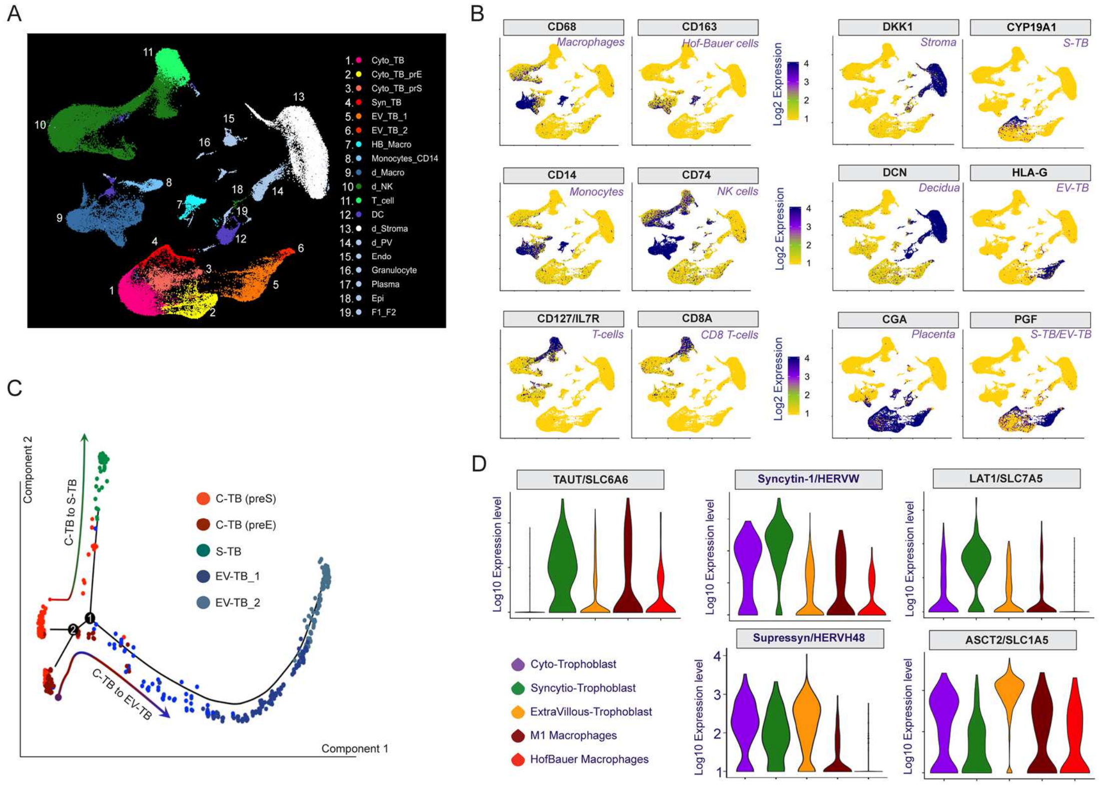

Figure 2.3: Defining lineage-specific SYN1, SUPYN, and ASCT2 expression from placental single-cell transcriptomics

(a) UMAP plot generated from published scRNA-seq data generated from 1st trimester placental explants. Colors denote CTB, STB, EVTB, 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)** Monocle2 single-cell trajectory analysis along an artificial temporal continuum using the top 500 CTB-, STB- and EVTB-defining differentially expressed genes. The transcriptome from each single cell represents a pseudotime point along an artificial time vector denoting progression of CTB to STB and EVTB respectively. **(d)** Violin plots denoting single-cell *SUPYN*, *ASCT2, ASCT1, LAT1,* and *TAUT* expression in multiple placental-cell lineages. *Also see* ***Fig. 2.1f.***

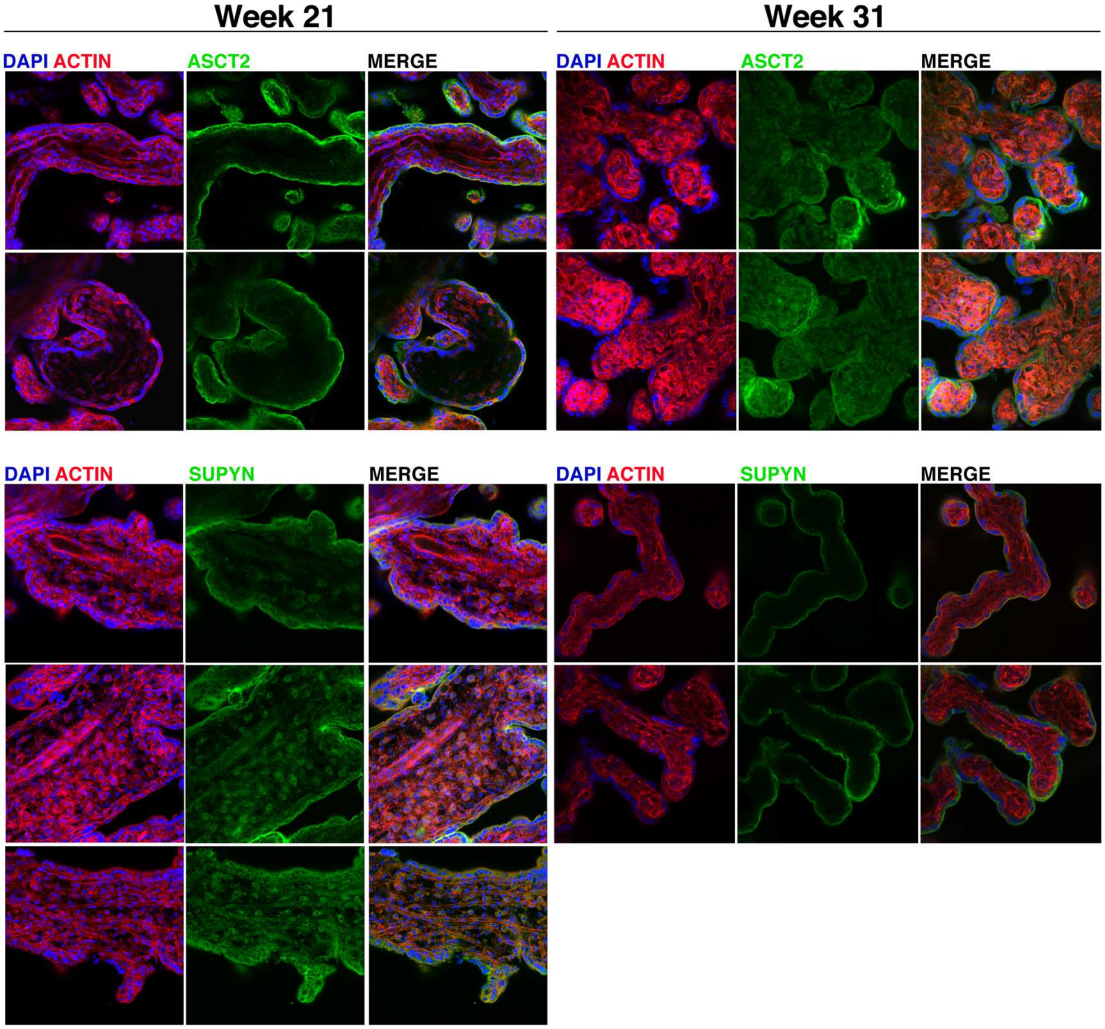

Figure 2.4: ASCT2 and SUPYN expression in 2nd and 3rd trimester human placenta.

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).

2.2.2 SUPYN CONFERS RESISTANCE TO RD114 ENVELOPE MEDIATED INFECTION

SUPYN expression during human embryonic and placental development, coincident with constitutively expressed ASCT2 **(Fig 2.1a),** suggests SUPYN may interact with ASCT2 throughout fetal development and confer resistance to RDR infection to the developing embryo. To begin testing this hypothesis, we first examined whether human placenta-derived cell lines Jar and JEG3 and the human ESC line H1 are resistant to RDRenv-mediated infection. We generated HIV-GFP viral particles pseudotyped with either the feline RD114env (HIV-RD114) or VSVg (HIV-VSVg), which allowed us to monitor the level of infection in cell culture based on GFP expression **(Fig 2.5)**[49]. These experiments revealed that Jar, JEG3, and H1 cells are susceptible to HIV-VSVg, as previously reported[50-54], but highly resistant to HIV-RD114 infection **(Fig 2.6a, b).** Concurrently infected 293T cells were similarly susceptible to infection by HIV-RD114 and HIV-VSVg **(Fig 2.6a, b)**.

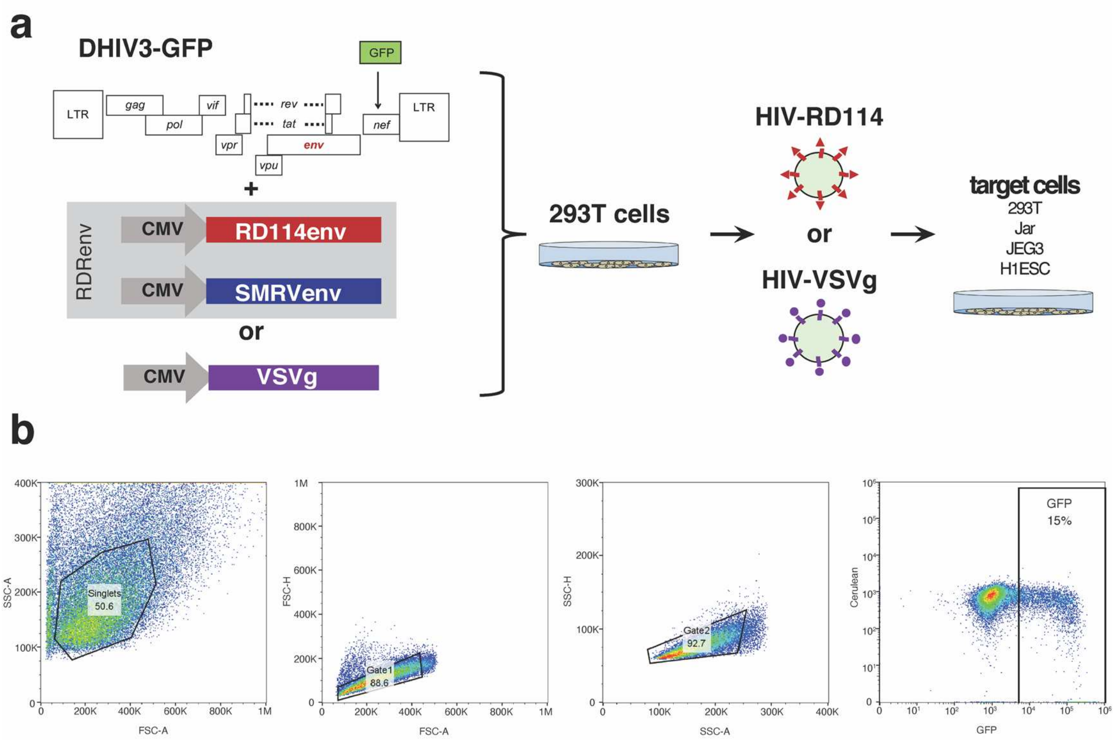

Figure 2.5: Reporter virus production and Flow Cytometry analysis scheme.
a, Env packaged HIV-GFP reporter virus particles were generated by co-transfecting 293T cells with DHIV3-GFP plasmid and a CMV promoter-driven glycoprotein encoding plasmid. Virus containing supernatant was then applied to target cells. RD114env and SMRVenv are representatives of the RDR interference group. **b,** Sequential gating scheme to assess reporter virus infection rate.

To test whether SUPYN contributes to the HIV-RD114 resistance phenotype, we repeated these infection experiments in Jar cells engineered to stably express short hairpin RNAs depleting ~80% of *SUPYN*[16] and *SYN1*[55] mRNAs respectively **(Fig 2.7a).** Depletion of SUPYN in Jar cells resulted in a significant increase in susceptibility to HIV-RD114 infection **(Fig 2.6c)**, but did not affect infection by HIV-VSVg **(Fig 2.6c**). Importantly, SYN1 depletion from Jar cells did not increase susceptibility to HIV-RD114 infection **(Fig 2.6c**).

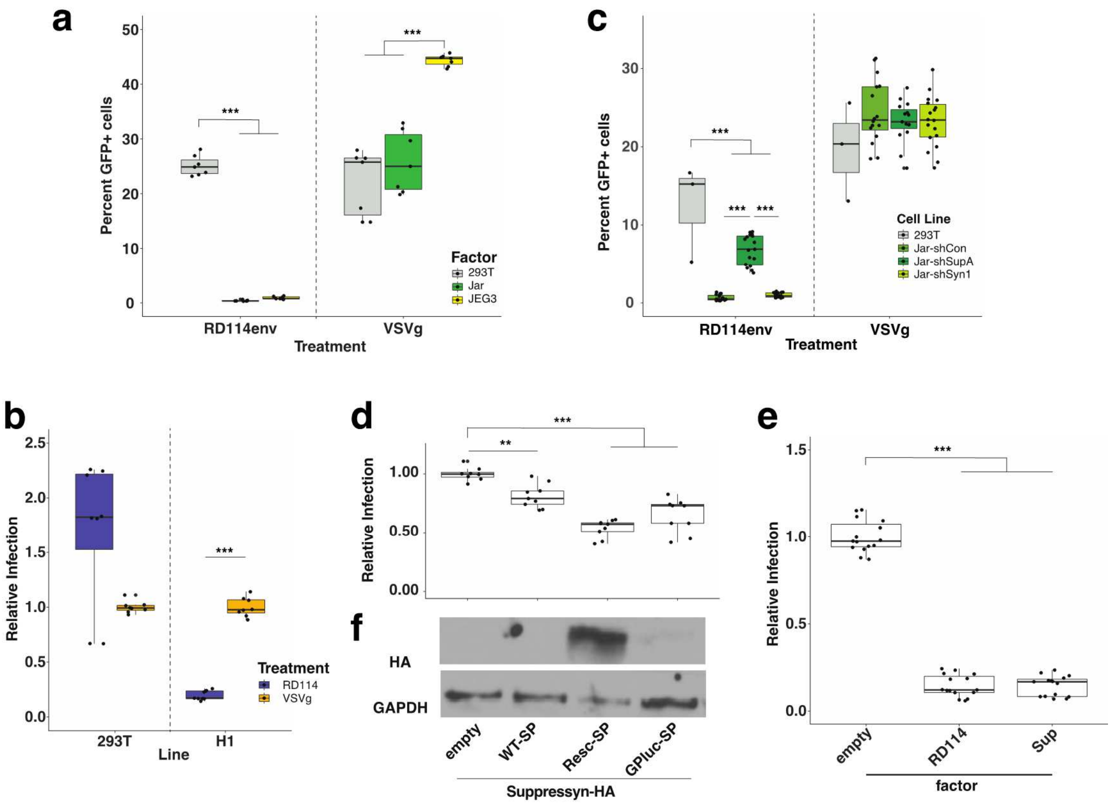

Figure 2.6: SUPYN confers resistance to RDR env mediated infection
(a, c) Proportion of infected (GFP+) 293T (grey), JEG3 (yellow), Jar (green), and shRNA-transduced Jar (green) cells infected with HIV-RD114 or HIV-VSVg. **(b)** Relative infection rate of 293T and H1-ESCs normalized to mean proportion of HIV-VSVg-infected cells **(d, e)** Relative infection rates of GFP+ 293T cells transfected with **(d)** wild-type (WT-SP), rescue (Resc-SP), *Gaussia princeps* luciferase signal peptide (GPluc), or **(e)** unmodified (Sup) SUPYN, and RD114env overexpression constructs. Relative infection was determined by normalizing indicated constructs to empty vector. **(f)** Western Blot analysis (αHA, αGAPDH) of 293T cell lysates transfected with indicated constructs. All assays were performed at least 3 times with a minimum of 2 technical replicates. ***adj. *p<0.001*; **adj. *p<0.01*; Tukey HSD.

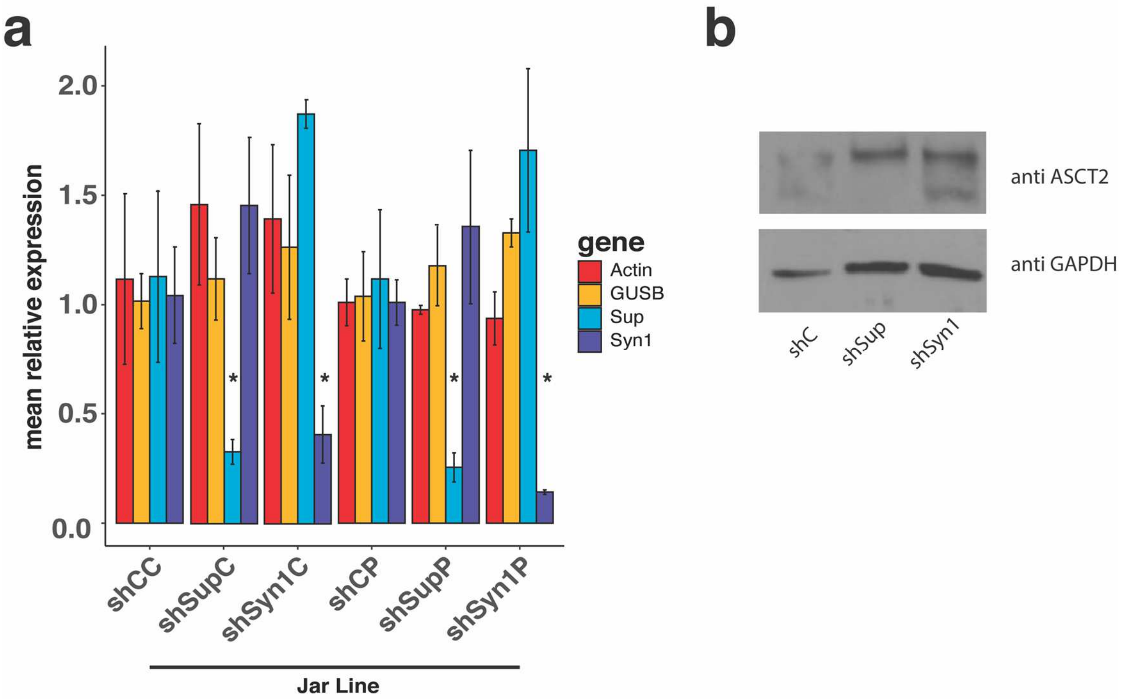

Figure 2.7: Characterization of shRNA transduced Jar cells and validation of env overexpression constructs.
(a, b) SUPYN and SYN1 knock down was validated by qPCR. Bar plots represent mean gene expression in Jar-shSupC, -shSyn1C, Jar-shSupP, and Jar-shSyn1P normalized to Jar-shCC and Jar-shCP respectively (n=3). Error bars represent ± standard error mean. *$p<0.1$; Wilcox rank sum test. **(b)** Western Blot analysis (αGAPDH, αASCT2) shRNA-transduced JEG3 cell lysates.

To account for possible off-target effects of SUPYN targeting siRNAs, we transfected Jar-shSup cells with siRNA-resistant, HA-tagged SUPYN rescue constructs (Sup-rescSP and Sup-lucSP) and infected with HIV-RD114. Both Sup-rescSP and Sup-lucSP significantly rescued resistance to HIV-RD114 infection **(Fig 2.6d).** Western Blot analysis of transfected cell lysates showed Sup-rescSP was more abundantly expressed than Sup-lucSP, which may account for the stronger resistance phenotype to HIV-RD114 infection **(Fig 2.6f).**

To test if SUPYN expression alone is sufficient to confer protection against HIV-RD114 infection, we transfected 293T cells, which are susceptible to RD114env-mediated infection, with SUPYN or RD114env overexpression constructs and subsequently infected with HIV-RD114 and HIV-VSVg respectively. Expression of RD114env and SUPYN resulted in ~80% reduction in the level of HIV-RD114 infection **(Fig 2.6e; Fig 2.8a)**, but had no significant effect on HIV-VSVg infectivity **(Fig 2.8b)**. Taken together, our KD and overexpression experiments indicate SUPYN expression is both necessary and sufficient to confer resistance to RD114env-mediated infection.

2.2.3 SUPPRESSYN RESTRICTS RDR INFECTION THROUGH RECEPTOR INTERFERENCE

Our RD114env-specific resistance phenotype **(Fig 2.8a, b)** strongly suggests SUPYN functions by receptor interference. If so, this protective effect should extend to infection mediated by other RDRenv[3,15,56] since they all use ASCT2 as their receptor. To test this prediction, we generated HIV-GFP reporter virions pseudotyped with Squirrel Monkey Retrovirus (SMRV) env (HIV-SMRVenv)[15] **(Fig 2.5)** and infected 293T cells previously transfected with SUPYN, SMRVenv or an empty vector. Cells expressing SUPYN or SMRVenv showed an ~80% reduction of HIV-SMRVenv infected cells **(Fig 2.8c)**. Thus, SUPYN expression is capable of restricting infection mediated by multiple RDRenv.

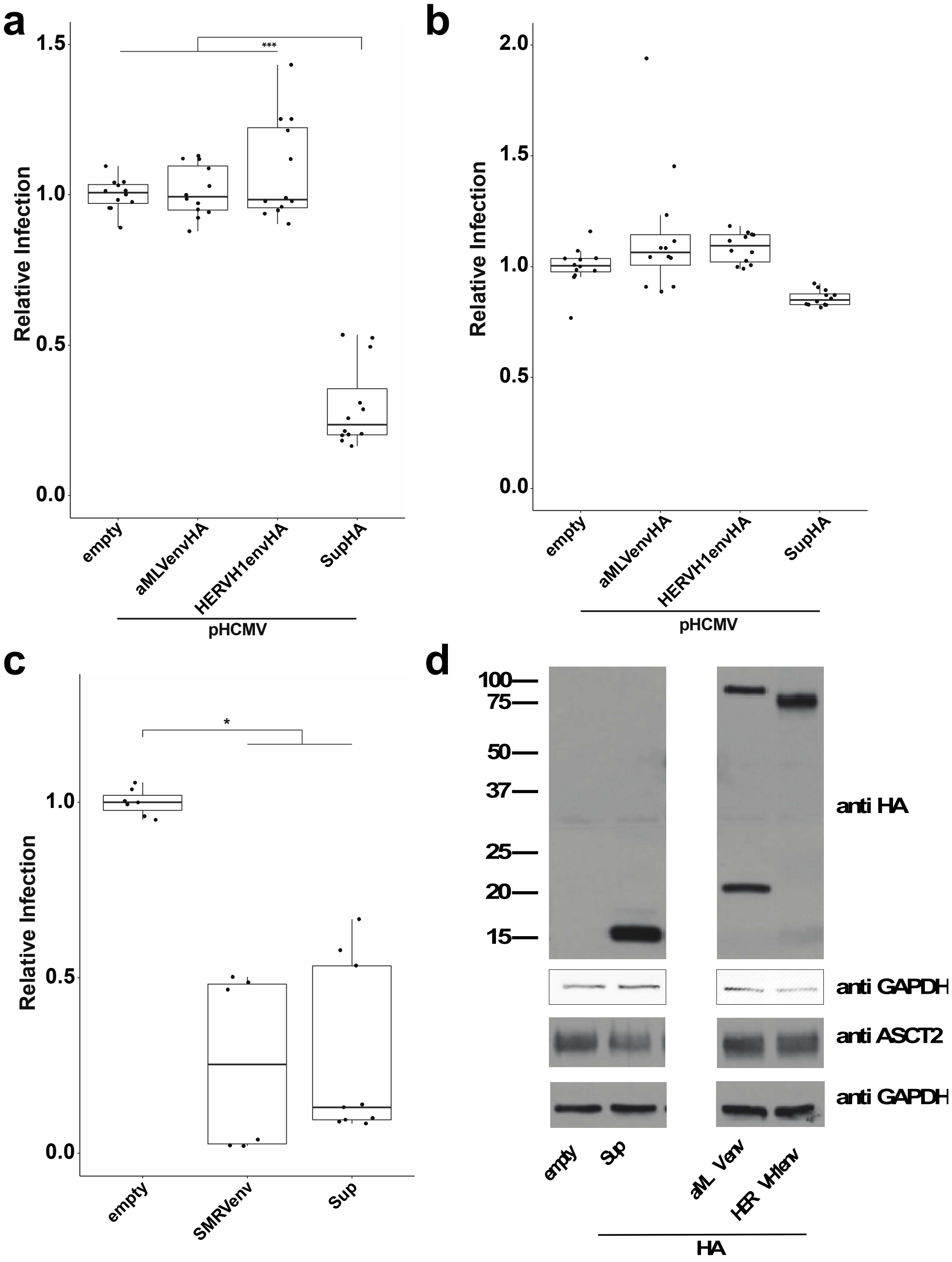

Figure 2.8: SUPYN expression is sufficient to specifically restrict RDRenv-mediated infection

(a, b, c) 293T cells, transfected with SMRVenv, SUP, SUP-HA and HA-tagged env, were infected with HIV-RD114 **(a)**, HIV-VSVg **(b)**, and HIV-SMRVenv **(c)** respectively. Relative infection rates were determined by normalizing GFP+ counts to empty vector. All assays were performed at least 3 times with a minimum of 1 technical replicates. **(d)** Western Blot analysis (αHA, αGAPDH, αASCT2) of 293T cell lysates following transfection with indicated constructs. ***adj. *p<0.001*; *adj. *p<0.05*; Tukey HSD,

Another prediction of RDR restriction via receptor interference is that it should be a property of envelope binding ASCT2, but not those using other cellular receptors. Consistent with this prediction, expressing HA-tagged envelopes from amphotrophic murine leukemia virus or human endogenous retrovirus H, neither of which are expected to interact with ASCT2[57-59], had no effect on HIV-RD114 nor HIV-VSVg infection in 293T cells. Conversely, HA-tagged SUPYN strongly restricted HIV-RD114 **(Fig 2.8a, b)**, yet all tested env were expressed at comparable levels **(Fig 2.8d).** Furthermore, we observed that SUPYN overexpression did not significantly impact ASCT2 expression levels in 293T cells **(Fig 2.8d).** This result suggests that if SUPYN acts by receptor interference, its interaction with ASCT2 does not result in ASCT2 degradation, which is consistent with some instances of receptor interference[60-62]. We also noted that SUPYN knock down in Jar cells seemed to result in the specific loss of a non-glycosylated ASCT2 isoform **(Fig 2.7b)**, which is consistent with previous observations[17]. While ASCT2 glycosylation may impact RDR infection susceptibility in mouse and hamster cells[63,64], it is unclear if glycosylation of human ASCT2 impacts RDR-env mediated infection. Nonetheless, all these observations converge on the model that SUPYN restricts against RDR infection through receptor interference.

2.2.4 SUPYN EMERGED IN A CATARRHINE ANCESTOR AND EVOLVED UNDER FUNCTIONAL CONSTRAINT

Little has been reported about the evolutionary origin of *SUPYN*. It was originally identified as derived from a member of the HERV-Fb endogenous retrovirus family (also known as HERVH48 in DFAM[65]) inserted on human chromosome 21q22.3 with an ortholog in chimpanzee[16]. Using comparative genomics (see Methods), we found that this HERVH48 insertion is present at an 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 2.9a; Fig 2.10).** These data indicate the endogenous retrovirus that gave rise to *SUPYN* inserted in the common ancestor of catarrhine primates ~20-38 million years ago[66] **(Fig 2.9a)**.

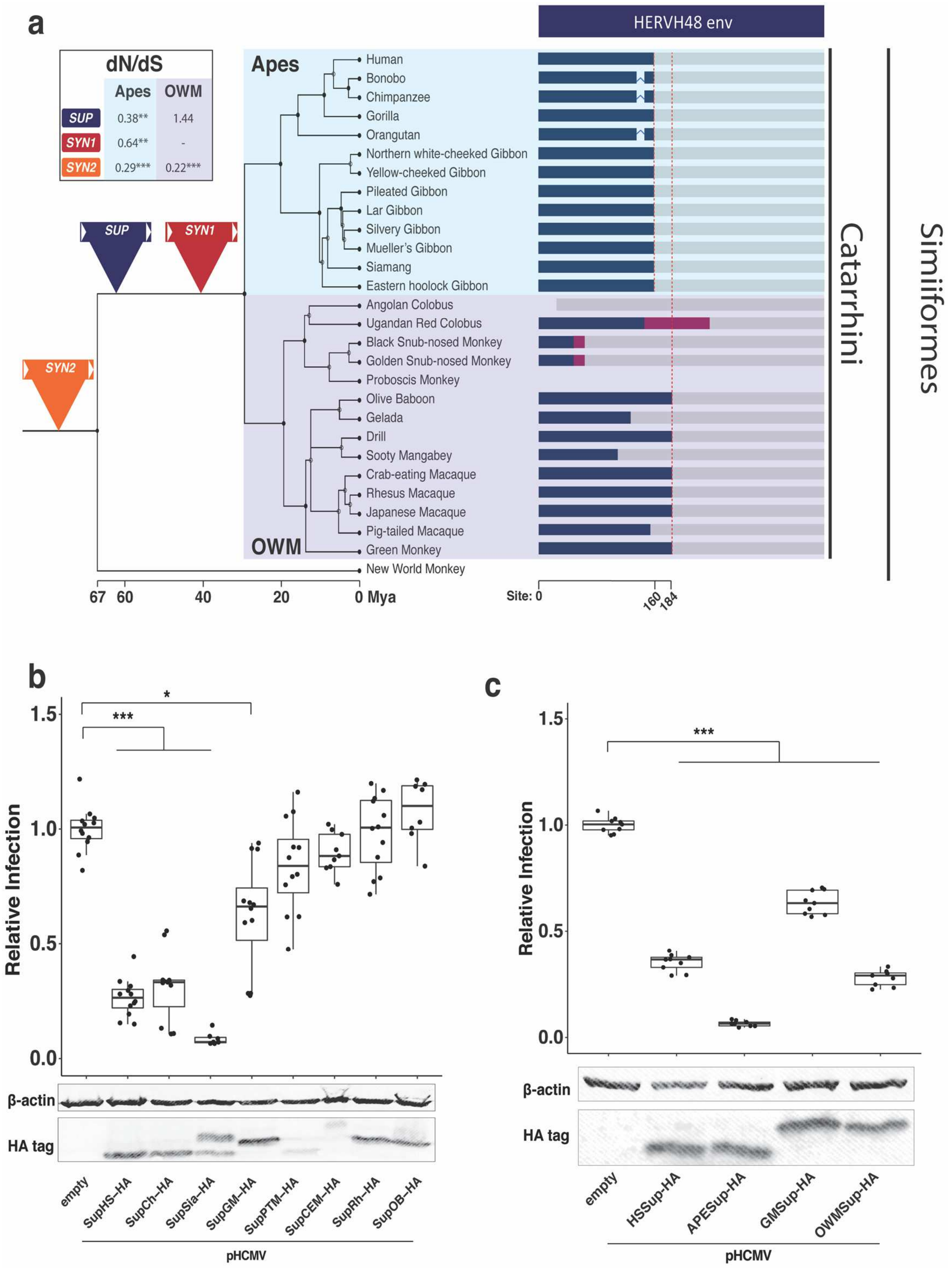

Figure 2.9: SUPYN is evolutionarily conserved in Catarrhinne primates and has antiviral activity in Hominoids.

(a) Consensus primate phylogeny with cartoon representation of intact *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 downstream HERVH48env sequence. *SUPYN*-, *SYN1*-, and *SYN2*-labeled triangles denote ancestral lineage where ERVenv acquired. Lineage specific *SUPYN*, *SYN1* and *SYN2* dN/dS values are shown in box. **(b, c)** 293T cells transfected with primate (**b**) or ancestral (**c**) SupHA constructs were infected with HIV-RD114. Relative infection rates were determined by normalizing GFP+ counts to empty vector. All assays were performed at least 3 times with a minimum of 2 technical replicates ***adj. *p<0.001*; *adj. *p<0.05*; Tukey HSD.

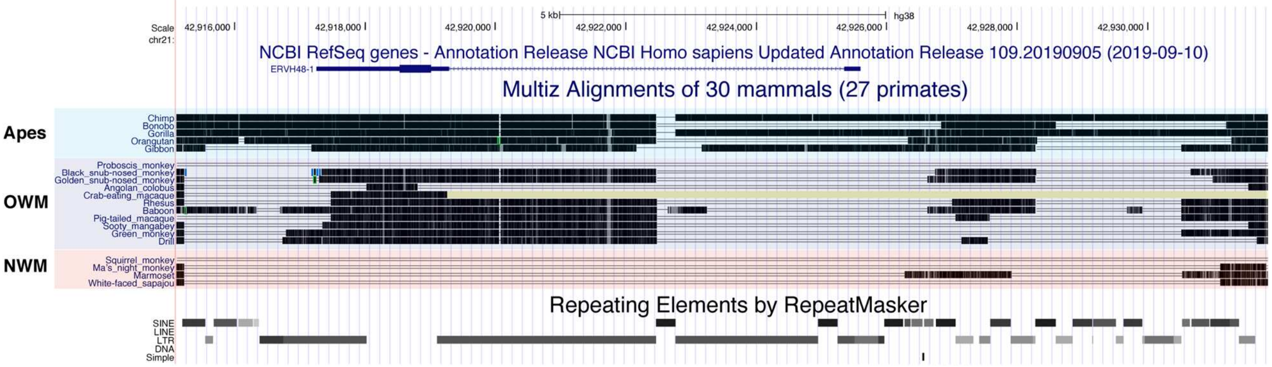

Figure 2.10: SUPYN locus conservation in primates.
UCSC genome Browser snapshot of SUPYN-coding with surrounding sequence. NCBI RefSeq gene, Simiforme primates of the 30-species primate whole genome alignment, and RepeatMasker repetitive element tracks are shown.

All primates with HERVH48 orthologs also share a nonsense mutation which would have truncated the ancestral encoded env protein at site 185 in the common ancestor of catarrhine primates. Hominoids share an additional nonsense mutation further truncating the protein to the 160-aa SUPYN-encoding ORF currently annotated in the human reference genome **(Fig 2.9a; Fig 2.11).** The *SUPYN* ORF is almost perfectly conserved in length across hominoids, but not in OWM where some species display further truncating and frameshifting mutations, suggesting *SUPYN* may have evolved under different evolutionary regimes in hominoids and OWMs. To test this idea, we analyzed the ratio (ω) of nonsynonymous (dN) to synonymous (dS) substitution rates using codeml[67], 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 ($\omega = 0.38$; $p = 1.47\text{E-}02$), but did not depart from neutral evolution in OWMs ($\omega = 1.44; p = 0.29$) **(Fig 2.9a).** For comparison, we performed the same type of analysis for *SYN1* and *SYN2*, primate-specific *env*-derived genes presumably involved in placentation[47,68,69]. Consistent with previous reports[70,71], we found that both *SYN1* ($\omega = 0.64; p = 0.0180$) and *SYN2* ($\omega = 0.29$; $p = 3.22$E-08) evolved under purifying selection during hominoid evolution **(Fig 2.9a)**. In OWMs, *SYN2* also evolved under purifying selection ($\omega = 0.22$, p = 2.78E-08), while *SYN1* was lost through an ancestral deletion[26] **(Fig 2.9a).** These results suggest that the level of functional constraint acting on *SUPYN* during hominoid evolution is comparable to that seen on other *env*-derived genes with placental function.

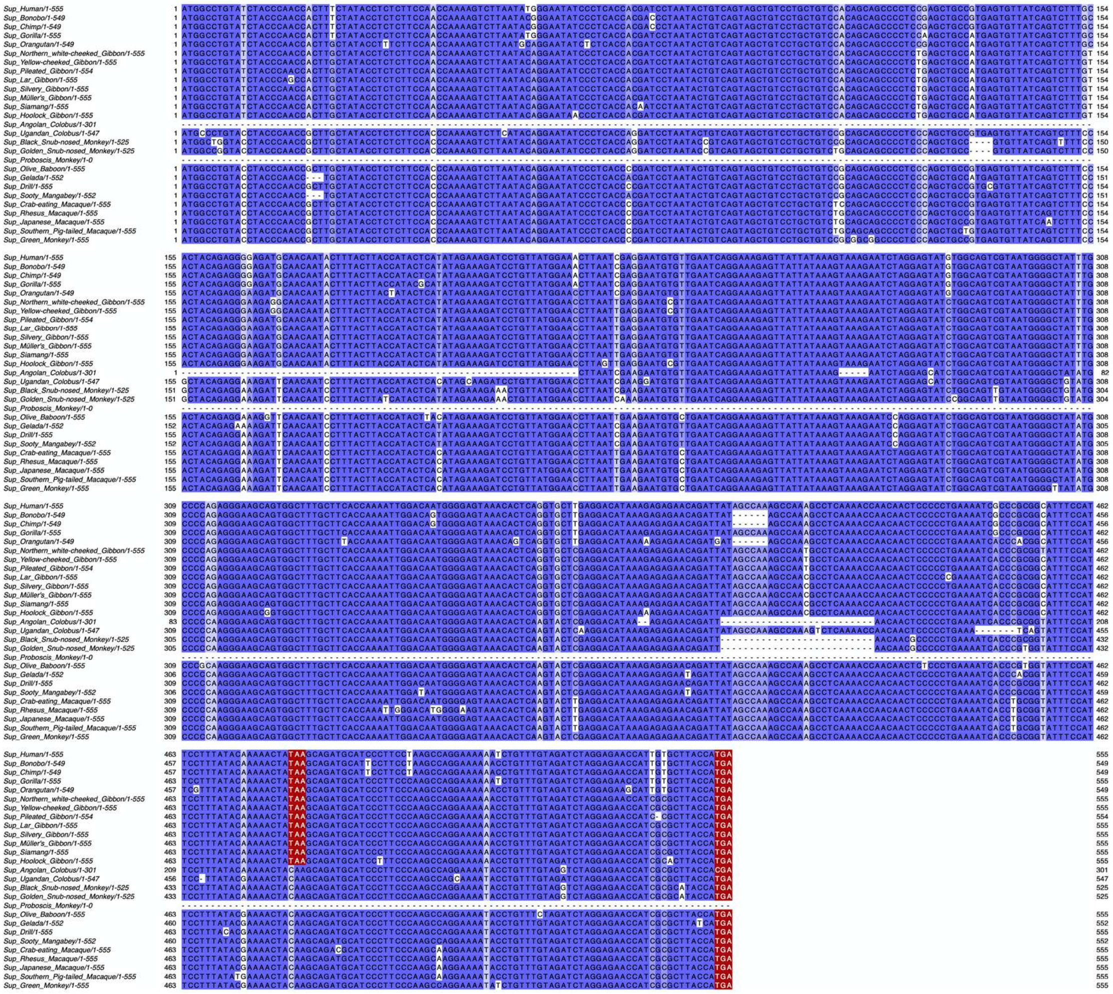
Sup_Human/1-555
Sup_Bonobo/1-549
Sup_Chimp/1-549
Sup_Gorilla/1-555
Sup_Orangutan/1-549
Sup_Northern_white-cheeked_Gibbon/1-555
Sup_Yellow-cheeked_Gibbon/1-555
Sup_Pileated_Gibbon/1-554
Sup_Lar_Gibbon/1-555
Sup_Silvery_Gibbon/1-555
Sup_Müller's_Gibbon/1-555
Sup_Siamang/1-555
Sup_Hoolock_Gibbon/1-555
Sup_Angolan_Colobus/1-301
Sup_Ugandan_Colobus/1-547
Sup_Black_Snub-nosed_Monkey/1-525
Sup_Golden_Snub-nosed_Monkey/1-525
Sup_Proboscis_Monkey/1-0
Sup_Olive_Baboon/1-555
Sup_Gelada/1-552
Sup_Drill/1-555
Sup_Sooty_Mangabey/1-552
Sup_Crab-eating_Macaque/1-555
Sup_Rhesus_Macaque/1-555
Sup_Japanese_Macaque/1-555
Sup_Southern_Pig-tailed_Macaque/1-555
Sup_Green_Monkey/1-555

Figure 2.11: 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.

2.2.5 SUPYN ANTIVIRAL ACTIVITY IS CONSERVED ACROSS HOMINOID PRIMATES

To assess whether primate *SUPYN* orthologs have antiviral activity, we generated and transfected 293T cells with HA-tagged overexpression constructs for the orthologous SUPYN sequences of chimp, siamang, African green monkey, pigtailed macaque, crab-eating macaque, Rhesus macaque, and olive baboon and challenged these cells with HIV-RD114 virions. Both chimp and siamang SUPYN proteins displayed antiviral activity with potency comparable to and greater than human SUPYN, respectively **(Fig 2.9b).** By contrast, only one (African green monkey) of the five OWM orthologous SUPYN proteins exhibited a modest but significant level of antiviral activity **(Fig 2.9b, 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 and/or their inability to bind the human ASCT2 receptor due to *SUPYN* and ASCT2 sequence divergence **(Fig 2.11; Fig 2.12)**. To gain further insight into the evolutionary origins of SUPYN antiviral activity, we reconstructed SUPYN sequences predicted for the common ancestor of hominoid and OWM (see Methods) and assayed their antiviral activity by expressing them in 293T cells. Both ancestral proteins were expressed at levels comparable to human SUPYN and exhibited significant antiviral activity **(Fig 2.9c).** These data indicate that SUPYN antiviral activity against RDRenv-mediated infection is an ancestral trait, which has been preserved over ~20 million years of hominoid evolution but may have been lost in some OWM lineages.

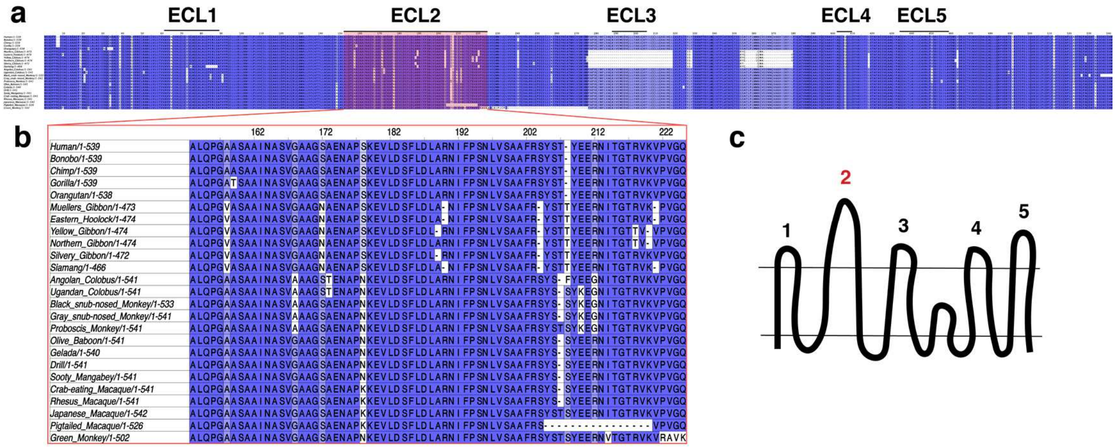

Figure 2.12: Conservation of ASCT2 RDR env binding region across Catarrhine primates.
(a) Amino acid sequence alignment of ASCT2 from Catarrhine primates. Extracellular loops (ECL), described by Marin et al. 2003, are indicated by black lines. ECL2, containing the RDRenv-binding region, is highlighted in red and amino acid sequence is shown in **(b)**. Amino acid sequences are shaded blue based on minimum sequence identity thresholds of 45% (light), 75% (medium) and 80% (dark) respectively. **(c)** ASCT2 protein topology is represented as described by Marin et al. 2003. Numbering corresponds to ECLs.

2.3 DISCUSSION

Our expression and selection analyses **(Fig 2.1; Fig 2.9)** firmly establish that *SUPYN* is a bona fide gene encoding a truncated envelope of retroviral origin that is highly expressed in the human preimplantation embryo and throughout placental development. Virological assays in human cell culture **(Fig. 2.6; Fig 2.8)** indicate SUPYN is necessary and sufficient to confer resistance to RDRenv-mediated infection, likely by interfering with the receptor (ASCT2) utilized by this diverse group of retroviruses. The expression profile of SUPYN **(Fig 2.1)** and the RD114 resistance phenotype of human ESCs and placental cells **(Fig. 2.6)** suggest *SUPYN* may provide protection against zoonotic retroviral infection of the developing embryo and perhaps retroviral invasion of the developing germline. The observation that extant, infectious RDRs are absent in hominoids[15] lends further support to a model in which SUPYN may have helped confer resistance to RDRs in Hominoids.

Like *SYN1*, *SUPYN* emerged in the common ancestor of catarrhine primates and was preserved by natural selection in hominoids. This parallel evolutionary path and the pattern of expression of *SUPYN* and *SYN1* in the placenta remain compatible with a model in which *SUPYN* acts as a negative modulator of *SYN1* fusogenic activity[16,17]. The developmental and antiviral functions of SUPYN are not mutually exclusive and may even be interlocked. Indeed, Syncytins, including SYN1, are fully functional envelopes that can be incorporated into heterologous retroviral particles and exosomes originating from the placenta[50-55,72]. 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 (and possibly the mother) to a wide variety of invasive genetic elements. Both exogenous and endogenous retroviral particles could be serendipitously enveloped by SYN1 throughout pregnancy. As such, it is tempting to speculate that SUPYN has been maintained by natural selection to shield the developing embryo from the 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.

This study also serves as a proof of principle that truncated envelope peptides expressed from relics of retroviruses fossilized in the human genome can exert and retain antiviral activities for millions of years. In fact, a preliminary search (see methods) for human endogenous retrovirus-derived *env* identified 30 conserved candidate open reading frames, seven of which had a significant signature of purifying selection **(Table 2.2)**.

Furthermore, Gag (capsid)-derived proteins encoded by endogenous retroviruses are also capable of retroviral restriction[33,73,74]. Thus, it is possible that our genomes encode a vast reservoir of retroviral-derived proteins with the ability to restrict various zoonotic agents, including non-retroviral pathogens (e.g. coronaviruses, intracellular bacteria) that use cell surface receptors to infect human cells.

Table 2.2: Summary of identified ERV env open reading frame candidates

ORF ID	genome loccation (hg38)	ERVenv chr	ERVenv start	ERVenv stop	HERV ID	Gene Overlap	Conservation	dN/dS	dN/dS p-value
hg19_chr2_71620367-71623016A	chr2:71393363-71393819	chr2	71393363	71393819	HERVK22	ZNF638	Catarrhini	0.71	1.98E-01
hg19_chr2_119640978-119643357	chr2:118884631-118885087	chr2	118884631	118885087	HERV9NC	-	Catarrhini	1.06	8.06E-01
hg19_chr3_44374968-44388273B	chr3:44334552-44334804	chr3	44334552	44334804	MER84	-	Simiformes	0.71	2.77E-01
hg19_chr3_121322365-121325892B	chr3:121605598-121605871	chr3	121605598	121605871	PABL-B	FBXO40 int	Simiformes	0.96	8.93E-01
hg19_chr4_53609324-53611916	chr4:52743828-52745574	chr4	52743828	52745574	MER34	ERVMER34-1	Simiformes	0.67	6.48E-06
hg19_chr4_56804097-56806740	chr4:55939277-55939544	chr4	55939277	55939544	N/A	-	Simiformes	0.56	4.60E-02
hg19_chr4_154609386-154612303A	chr4:153689502-153689766	chr4	153689502	153689766	HERVK9	-	Catarrhini	0.3	1.80E-02
hg19_chr5_43569150-43571691	chr5:43569855-43570197	chr5	43569855	43570197	PRIMA41	-	Catarrhini	1.06	8.62E-01
hg19_chr5_56815370-56818833C	chr5:57520623-57521472	chr5	57520623	57521472	HERV17	CTD-2023N9	Catarrhini	1.19	3.43E-01
hg19_chr5_58609110-58611420A	chr5:59314073-59314292	chr5	59314073	59314292	N/A	-	Simiformes	0.59	1.29E-01
hg19_chr5_150366488-150368669	chr5:150987303-150987795	chr5	150987303	150987795	LTR46	-	Simiformes	1.16	4.88E-01
hg19_chr6_11102913-11106510	chr6:11103693-11105310	chr6	11103693	11105310	MER50	Syn2	Simiformes	0.36	3.66E-15

hg19_chr6_28043414-28046665	chr6:28077393-28077606	chr6	28077393	28077606	N/A	-	Primate	1.2	4.00E-01
hg19_chr7_64450342-64454344	chr7:64991211-64993032	chr7	64991211	64993032	HERV3	ERV3-1	Catarrhini	0.41	8.87E-07
hg19_chr7_99494522-99497192	chr7:99897888-99898185	chr7	99897888	99898185	N/A	TIRM4 int	Simiformes	0.93	8.10E-01
hg19_chr8_41448290-41452025	chr8:41592988-41593375	chr8	41592988	41593375	HERVe_a	GPAT-4 int	Catarrhini	1.2	5.83E-01
hg19_chr9_90651934-90655861A	chr9:88039262-88039946	chr9	88039262	88039946	HERVIP10B3	-	Catarrhini	1.06	8.41E-01
hg19_chr9_90651934-90655861B	chr9:88038892-88039357	chr9	88038892	88039357	HERVIP10B3	-	Catarrhini	0.64	1.59E-01
hg19_chr9_125251197-125253596	chr9:122489901-122490144	chr9	122489901	122490144	HERVL66	-	Catarrhini	0.78	6.11E-01
hg19_chr11_62136133-62144843C	chr11:62375427-62375706	chr11	62375427	62375706	HERVK	ASRGL int	Hominoid	0.46	7.06E-02
hg19_chr12_68936283-68942263B	chr12:68545400-68545739	chr12	68545400	68545739	Harlequin	-	Catarrhini	0.77	4.00E-01
hg19_chr14_32709691-32713257	chr14:32242538-32242850	chr14	32242538	32242850	N/A	-	Simiformes	0.75	3.18E-01
hg19_chr14_93088234-93092227	chr14:92622884-92624900	chr14	92622884	92624900	HERVIP10B3	RIN3 int	Simiformes	0.96	6.89E-01
hg19_chr19_53516345-53519738	chr19:53014090-53015524	chr19	53014090	53015524	MER66	ERVV1	Simiformes	0.46	7.38E-13
hg19_chrX_4688854-4691905	chrX:4772543-4772834	chrX	4772543	4772834	HERVL66	FTX lncRNA	Catarrhini	0.36	3.99E-03
hg19_chrX_62499347-62502899	chrX:63281470-63281866	chrX	63281470	63281866	N/A	-	Simiformes	0.67	5.78E-02

hg19_chrX_73356170-73358525	chrX:74137210-74137582	chrX	74137210	74137582	HUERS-P3b	-	Catarrhini	0.9	8.10E-01
hg19_chrX_99823301-99826264	chrX:100569921-100570197	chrX	100569921	100570197	HERVP71A	-	Catarrhini	1.7	4.71E-01
hg19_chrX_100051694-100054754A	chrX:100797753-100798005	chrX	100797753	100798005	PRIMA41	-	Hominoid	0.13	5.11E-02
hg19_chrX_100051694-100054754B	chrX:100798005-100798332	chrX	100798005	100798332	PRIMA41	-	Hominoid	0.87	8.00E-01

2.4 MATERIALS AND METHODS

2.4.1 Single cell RNAseq data analysis

We mined published single cell transcriptome datasets of human pre-implantation embryos isolated at developmental stages ranging from oocyte to blastocyst (PMID: 23934149) and human placenta (PMID: 30042384, GSE89497), which were generated on various Illumina platforms. Reads were mapped to the human genome (hg19) with STAR[75] using the following settings *--alignIntronMin 20 --alignIntronMax 1000000 --chimSegmentMin 15 --chimJunctionOverhangMin 15 --outFilterMultimapNmax 20*. Only uniquely mapped reads were considered for expression calculations. Gene level counts were obtained using *featureCounts*[76] run with RefSeq annotations. Gene expression levels were calculated at Transcript Per Million (TPM) from counts mapped over the entire gene (defined as any transcript located between the Transcription Start Site (TSS) and Transcription End Site (TES)). Only cells that met the following criteria were included in this analysis: (1) Cells must express at least 5000 genes. (2) Genes must be expressed in at least 1% of cells. (3) Genes must meet a log2 TPM > 1 threshold. We clustered cells meeting these criteria using the default parameters of the Seurat (v2.3[77,78]) package implemented in R. Seurat applies the most variable genes to get top principle components that are used to discriminate cell clusters in tSNE or UMAP plots. In our analyses, 10 principle components were chosen to define cell cluster. Major clusters corresponding to CTB, STB, EVTB, Macrophages, and stromal cells were identified based on the expression of known marker genes. Monocle2[79] 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.

2.4.2 Analysis of 10X Genomics datasets

Data generated on the 10X Genomics scRNAseq platforms were processed in the following way. The processed data matrix from (PMID:30429548) was first fetched from the E-MTAB-6701 entry. Normalized counts and cell-type annotations were used as provided by the original publications. Seurat (v3.1.1), implemented in R (v3.6.0), 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 ≥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. (4) 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 UMAP plots. Cell types were assigned based on the annotations provided by the original publication.

2.4.3 ChIP-seq data analysis

Various ChIP-seq datasets representing Histone modifications and Transcription factors in Human embryonic stem cells and their differentiation were fetched from (PMID: 25693565, GSE61475). We obtained the H3K27Ac (PMID:31294776, GSE127288) for CTB to STB primary cultures, H3K4Me1 for trophoblasts (PMID:30231016, GSE118289), H3K4Me3, H3K27Me3 for differentiated trophoblasts (PMID: 29078328, GSE105258), and GATA2/3, TFAP2A/C (PMID: 29078328, GSE105081)

ChIP-seq datasets in raw fastq format. ChIP-seq reads were aligned to the hg19 human reference genome using the Bowtie2[80] using *--very-sensitive-local* mode. All reads with MAPQ < 10 and PCR duplicates were removed using Picard and *samtools*[81]. All the ChIP-seq peaks were called by MACS2 **[Gaspar. BioRxiv. 2018]** with the parameters in narrow mode for TFs and broad mode for histone modifications keeping FDR < 1%. ENCODE-defined blacklisted regions[82] were excluded from called peaks. We then intersected these peak sets with repeat elements from hg19 repeat-masked coordinates using bedtools *intersectBed*[83] with a 50% overlap. To visualize over Refseq genes (hg19) using IGV[84], raw ChIP-seq signals were obtained with MACS2, using the parameters: *-g hs -q 0.01 -B*. The conservation track was visualized through the UCSC genome browser[26] under net/chain alignment of given non-human primates (NHPs) and merged beneath the IGV tracks.

2.4.4 Cell culture

293T cells (provided by Nels Elde) were cultured in DMEM containing 10% Fetal Bovine serum (FBS) (GIBCO). Jar cells (provided by Carolyn Coyne) were cultured in RPMI containing 10% FBS. JEG3 cells were cultured in MEM (GIBCO) containing 10% FBS. Culture medium for these cell lines was supplemented with sodium pyruvate (GIBCO), glutamine (GIBCO), and Penicillin Streptomycin (GIBCO) according to manufacturer specifications. H1-ESCs (obtained from WiCell) were grown on Matrigel (Corning, 356277) coated plates in MTESR+ (Stemcell) growth-media and subcultured using Accutase (Innovative Cell Techonologies, AT-104) and MTESR+ supplemented with CloneR (Stemcell). All cell lines were cultured at 37C and 5% CO_2.

2.4.5 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). SUPYN and HERVH1env ORFs were PCR amplified using Q5 polymerase (NEB) from HeLa and 293T genomic DNA respectively and cloned into a TOPO vector (ThermoFisher).

To generate siRNA-resistant SUPYN rescue constructs, we replaced the native signal peptide sequence (which is targeted by siRNAs used in this study) with (1) a *Gaussia princeps* luciferase SP (Sup-lucSP) [85,86] and (2) a codon optimized shSup resistant SUPYN rescue construct (Sup-rescSP).

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), and cloned into EcoRI digested pHCMV backbones using the InFusion cloning kit (Takara Bio, 638920).

pHIV7 lentiviral constructs were cloned using the pHIV7-U6-shW3 plasmid[55] (provided by Lars Aagaard) 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 and KpnI. 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 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.

2.4.6 Antibodies

All antibodies used in this study are commercially available. α-GAPDH, α-βactin, α-HA, α-ASCT2 primary antibodies were purchased from Cell Signaling Technology. α-Mouse and α-Rabbit HRP conjugated secondary antibodies were purchased from Cell Signaling Technology. IRDye secondary antibodies were purchased from Licor. α-SUPYN primary antibody was purchased from Phoenix Pharma. Alexa-fluor conjugated secondary antibody was purchased from Invitrogen.

2.4.7 Western Blot

Whole cell extracts from cultured cell lines were prepared using 1x GLO lysis buffer (Promega). One third volume of 4x Laemli buffer was added to one volume whole cell extract samples, then incubated at 95C for 5 minutes, and sonicated for 15 minutes at 4C (amplitude 100; pulse interval 15 sec on 15 sec off). Approximately 30ug of protein were separated by SDS-PAGE (BioRad 12% gel), transferred to PVDF membrane (BioRad), blocked according to antibody manufacturers specification, and incubated overnight in appropriate primary antibody then incubated in IRDye (Licor) or peroxidase conjugated goat anti-mouse or anti-rabbit antibodies (Cell Signaling technology) for 1hour at room temperature. Protein was then detected using ECL reagent (BioRad) or the Licor Odyssey imaging system.

2.4.8 IF microscopy

Placental tissues were fixed in 4% PFA (in 1x PBS) for 30min, permeabilized with 0.25% Triton X-100 for 30min (on a rocker), washed with 1x PBS and then incubated with primary anti-Suppressyn antibody at 1:200 in 1xPBS for 2-4h at RT. These samples were incubated with Alexa-fluor conjugated secondary antibody (Invitrogen) diluted 1:1000 and counterstained with actin (or CD163). DAPI was included in our PBS and then mounted in Vectashield mounting medium with DAPI (H-1200).

2.4.9 Virus production

Low passage 293T cells were used to produce all lentiviral particles. DHIV3-GFP and env-expression plasmids were co-transfected at a mass ratio of 2:1 using lipofectamine 2000 (ThermoFisher). shRNA encoding lentiviral particles were produced by cotransfecting pHIV7, psPAX2, pVSVg according to BROAD institute lentiviral production protocol 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 0.45um pore filter, and stored at -80C.

2.4.10 Infection Assays

HEK 293T cells were transfected with env-overexpression constructs using Lipofectamine 2000 (*Invitrogen*) and incubated 24hrs._Transfected cells were infected with reporter virus by applying virus (HIV-RD114, HIV-VSVg, HIV-SMRVenv) stocks in the presence of polybrene (*Santa Cruz Bio*) at a final concentration of 4ug/mL. After 6-8hrs, virus stock was replaced with fresh growth media. Infected cells were maintained for 72hrs, replacing media when necessary, and harvested with trypsin (293T). Detached cells were suspended in fresh growth media, strained and analyzed by flow cytometry.

2.4.11 Placental cell shRNA transduction

Placenta-derived cell lines were treated with pHIV-shRNA-virus-containing supernatant and incubated for 72hrs 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) at a final concentration of 3.5 ug/mL for 7 days, then cultured in regular growth media.

2. 4.12 RT-qPCR

RNA was isolated from cultured cells using the RNeasy Mini Kit (Qiagen) and an on column dsDNAse digestion was performed. 1-3 ug of total RNA were used to generate cDNA with the maxima cDNA synthesis with dsDNAse kit (ThermoFisher). qPCR reactions were performed using the LC480 with Sybr Green PCR master mix (Roche) according to manufacturer's protocol. Gene expression was then quantified using the ΔΔCT method. 18S expression was used as the reference housekeeping gene.

2.4.13 Envelope evolutionary sequence analyses

Orthologous SUPYN, SYN1, and SYN2 sequences were extracted from the 30-species MULTIZ alignment[26] and formatted for sequence alignment using the phast package[87]. These and additional syntenic SUPYN and SYN2 open reading frame sequences were validated/identified by BLASTn[88] search with default settings of publicly available Catarrhine primate genomes (ncbi.nih.gov). The Carbone Lab (OHSU) 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[81] and JalView[89].

Orthologous env sequences (>90bp length) encoding the mature sequence downstream of the signal peptide cleavage site, were aligned using MEGA7[90] and manually converted to PHYLIP format. A newick tree was generated based on this alignment using the maximum likelihood algorithm implemented in MEGA7. The *codeml* program implemented in the PAML package was then run to calculate *dN/dS* values and log likelihood (LnL) scores generated under models M0, M1, M2, M7 and M8[67]. Chi-square tests comparing LnL scores generated under models of neutral evolution and selection were performed.

Ancestral ape and OWM SUPYN sequences were reconstructed using the *baseml* program implemented in the PAML package. For the ancestral ape sequence, a newick tree was generated for the 13 ape species shown in Figure 4a, using a maximum likelihood algorithm implemented in MEGAX[91-93]. The *baseml* program was run using nucleotide substitution models 3-7 (F84, HKY85, T92, TN93, REV) and a reconstruction was generated for each node on the tree. For the ancestral OWM SUPYN sequence, a newick tree was generated the same way, this time using the 6 old world monkey sequences with the most complete open reading frame (184 amino acids) and the 13 Ape SUPYN open reading frame sequences (up to their stop codon). The *baseml* program was run using models 3, 4, 5, 6, and 7, and a reconstruction was generated for each node on the tree, including the one encompassing the OWM and Ape clades respectively.

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

Candidate envelope open reading frames were identified by performing tBLASTn[88] searches of the hg19 human genome assembly using envelope amino acid sequences, taken from the Repbase collection and published retroviral envelope sequences, as a query. Collected hits were used as a query to repeat a tBLASTn search, initially yielding 82715 candidate open reading frames. This list of candidates was filtered using the following criteria. (1) Only open reading frames with a length ≥100aa. (2) Hits starting a position ≥300aa were removed because such open reading frames are predicted to encode a portion of the envelope transmembrane domain, which does not play a role in receptor binding. (3) After these processing steps, our list was further concatenated to only include unique genome coordinates (n=2183). The position of these candidate sequences was then intersected[83] with conserved elements genome positions, which are reported in the 20-species primate alignment track[26], to identify candidate env open reading frames with evidence of sequence conservation in primates (see Supplemental Table 2). Candidate sequences were processed as described in 2.4.13 to identify sequences with a signature of purifying selection.

2.4.15 Statistical Analyses

Wilcox rank sum and Tukey Honest Statistical Difference tests were implemented in R. Boxplots and barplots were generated using ggplot2[94] implemented in R.

REFERENCES

1. Warren, C. J. & Sawyer, S. L. How host genetics dictates successful viral zoonosis. *PLoS Biol.* **17,** e3000217 (2019).
2. Feschotte, C. & Gilbert, C. Endogenous viruses: insights into viral evolution and impact on host biology. *Nat. Rev. Genet.* **13,** 283–296 (2012).
3. Johnson, W. E. Endogenous Retroviruses in the Genomics Era. *Annu Rev Virol* **2,** 135–159 (2015).
4. Metegnier, G. *et al.* Comparative paleovirological analysis of crustaceans identifies multiple widespread viral groups. *Mob DNA* **6,** 16 (2015).
5. Aiewsakun, P. & Katzourakis, A. Endogenous viruses: Connecting recent and ancient viral evolution. *Virology* **479-480,** 26–37 (2015).
6. Gilbert, C. & Feschotte, C. Genomic fossils calibrate the long-term evolution of hepadnaviruses. *PLoS Biol.* **8,** e1000495 (2010).
7. Niewiadomska, A. M. & Gifford, R. J. The extraordinary evolutionary history of the reticuloendotheliosis viruses. *PLoS Biol.* **11,** e1001642 (2013).
8. Aswad, A. & Katzourakis, A. Convergent capture of retroviral superantigens by mammalian herpesviruses. *Nat Commun* **6,** 8299 (2015).
9. Blanc, G., Gallot-Lavallée, L. & Maumus, F. Provirophages in the Bigelowiella genome bear testimony to past encounters with giant viruses. *Proc. Natl. Acad. Sci. U.S.A.* **112,** E5318–26 (2015).
10. Hayward, A., Cornwallis, C. K. & Jern, P. Pan-vertebrate comparative genomics unmasks retrovirus macroevolution. *Proc. Natl. Acad. Sci. U.S.A.* **112,** 464–469 (2015).
11. Han, G.-Z. & Worobey, M. A primitive endogenous lentivirus in a colugo: insights into the early evolution of lentiviruses. *Molecular Biology and Evolution* **32,** 211–215 (2015).
12. Diehl, W. E., Patel, N., Halm, K. & Johnson, W. E. Tracking interspecies transmission and long-term evolution of an ancient retrovirus using the genomes of modern mammals. *Elife* **5,** e12704 (2016).
13. Aiewsakun, P. & Katzourakis, A. Marine origin of retroviruses in the early Palaeozoic Era. *Nat Commun* **8,** 13954 (2017).
14. Montiel, N. A. An updated review of simian betaretrovirus (SRV) in macaque hosts. *J. Med. Primatol.* **39,** 303–314 (2010).
15. Sinha, A. & Johnson, W. E. ScienceDirect Retroviruses of the RDR superinfection interference group: ancient origins and broad host distribution of a promiscuous Env gene. *Current Opinion in Virology* **25,** 105–112 (2017).
16. Sugimoto, J., Sugimoto, M., Bernstein, H., Jinno, Y. & Schust, D. A novel human endogenous retroviral protein inhibits cell-cell fusion. *Sci Rep* **3,** 1462–8 (2013).
17. Sugimoto, J. *et al.* Suppressyn localization and dynamic expression patterns in primary human tissues support a physiologic role in human placentation. *Sci Rep* **9,** 19502–12 (2019).
18. Mostafa, A., Abdelwhab, E. M., Mettenleiter, T. C. & Pleschka, S. Zoonotic Potential of Influenza A Viruses: A Comprehensive Overview. *Viruses* **10,** 497 (2018).

19. Sharp, P. M. & Hahn, B. H. Origins of HIV and the AIDS pandemic. *Cold Spring Harb Perspect Med* **1,** a006841–a006841 (2011).

20. Baseler, L., Chertow, D. S., Johnson, K. M., Feldmann, H. & Morens, D. M. The Pathogenesis of Ebola Virus Disease. *Annu Rev Pathol* **12,** 387–418 (2017).

21. Song, Z. *et al.* From SARS to MERS, Thrusting Coronaviruses into the Spotlight. *Viruses* **11,** 59 (2019).

22. Andersen, K. G., Rambaut, A., Lipkin, W. I., Holmes, E. C. & Garry, R. F. The proximal origin of SARS-CoV-2. *Nat. Med.* **100,** 1–3 (2020).

23. Henzy, J. E. & Johnson, W. E. Pushing the endogenous envelope. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.* **368,** 20120506–20120506 (2013).

24. van der Kuyl, A. C., Dekker, J. T. & Goudsmit, J. Discovery of a new endogenous type C retrovirus (FcEV) in cats: evidence for RD-114 being an FcEV(Gag-Pol)/baboon endogenous virus BaEV(Env) recombinant. *J. Virol.* **73,** 7994–8002 (1999).

25. Green, B. J., Lee, C. S. & Rasko, J. E. J. Biodistribution of the RD114/mammalian type D retrovirus receptor, RDR. *J Gene Med* **6,** 249–259 (2004).

26. Haeussler, M. *et al.* The UCSC Genome Browser database: 2019 update. *Nucleic Acids Res.* **47,** D853–D858 (2019).

27. Arora, N., Sadovsky, Y., Dermody, T. S. & Coyne, C. B. Microbial Vertical Transmission during Human Pregnancy. *Cell Host and Microbe* **21,** 561–567 (2017).

28. Kinney, J. S. & Kumar, M. L. Should we expand the TORCH complex? A description of clinical and diagnostic aspects of selected old and new agents. *Clin Perinatol* **15,** 727–744 (1988).

29. Delorme-Axford, E., Sadovsky, Y. & Coyne, C. B. The Placenta as a Barrier to Viral Infections. *Annu Rev Virol* **1,** 133–146 (2014).

30. Ander, S. E., Diamond, M. S. & Coyne, C. B. Immune responses at the maternal-fetal interface. *Sci Immunol* **4,** eaat6114 (2019).

31. Heidmann, A. D. C. L. T., Lavialle, C. & Heidmann, T. From ancestral infectious retroviruses to bona fide cellular genes: Role of the captured syncytins in placentation. *Placenta* **33,** 663–671 (2012).

32. Malfavon-Borja, R. & Feschotte, C. Fighting Fire with Fire: Endogenous Retrovirus Envelopes as Restriction Factors. *J. Virol.* **89,** 4047–4050 (2015).

33. Frank, J. A. & Feschotte, C. Co-option of endogenous viral sequences for host cell function. *Current Opinion in Virology* **25,** 81–89 (2017).

34. Petropoulos, S. *et al.* Single-Cell RNA-Seq Reveals Lineage and X Chromosome Dynamics in Human Preimplantation Embryos. *Cell* **165,** 1012–1026 (2016).

35. Yan, L. *et al.* Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. *Nat. Struct. Mol. Biol.* **20,** 1131–1139 (2013).

36. Niakan, K. K., Han, J., Pedersen, R. A., Simon, C. & Pera, R. A. R. Human pre-implantation embryo development. *Development* **139,** 829–841 (2012).

37. Barakat, T. S. *et al.* Functional Dissection of the Enhancer Repertoire in Human Embryonic Stem Cells. *Cell Stem Cell* **23,** 276–288.e8 (2018).

38. Gao, L. *et al.* Chromatin Accessibility Landscape in Human Early Embryos and Its Association with Evolution. *Cell* **173,** 248–259.e15 (2018).

39. Wu, J. *et al.* Chromatin analysis in human early development reveals epigenetic transition during ZGA. *Nature* **557,** 256–260 (2018).

40. Krendl, C. *et al.* GATA2/3-TFAP2A/C transcription factor network couples human pluripotent stem cell differentiation to trophectoderm with repression of pluripotency. *Proc. Natl. Acad. Sci. U.S.A.* **114,** E9579–E9588 (2017).

41. Tsankov, A. M. *et al.* Transcription factor binding dynamics during human ES cell differentiation. *Nature* **518,** 344–349 (2015).

42. Vento-Tormo, R. *et al.* Single-cell reconstruction of the early maternal-fetal interface in humans. *Nature* **563,** 347–353 (2018).

43. Liu, Y. *et al.* Single-cell RNA-seq reveals the diversity of trophoblast subtypes and patterns of differentiation in the human placenta. *Cell Res.* **28,** 819–832 (2018).

44. Dunn-Fletcher, C. E. *et al.* Anthropoid primate-specific retroviral element THE1B controls expression of CRH in placenta and alters gestation length. *PLoS Biol.* **16,** e2006337 (2018).

45. Kwak, Y.-T., Muralimanoharan, S., Gogate, A. A. & Mendelson, C. R. Human Trophoblast Differentiation Is Associated With Profound Gene Regulatory and Epigenetic Changes. *Endocrinology* **160,** 2189–2203 (2019).

46. Holder, B. S., Tower, C. L., Abrahams, V. M. & Aplin, J. D. Syncytin 1 in the human placenta. *Placenta* **33,** 460–466 (2012).

47. Mi, S. *et al.* Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. *Nature* **403,** 785–789 (2000).

48. Frendo, J. L. *et al.* Direct Involvement of HERV-W Env Glycoprotein in Human Trophoblast Cell Fusion and Differentiation. *Molecular and Cellular Biology* **23,** 3566–3574 (2003).

49. Sandrin, V., Muriaux, D., Darlix, J. L. & Cosset, F. L. Intracellular Trafficking of Gag and Env Proteins and Their Interactions Modulate Pseudotyping of Retroviruses. *J. Virol.* **78,** 7153–7164 (2004).

50. Dottori, M., Tay, C. & Hughes, S. M. Neural development in human embryonic stem cells-applications of lentiviral vectors. *J. Cell. Biochem.* **112,** 1955–1962 (2011).

51. Gropp, M. *et al.* Stable genetic modification of human embryonic stem cells by lentiviral vectors. *Mol. Ther.* **7,** 281–287 (2003).

52. Sakata, M. *et al.* Analysis of VSV pseudotype virus infection mediated by rubella virus envelope proteins. *Sci Rep* **7,** 11607 (2017).

53. Delorme-Axford, E. *et al.* Human placental trophoblasts confer viral resistance to recipient cells. *Proc. Natl. Acad. Sci. U.S.A.* **110,** 12048–12053 (2013).

54. Vidricaire, G., Tardif, M. R. & Tremblay, M. J. The low viral production in trophoblastic cells is due to a high endocytic internalization of the human immunodeficiency virus type 1 and can be overcome by the pro-inflammatory

cytokines tumor necrosis factor-alpha and interleukin-1. *J. Biol. Chem.* **278,** 15832–15841 (2003).

55. Aagaard, L. *et al.* Silencing of endogenous envelope genes in human choriocarcinoma cells shows that envPb1 is involved in heterotypic cell fusions. *J. Gen. Virol.* **93,** 1696–1699 (2012).

56. Sommerfelt, M. A. & Weiss, R. A. Receptor interference groups of 20 retroviruses plating on human cells. *Virology* **176,** 58–69 (1990).

57. de Parseval, N., Casella, J.-F., Gressin, L. & Heidmann, T. Characterization of the Three HERV-H Proviruses with an Open Envelope Reading Frame Encompassing the Immunosuppressive Domain and Evolutionary History in Primates. *Virology* **279,** 558–569 (2001).

58. van Zeijl, M. *et al.* A human amphotropic retrovirus receptor is a second member of the gibbon ape leukemia virus receptor family. *Proc Natl Acad Sci USA* **91,** 1168–1172 (1994).

59. Miller, D. G., Edwards, R. H. & Miller, A. D. Cloning of the cellular receptor for amphotropic murine retroviruses reveals homology to that for gibbon ape leukemia virus. *Proc Natl Acad Sci USA* **91,** 78–82 (1994).

60. Liu, M. & Eiden, M. V. The receptors for gibbon ape leukemia virus and amphotropic murine leukemia virus are not downregulated in productively infected cells. *Retrovirology* **8,** 53–14 (2011).

61. Jobbagy, Z., Garfield, S., Baptiste, L., Eiden, M. V. & Anderson, W. B. Subcellular redistribution of Pit-2 P(i) transporter/amphotropic leukemia virus (A-MuLV) receptor in A-MuLV-infected NIH 3T3 fibroblasts: involvement in superinfection interference. *J. Virol.* **74,** 2847–2854 (2000).

62. Kim, J. W. & Cunningham, J. M. N-linked glycosylation of the receptor for murine ecotropic retroviruses is altered in virus-infected cells. *J. Biol. Chem.* **268,** 16316–16320 (1993).

63. Marin, M., Tailor, C. S., Nouri, A. & Kabat, D. Sodium-dependent neutral amino acid transporter type 1 is an auxiliary receptor for baboon endogenous retrovirus. *J. Virol.* **74,** 8085–8093 (2000).

64. 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 neutral amino acid transporters determine their retroviral receptor functions. *J. Virol.* **77,** 2936–2945 (2003).

65. Hubley, R. *et al.* The Dfam database of repetitive DNA families. *Nucleic Acids Res.* **44,** D81–9 (2016).

66. Perelman, P. *et al.* A molecular phylogeny of living primates. *PLoS Genet* **7,** e1001342 (2011).

67. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution* **24,** 1586–1591 (2007).

68. Malassiné, A. *et al.* Expression of the fusogenic HERV-FRD Env glycoprotein (syncytin 2) in human placenta is restricted to villous cytotrophoblastic cells. *Placenta* **28,** 185–191 (2007).

69. Mallet, F. *et al.* The endogenous retroviral locus ERVWE1 is a bona fide gene involved in hominoid placental physiology. *Proc Natl Acad Sci USA* **101,** 1731–1736 (2004).
70. Bonnaud, B. *et al.* Evidence of selection on the domesticated ERVWE1 env retroviral element involved in placentation. *Molecular Biology and Evolution* **21,** 1895–1901 (2004).
71. de Parseval, N. *et al.* Comprehensive search for intra- and inter-specific sequence polymorphisms among coding envelope genes of retroviral origin found in the human genome: genes and pseudogenes. *BMC Genomics* **6,** 117–11 (2005).
72. Tang, Y. *et al.* Endogenous Retroviral Envelope Syncytin Induces HIV-1 Spreading and Establishes HIV Reservoirs in Placenta. *Cell Rep* **30,** 4528–4539.e4 (2020).
73. Saha, A. *et al.* A trans-dominant form of Gag restricts Ty1 retrotransposition and mediates copy number control. *J. Virol.* **89,** 3922–3938 (2015).
74. Nishida, Y. *et al.* Ty1 retrovirus-like element Gag contains overlapping restriction factor and nucleic acid chaperone functions. *Nucleic Acids Res.* **43,** 7414–7431 (2015).
75. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21 (2013).
76. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30,** 923–930 (2014).
77. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* **36,** 411–420 (2018).
78. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177,** 1888–1902.e21 (2019).
79. Qiu, X. *et al.* Single-cell mRNA quantification and differential analysis with Census. *Nat. Methods* **14,** 309–315 (2017).
80. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9,** 357–359 (2012).
81. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
82. Amemiya, H. M., Kundaje, A. & Boyle, A. P. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Sci Rep* **9,** 9354 (2019).
83. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).
84. Robinson, J. T., Thorvaldsdóttir, H., Nature, W. W.2011. Integrative genomics viewer. *nature.com*
85. Luft, C. *et al.* Application of Gaussia luciferase in bicistronic and non-conventional secretion reporter constructs. *BMC Biochem.* **15,** 14 (2014).
86. Knappskog, S. *et al.* The level of synthesis and secretion of Gaussia princeps luciferase in transfected CHO cells is heavily dependent on the choice of signal peptide. *J. Biotechnol.* **128,** 705–715 (2007).

87. Hubisz, M. J., Pollard, K. S. & Siepel, A. PHAST and RPHAST: phylogenetic analysis with space/time models. *Brief. Bioinformatics* **12,** 41–51 (2011).
88. Ye, J., McGinnis, S. & Madden, T. L. BLAST: improvements for better sequence analysis. *Nucleic Acids Res.* **34,** W6–9 (2006).
89. Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M. & Barton, G. J. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25,** 1189–1191 (2009).
90. Kumar, S., Stecher, G. & Tamura, K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Molecular Biology and Evolution* **33,** 1870–1874 (2016).
91. Kumar, S., Stecher, G., Li, M., and, C. K. M. B.2018. MEGA X: molecular evolutionary genetics analysis across computing platforms. *academic.oup.com*
92. Tamura, K., evolution, M. N. M. B. A.1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *academic.oup.com*
93. Stecher, G., Tamura, K., and, S. K. M. B.2020. Molecular evolutionary genetics analysis (MEGA) for macOS. *academic.oup.com*
94. Wickham, H. 2016. ggplot2: Elegant Graphics for Data Analysis. *Springer-Verlag*

CHAPTER 3
DISCUSSION AND OPEN QUESTIONS

3.1 EVOLUTIONARY ARMS RACES SHAPE HOST-VIRUS EVOLUTION

Viruses and their target hosts are in a persistent evolutionary arms race that has shaped the evolution of virus and host alike. Selection drives the emergence of adaptive traits in the viral genome that allow the invading virus to successfully infect and adapt to the host- cell environment[1,2]. The virus may acquire point mutations, insertions, deletions or structural genome changes that introduce adaptations to viral proteins and regulatory sequences that improve the efficiency of infection, replication, virus release, immune evasion or transmissibility[2,3]. For many RNA viruses, including retroviruses, recombination allows viruses to acquire novel sequences from other viruses and hosts that increase pathogenicity or expand host tropism to new cellular environments or species[4-6].

Within the time-frame of viral infections afflicting a small number of host generations, effective innate and adaptive immune responses must combat these ever-changing invaders. Over evolutionary time, selection will favor adaptive changes to host immune factors that improve the detection of viruses and restriction of the viral life-cycle. In a stroke of evolutionary irony, viruses can provide the host with the means to limit their propagation. Viral regulatory and protein coding sequences that have entered the host germline genome (predominantly retroviruses in vertebrates) and become fixed in the population may be co-opted to combat viral infection[7,8]. While this work predominantly focuses on a single human ERV-derived gene, insights gained from SUPYN may inform future studies that extend to vertebrates as a whole.

3.2 SUPYN LIKELY ACTS THROUGH RECEPTOR INTERFERENCE

Our work indicates SUPYN can restrict infection of the developing fetus by potentially zoonotic Type-D and RD114 retroviruses (RDR). Existing literature[9,10] strongly implies SUPYN restricts SYN1-mediated cell-fusion by directly interacting with ASCT2. Consistent with these reports, our data suggest SUPYN likely interferes with receptor binding by RDRenv and consequent cell entry **(Fig. 2.8)**. Preliminary experiments also suggest secreted SUPYN may confer modest protection to co-cultured cells not expressing SUPYN **(data not shown)**. While these data imply that virions decorated with RDRenv should not be able to bind to ASCT2 when SUPYN is expressed in target cells, our lentiviral infection reporter system does not formally show that SUPYN interferes with RDRenv receptor-binding. It is possible, though unlikely, that SUPYN may interfere with infection at some stage after ASCT2 binding and prior to reporter gene expression. Further experiments, using labeled virions or tagged recombinant RDRenv surface domains, will be necessary to determine if SUPYN expression results in reduced RDRenv binding to ASCT2.

3.3 EVOLUTIONARY CONSEQUENCES OF SUPYN INTERACTION WITH THE ASCT2 RDR-ENV BINDING INTERFACE

Host species that are subject to pervasive and persistent viral infection commonly exhibit rapid evolution in gene products that directly interact with viral proteins[1]. This rapid evolution manifests as sequence variation at the host-virus interface across related species[1]. If the putative interaction between SUPYN and ASCT2 prevents viral receptor binding and entry, we would expect residues at the interface to be under purifying selection. Thus, we would predict that ASCT2 sequence variation in extracellular loop 2 (ECL2) (**Fig. 2.12),** which is the binding region for RDRenv, would be less divergent in Apes compared to OWMs. Similarly, it is possible that species lacking *SUPYN* would be expected to exhibit signatures of rapid evolution at sites within ASCT2-ECL2. Our preliminary sequence analyses across primates, rodents and bats, which have all been infected by RDRs, are consistent with this prediction (**Fig 3.1**). We found that sites in ECL2 are under positive selection across rodents, bats and primates. Though there is a notable absence of sequence variation within Catarrhini, particularly within Apes, where SUPYN emerged and was retained. If the interaction between SUPYN and ASCT2 is evolutionarily significant in the human population, we would expect to see low sequence variation at the respective binding interfaces. Preliminary analysis of ASCT2 sequence variation in the gnomad exome and genome sequencing datasets [https://www.biorxiv.org/content/10.1101/531210v4.article-metrics] indicates there is no evidence of sequence variation above a 1% allele frequency in ASCT2 ECL2. These data are consistent with the hypothesis that SUPYN co-option may have resulted in ASCT2-ECL2 sequence fixation.

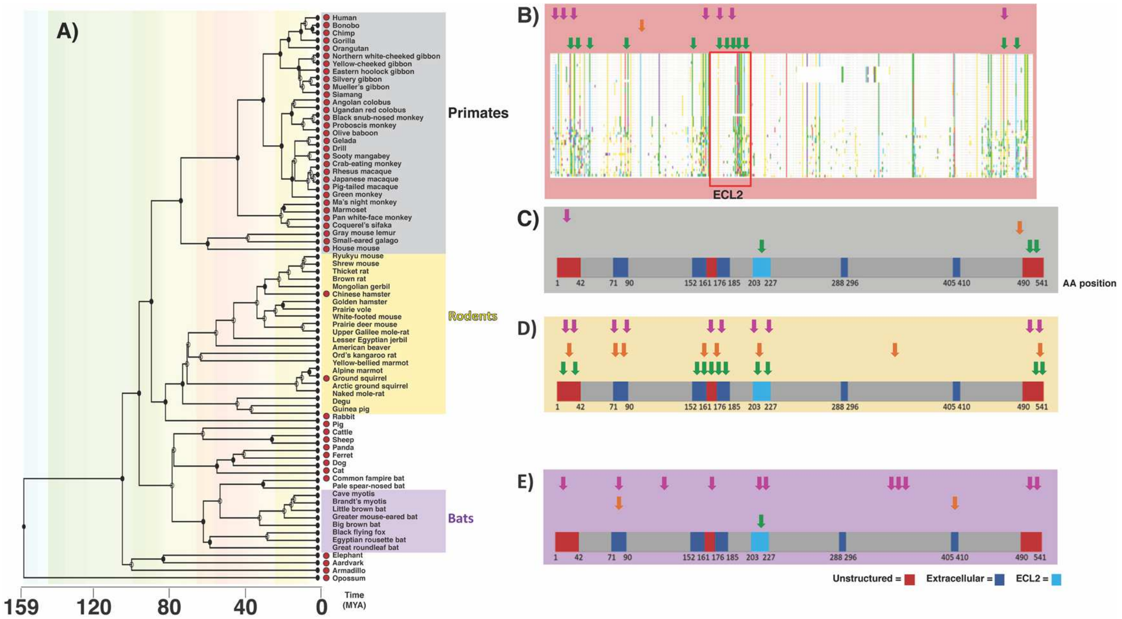

Figure 3.1: Site-specific selection analysis of ASCT2 in mammals
A) Phylogeny of all analyzed species. Red circles denote species used in mammal-wide selection analysis in **(B)**. Primate (purple), rodent (yellow), and bat (purple) clades are highlighted. **(B)** Mammal-wide alignment is shown. Colored shading in alignment represents conservation at a 70% similarity cutoff. **(C, D, E)** Cartoon depictions of ASCT2 open reading frames of primates **(C)**, rodents **(D)** and bats **(E)** are shown with

unstructured (red), extracellular (navy) and ECL2 (light blue) regions highlighted. Sites exhibiting a significant signature of rapid evolution (*p* < 0.05) are indicated by arrows at indicated sequence positions. Arrow color represents a MEME (pink), FEL (orange), and PAML (green) selection analyses.

3.4 POTENTIAL CONSERVATION OF SUPYN EXPRESSION AND FUNCTION

Our evolutionary sequence analyses **(Fig. 2.9)** and infection assays **(Fig. 2.6)** imply *SUPYN* antiviral activity against RDRenv is likely conserved in Apes and partially conserved in OWMs. Preliminary analysis of transcriptome datasets, generated from iPSC culture of, chimp, gorilla, and orangutan, indicate *SUPYN* expression in pluripotent stem cells is conserved in Apes (data not shown). However, these experiments suffer from the following two limitations: (1) *SUPYN* expression has not been extensively characterized in primates. Thus, we do not know if and in what tissues or developmental contexts *SUPYN* is expressed. While *SUPYN* appears to be evolving neutrally in OWMs, it is possible that a subset of OWMs may express functional SUPYN. (2) The antiviral activity of primate SUPYNs were tested within the context of human cells and ASCT2. It is possible that the interaction between endogenously expressed SUPYN and ASCT2 might not result in the same resistance phenotype in the native host. These limitations can be resolved by first characterizing when and where *SUPYN* is expressed during fetal development in catarrhine primates. Once endogenous *SUPYN* expression has been validated, it should then be possible to determine if *SUPYN* is capable of conferring resistance to RDR infection. This may be achieved by co-expressing ASCT2 with SUPYN from individual primates in a heterologous system or by knocking down endogenously expressed SUPYN in primate cells and testing for changes in infection susceptibility. Such experiments will be required to more fully understand the extent of *SUPYN* functional conservation.

3.5 SUPYN MAY RESTRICT GENOME INVASION BY RD-LIKE RETROVIRUSES

Previous work showed that HERV-T co-option likely resulted in the death of the HERV-T gamma-retroviral lineage in humans[11]. Previous reports and our work imply HERVW and HERVH48 are likely ancestral members of the RDR interference group because SYN1 and SUPYN are known to interact with ASCT2[10,12]. The apparent antiviral activity of SUPYN in the developing placenta and potentially within the preimplantation embryo implies that the human genome may be protected from recurrent germline invasion by RDRs. If *SUPYN* has provided evolutionarily significant protection, then the evolutionary conservation of *SUPYN* in Catarrhini and Apes might have resulted in the accumulation of fewer RDR-like insertions compared to genomes lacking *SUPYN*. Further, Ape genomes may be more resistant to RDR invasion than OWM genomes because *SUPYN* has been under functional constraint; whereas *SUPYN* is absent or degraded as a result of neutral evolution in the OWM lineage. A brief search of the literature lends credence to this hypothesis. Grandi et al.[13] found that Rhesus Macaques acquired a larger number of lineage-specific ERVW insertions (n = 66) compared to apes (2-6). These observations are consistent with the idea that SUPYN may have inoculated the ape genome from RDR invasion. To address this question, it would be valuable to identify and compare the number of HERVH48 insertions in the genomes of Apes vs OWMs.

3.6 POTENTIAL MULTIFUNCTIONALITY OF SYNCYTINS

The majority of studies on co-opted EVE-derived genes found in vertebrates describe a single gene function. Recently co-opted EVEs typically restrict virus entry, replication, or assembly[8]. Conversely, genes derived from more ancient EVEs (i.e. those lacking a known extant exogenous counterpart) fulfill a host cell function unrelated to virus restriction[8,14-16]. The dual function of SUPYN in placental development and restriction of RDRenv-mediated entry represents an interesting instance where an EVE-derived protein-coding sequence has seemingly been co-opted for multiple functions. This multifunctionality raises an interesting question regarding the evolution of *SUPYN* and Syncytins as a whole. Is the dual activity of SUPYN a general feature of ERVenv that have been co-opted as a result of their receptor binding activity? *SUPYN* may have initially emerged as an antiviral factor to protect against HERVH48 infection. This need to protect against HERVH48 infection may have provided the evolutionary space for SUPYN to be repurposed as a modulator of SYN1 during placental development. This hypothesis is supported by three observations: (1) *SYN1* was also acquired in the catarrhine lineage. (2) Both *SUPYN* and *SYN1* have been under evolutionary constraint in Apes. (3) *SUPYN* evolved neutrally in OWM where *SYN1* was lost **(Fig 2.9)**.

The dual functionality of SUPYN may be shared by syncytins, which mediate cell fusion in the developing placenta. The interaction between a syncytin and receptor would be expected to result in some degree of resistance to viral infection in the developing placenta. Indeed, our in vitro experiments testing the antiviral activity of overexpressed *SYN1* support this hypothesis **(data not shown)**. Given that virus families like RDRs tend to utilize common target-receptor proteins to gain host-cell entry, it is possible that Syncytins and placentally expressed ERVenv, which are derived from diverse retroviral families[17,18], may protect the developing germline against multiple viruses at once. It would be interesting to see if further placentally expressed ERVenv confer resistance to infection in humans or other mammals.

3.7 IMPLICATIONS FOR EVE CO-OPTION AS ANTIVIRAL FACTORS

SUPYN serves as a proof of principle that ERVenv can function as antiviral restriction factors in humans and implies our genomes may harbor further ERVenv with antiviral activity. The case of *SUPYN* also illustrates that env-coding sequences need not be full-length to be functional. In our preliminary tBLASTn searches of the human genome, we identified ~1700 unique candidate ERVenv with a minimum ORF length of 100aa. We then intersected the location of our BLAST-hits with the UCSC 30-primate species conserved element track[19], to identify ORFs with some evidence of evolutionary constraint. Using this approach, we found 30 env ORFs overlapping with conserved elements, 13 of which overlap with annotated genes (Table3.1). These sequences stem from beta-, gamma-, and spuma-like retroviruses (ERV1, HERVL, ERVK). Seven of these ORFs exhibit a significant signature of purifying selection, four of which were previously annotated as ERVenv-derived genes. These results imply that our search identified three novel ERV ORFs that may have some host-cell function, perhaps in restricting retroviral infection. The low number of novel ERV ORFs suggests that using conserved elements as filter is conservative but capable of identifying novel ORFs with a somewhat robust signature of evolutionary constraint. This approach can identify *rec*-encoding ORFs, which are chaperone-proteins that canonically ensure unspliced retroviral RNAs are accurately trafficked out of the nucleus[20]. In fact, our search revealed one *rec*-coding sequence in our list of 35 conserved ORFs.

Apart from these more deeply conserved env, our genome is also littered with young HERV insertions that encode intact env sequences where signatures of evolutionary constraint cannot be detected, due to lack of evolutionary time**.** HERVs that entered the genomes of Apes or humans may be expressed and have the capacity to restrict retroviral infection. For example, HERVH and HERVK(HML2) insertions have been reported to be transcribed during early embryonic development and in tissues other than the placenta[21-24]. HERVenv expression has also been described in healthy and cancerous immune cells[25-28]. In many cases, HERV expression has been linked to autoimmune diseases[29-31] like multiple sclerosis[32] and lupus[33,34]**.** Localized HERVenv expression in immune cells would be consistent with a potential function as a restriction factor because immune cells are commonly targeted by retroviruses[35-37]. It is possible that HERVenv expression, like HERVK or -W, in immune cells may provide an added layer of targeted resistance to retroviral infection. In fact, a recent report implied that HERVK(HML-2)env may be capable of interfering with HIV replication[38], though, it is unclear by what mechanism this env functions or whether this activity is significant *in vivo*. Analysis of large tissue-level gene expression datasets, like GTEx (https://gtexportal.org/home/), and other transcriptome datasets for evidence of envORF expression is likely to identify further candidate env that can be experimentally tested for host-cell function. These analyses can be supplemented by mining existing proteome datasets, such as the Protein Atlas[39], to identify protein-level envORF expression.

Beyond ERVenv, our genomes also contain remnants of retroviral proteins, like *gag* and *pro*, as well as isolated protein-coding sequences derived from other non-retroviral families[8,40-42]. Work conducted in other eukaryotic systems has shown such sequences can be repurposed to serve a myriad of functions including defense against exogenous viral infection. *Gag*-derived proteins, encoded in yeast and sheep, have been repurposed to interfere with capsid assembly[8,15,16]. More ancient *gag*-derived *arc* genes have been shown to play a role in neuron signaling in both vertebrates[15] and invertebrates[16] by packaging RNA in capsid like structures that are transmitted between neurons. ERV-encoded RNA chaperones, like HERVK(HML2) *rec*, have also been suggested to interfere with influenza replication in ESCs[21]. Though it is unclear how rec functions in this capacity. Non-retroviral endogenous borna-like nucleoprotein has been shown to interfere with borna-disease virus replication in squirrel cells[43]. These individual examples imply that viral protein coding sequences other than env can confer resistance to exogenous viruses.

3.8 CONCLUSION

Beyond studies focusing on individual examples of EVE co-option, little work has been done to systematically screen for EVE-derived sequences with host cell function. Our genomes may be equipped with many more EVE-encoded genes with undiscovered functions. While this work has predominantly focused on retroviral *env*, particularly within the context of antiviral activity, further EVE-derived protein-coding sequences, including gag, pol, helicase and reverse transcriptase, may have been preserved by natural selection to serve as restriction factors or have further undiscovered host-cell functions. By combining sequence-homology based genome searches with evolutionary and expression data, future studies will likely identify further EVE-derived candidate genes that can be subsequently experimentally tested. This integrative approach is likely to be applicable not only to humans but can be applied to any available vertebrate genome with existing expression data. EVE-co-option is a complex process that can take many forms. It is clear that the evolutionary pressure posed by exogenous and potentially zoonotic viruses likely results in the emergence of novel EVE-derived restriction factors. This work illustrates how endogenous viral sequences have and are likely to continue to contribute to our antiviral defenses.

REFERENCES

1. Daugherty, M. D. & Malik, H. S. Rules of engagement: molecular insights from host-virus arms races. *Annu. Rev. Genet.* **46,** 677–700 (2012).
2. Simmonds, P., Aiewsakun, P. & Katzourakis, A. Prisoners of war - host adaptation and its constraints on virus evolution. *Nat. Rev. Microbiol.* **54,** 156–328 (2018).
3. Vijaykrishna, D., Mukerji, R. & Smith, G. J. D. RNA Virus Reassortment: An Evolutionary Mechanism for Host Jumps and Immune Evasion. *PLoS Pathog* **11,** e1004902 (2015).
4. Diehl, W. E., Patel, N., Halm, K. & Johnson, W. E. Tracking interspecies transmission and long-term evolution of an ancient retrovirus using the genomes of modern mammals. *Elife* **5,** e12704 (2016).
5. Henzy, J. E. & Johnson, W. E. Pushing the endogenous envelope. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.* **368,** 20120506–20120506 (2013).
6. Worobey, M. & Holmes, E. C. Evolutionary aspects of recombination in RNA viruses. *J. Gen. Virol.* **80 (Pt 10),** 2535–2543 (1999).
7. Chuong, E. B., Elde, N. C. & Feschotte, C. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. *Science* **351,** 1083–1087 (2016).
8. Frank, J. A. & Feschotte, C. Co-option of endogenous viral sequences for host cell function. *Current Opinion in Virology* **25,** 81–89 (2017).
9. Sugimoto, J. *et al.* Suppressyn localization and dynamic expression patterns in primary human tissues support a physiologic role in human placentation. *Sci Rep* **9,** 19502–12 (2019).
10. Sugimoto, J., Sugimoto, M., Bernstein, H., Jinno, Y. & Schust, D. A novel human endogenous retroviral protein inhibits cell-cell fusion. *Sci Rep* **3,** 1462–8 (2013).
11. Blanco-Melo, D., Gifford, R. J. & Bieniasz, P. D. Co-option of an endogenous retrovirus envelope for host defense in hominid ancestors. *Elife* **6,** 11 (2017).
12. Blond, J. L. *et al.* An envelope glycoprotein of the human endogenous retrovirus HERV-W is expressed in the human placenta and fuses cells expressing the type D mammalian retrovirus receptor. *J. Virol.* **74,** 3321–3329 (2000).
13. Grandi, N., Cadeddu, M., Blomberg, J., Mayer, J. & Tramontano, E. HERV-W group evolutionary history in non-human primates: characterization of ERV-W orthologs in Catarrhini and related ERV groups in Platyrrhini. *BMC Evol. Biol.* **18,** 6 (2018).
14. Heidmann, O. *et al.* HEMO, an ancestral endogenous retroviral envelope protein shed in the blood of pregnant women and expressed in pluripotent stem cells and tumors. *Proc. Natl. Acad. Sci. U.S.A.* **114,** E6642–E6651 (2017).
15. Pastuzyn, E. D. *et al.* The Neuronal Gene Arc Encodes a Repurposed Retrotransposon Gag Protein that Mediates Intercellular RNA Transfer. *Cell* **172,** 275–288.e18 (2018).

16. Ashley, J. *et al.* Retrovirus-like Gag Protein Arc1 Binds RNA and Traffics across Synaptic Boutons. *Cell* **172,** 262–274.e11 (2018).

17. Funk, M. *et al.* Capture of a hyena-specific retroviral envelope gene with placental expression associated in evolution with the unique emergence among carnivorans of hemochorial placentation in Hyaenidae. *J. Virol.* (2018). doi:10.1128/JVI.01811-18

18. de Parseval, N. *et al.* Comprehensive search for intra- and inter-specific sequence polymorphisms among coding envelope genes of retroviral origin found in the human genome: genes and pseudogenes. *BMC Genomics* **6,** 117–11 (2005).

19. Haeussler, M. *et al.* The UCSC Genome Browser database: 2019 update. *Nucleic Acids Res.* **47,** D853–D858 (2019).

20. Magin, C., Lower, R. & Lower, J. cORF and RcRE, the Rev/Rex and RRE/RxRE homologues of the human endogenous retrovirus family HTDV/HERV-K. *J. Virol.* **73,** 9496–9507 (1999).

21. Grow, E. J. *et al.* Intrinsic retroviral reactivation in human preimplantation embryos and pluripotent cells. *Nature* **522,** 221–225 (2015).

22. Göke, J. *et al.* Dynamic transcription of distinct classes of endogenous retroviral elements marks specific populations of early human embryonic cells. *Cell Stem Cell* **16,** 135–141 (2015).

23. Wang, J. *et al.* Primate-specific endogenous retrovirus-driven transcription defines naive-like stem cells. *Nature* **516,** 405–409 (2014).

24. Izsvák, Z., Wang, J., Singh, M., Mager, D. L. & Hurst, L. D. Pluripotency and the endogenous retrovirus HERVH: Conflict or serendipity? *Bioessays* **38,** 109–117 (2016).

25. Brinzevich, D. *et al.* HIV-1 interacts with human endogenous retrovirus K (HML-2) envelopes derived from human primary lymphocytes. *J. Virol.* **88,** 6213–6223 (2014).

26. Depil, S., Roche, C., Dussart, P. & Prin, L. Expression of a human endogenous retrovirus, HERV-K, in the blood cells of leukemia patients. *Leukemia* **16,** 254–259 (2002).

27. Tamura, N., Iwase, A., Suzuki, K., Maruyama, N. & Kira, S. Alveolar macrophages produce the Env protein of a human endogenous retrovirus, HERV-E 4-1, in a subgroup of interstitial lung diseases. *Am. J. Respir. Cell Mol. Biol.* **16,** 429–437 (1997).

28. Hsiao, F. C., Lin, M., Tai, A., Chen, G. & Huber, B. T. Cutting edge: Epstein-Barr virus transactivates the HERV-K18 superantigen by docking to the human complement receptor 2 (CD21) on primary B cells. *J. Immunol.* **177,** 2056–2060 (2006).

29. Brütting, C., Emmer, A., Kornhuber, M. E. & Staege, M. S. Cooccurrences of Putative Endogenous Retrovirus-Associated Diseases. *Biomed Res Int* **2017,** 7973165–11 (2017).

30. Nadeau, M.-J., Manghera, M. & Douville, R. N. Inside the Envelope: Endogenous Retrovirus-K Env as a Biomarker and Therapeutic Target. *Front Microbiol* **6,** 1244 (2015).

31. Nakagawa, K. & Harrison, L. C. The potential roles of endogenous retroviruses in autoimmunity. *Immunol. Rev.* **152,** 193–236 (1996).
32. García-Montojo, M. *et al.* Syncytin-1/HERV-W envelope is an early activation marker of leukocytes and is upregulated in multiple sclerosis patients. *Eur. J. Immunol.* **10,** 127 (2020).
33. Tokuyama, M. *et al.* ERVmap analysis reveals genome-wide transcription of human endogenous retroviruses. *Proc. Natl. Acad. Sci. U.S.A.* **115,** 12565–12572 (2018).
34. Bengtsson, A. *et al.* Selective antibody reactivity with peptides from human endogenous retroviruses and nonviral poly(amino acids) in patients with systemic lupus erythematosus. *Arthritis Rheum.* **39,** 1654–1663 (1996).
35. Wilen, C. B., Tilton, J. C. & Doms, R. W. HIV: cell binding and entry. *Cold Spring Harb Perspect Med* **2,** a006866–a006866 (2012).
36. Coffin, J. M., Hughes, S. H. & Varmus, H. E. Retroviruses. (1997).
37. Montiel, N. A. An updated review of simian betaretrovirus (SRV) in macaque hosts. *J. Med. Primatol.* **39,** 303–314 (2010).
38. Terry, S. N. *et al.* Expression of HERV-K108 envelope interferes with HIV-1 production. *Virology* **509,** 52–59 (2017).
39. Thul, P. J. *et al.* A subcellular map of the human proteome. *Science* **356,** eaal3321 (2017).
40. Myers, K. N. *et al.* The bornavirus-derived human protein EBLN1 promotes efficient cell cycle transit, microtubule organisation and genome stability. *Sci Rep* **6,** 35548–12 (2016).
41. Horie, M. *et al.* Endogenous non-retroviral RNA virus elements in mammalian genomes. *Nature* **463,** 84–87 (2010).
42. Agut, H., Bonnafous, P. & Gautheret-Dejean, A. Laboratory and clinical aspects of human herpesvirus 6 infections. *Clin. Microbiol. Rev.* **28,** 313–335 (2015).
43. Fujino, K., Horie, M., Honda, T., Merriman, D. K. & Tomonaga, K. Inhibition of Borna disease virus replication by an endogenous bornavirus-like element in the ground squirrel genome. *Proc. Natl. Acad. Sci. U.S.A.* **111,** 13175–13180 (2014).