Title

Evolution and antiviral activity of a human protein of retroviral origin

Summary

We found that the human genome expresses a vast pool of envelope sequences of retroviral origin and provide proof of principle that such proteins can restrict zoonotic viruses.

Authors

John A. Frank[1], Manvendra Singh[1], Harrison B. Cullen[1], Raphael A. Kirou[1], Meriem Benkaddour-Boumzaouad[2], Jose L. Cortes[2,3], Jose Garcia-Perez[2,4], Carolyn B. Coyne[5], Cédric Feschotte[1*]

Affiliations

[1]Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY, USA

[2]GENYO. Centre for Genomics and Oncological Research: Pfizer/University of Granada/Andalusian Regional Government, PTS Granada, Spain.

[3]Eppendorf Iberica, Spain.

[4]MRC-Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Western General Hospital, Edinburgh, UK.

[5]Department of Molecular Genetics and Microbiology, Duke University School of Medicine, Durham, NC, USA

[*]Corresponding Author

Abstract

Endogenous retroviruses are abundant components of mammalian genomes descended from ancient germline infections. In several mammals, the envelope proteins encoded by these elements protect against exogenous viruses, but this activity has not been documented in human. We report that our genome harbors a large pool of envelope-derived sequences with the potential to restrict retroviral infection. To further test this, we characterize in detail the envelope-derived protein, *Suppressyn*. We found that *Suppressyn* is expressed in preimplantation embryos and developing placenta using its ancestral retroviral promoter. Restriction assays in cell culture show that *Suppressyn*, and its hominoid orthologs, can restrict infection by extant mammalian type D retroviruses. Our data support a generalizable model of retroviral envelope cooption for host immunity and genome defense.

Viruses pose a persistent threat to human health and may be potent evolutionary drivers of immune adaptation (*1–3*). Thus, it is critical to identify host factors that protect against viral infection. Endogenous retroviruses (ERV) are abundant components of mammalian genomes that descend from ancient retroviral germline infections (*1*). ERV sequences are a source of novel genes in evolution (*2*). ERV-derived envelopes (*env*) are particularly interesting because they can serve as antiviral factors by binding to and competing for target cell receptors of exogenous viruses (*1*). Env-derived proteins restricting against contemporary infectious retroviruses have been documented in mouse (Fv4), cat (Refrex-1), and sheep (enJSS6A1) but not in humans (*1*, *2*, *4*). Importantly, the receptor-binding surface domain of an *env* can be sufficient to restrict infection by this interference mechanism (*4*). Because the human genome contains thousands of remnant

46 retroviral sequences with coding potential, we hypothesized that it includes envelope sequences
47 with the capacity to restrict against modern viruses.
48
49 To assess the pool of potentially antiviral sequences, we scanned the human genome for *env*-
50 derived open reading frames (*envORF)* that encode at least 70 aa and are predicted to include the
51 receptor-binding surface domain (see Methods). This search identified a total of 1,507 unique
52 *envORFs,* including ~20 *env*-derived sequences currently annotated as human genes such as
53 *Supppressyn* (*SUPYN*), *Syncytin-1* and *Syncytin-2* (*5*, *6*). This number is consistent with a previous
54 genome-wide scan for ORFs derived from endogenous viral elements predicting 1,731 ORFs with
55 homology to retroviral envelopes (*7*). Next, we mined transcriptome datasets generated from
56 human embryos and various adult tissues **(Table S1)** and observed that ~44% of *envORF*s
57 (668/1507) showed evidence of RNA expression in at least one of the cell types surveyed **(Fig 1a).**
58 These analyses revealed three general trends about expressed *envORFs*: (i) like known *env*-derived
59 genes, *envORFs* exhibit tissue-specific expression patterns **(Fig 1a; fig S1, S2)**; (ii) the majority
60 of *envORFs* are expressed during early human development, often in stem and progenitor germ
61 cells, and/or in placenta **(Fig 1a; fig S1)**; (iii) with the exception of brain, *envORFs* are rarely
62 expressed in differentiated tissues under normal conditions **(Fig 1a; fig S1, S2)**. Since antiviral
63 factors are generally expressed in immune cells and/or induced upon immune stimulation or
64 infection, we also profiled *envORF* expression using transcriptome datasets generated from a
65 variety of immune cells, including resting, stimulated, and HIV-infected cells **(Table S1)**. These
66 analyses revealed that a substantial fraction of *envORFs* (145 loci from multiple families) are
67 expressed in immune cells and tend to be induced by immune simulation **(Fig 1b; fig S3).**
68 Consistent with previous reports (*8*), we also observed that a subset of *envORFs* are induced upon

69 HIV infection **(Fig 1b, c; fig. S3, S4)**. Together, these data suggest that the human genome harbors
70 a vast reservoir of *env*-derived sequences with coding potential for receptor-binding proteins, many
71 of which are expressed in a tissue-specific or infection-inducible fashion, suggesting that they may
72 function as antiviral factors.
73

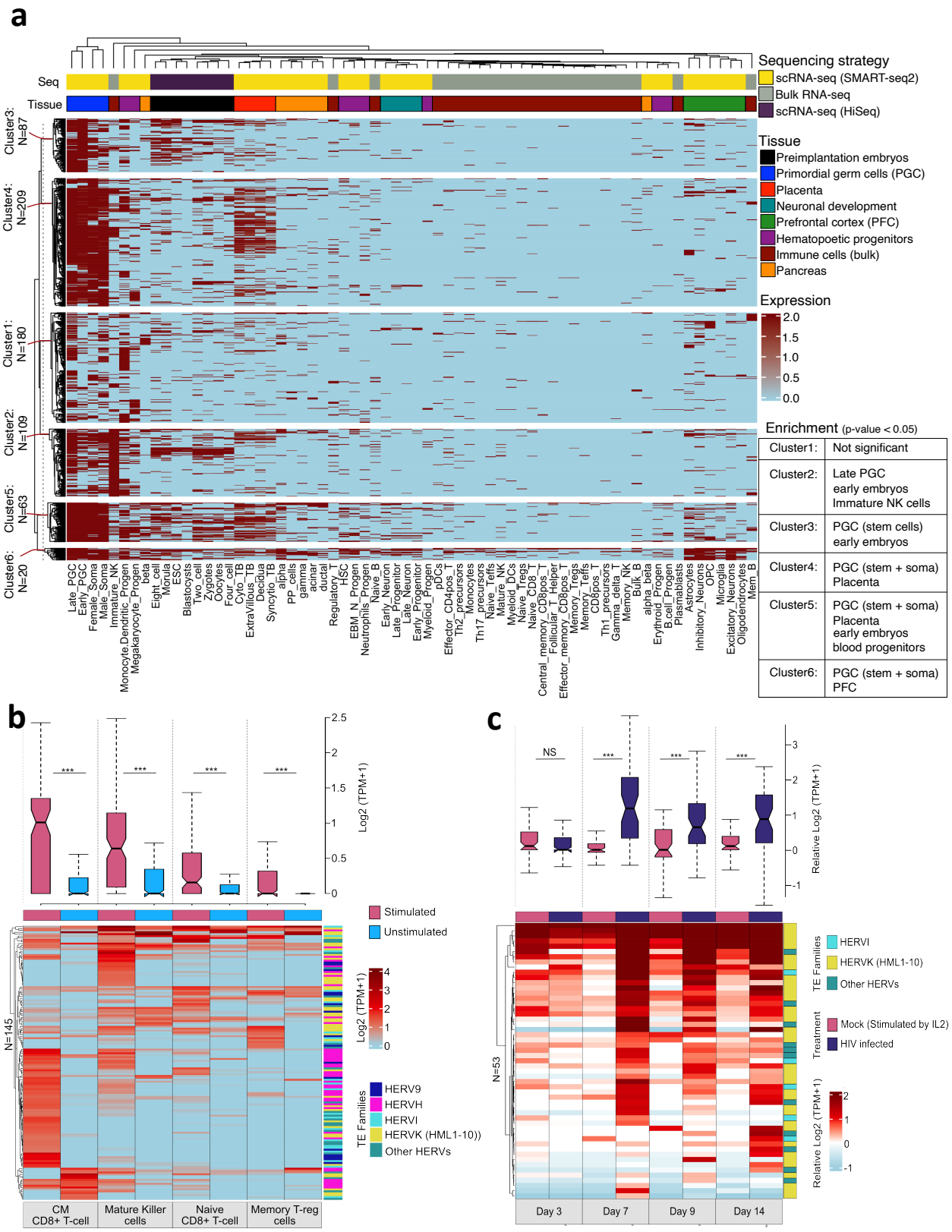


Figure 1: Expression profile of env-derived transcripts over a subset of human cell types.

76 **(a-c)** Heatmaps show *envORF* expression in 66 distinct cell-types from 8 independent datasets (a),
77 and in response to stimulation (b) and HIV infection in immune cells (c). Rows and columns
78 represent individual *envORF* loci and cell identity respectively. *EnvORF* loci were included in
79 heatmap if log2 CPM > 1 in at least one cell type. Bars located above heatmaps denote sequencing
80 strategy (a, top), tissue source (a: bottom) or treatment (b, c) with the same color scheme shown
81 in figures S1, S3, and S4 respectively. Rows and columns are ordered by hierarchical clustering
82 based on *envORF* expression. Significant *envORF* enrichment in indicated clusters was calculated
83 by hypergeometric test. **(b, c)** Boxplots represent the distribution of envORF expression levels
84 relative to unstimulated (b) and Mock-infected cells (c) (NS: not significant; ***$p < 0.01$;
85 Wilcoxon rank-sum test)
86

87 As a paradigm to test the antiviral activity of an envORF, we focused on SUPYN for two primary
88 reasons. First, SUPYN was reported to be expressed in the developing placenta, as we observe for
89 many envORFs, a tissue that can be vulnerable to viral infection and a barrier against vertical
90 transmission to the fetus (*9*). Second, *SUPYN* lacks a transmembrane domain but retains the ability
91 to bind the amino acid transporter ASCT2 (also known as SLC1A5), which is the receptor for a
92 diverse group of retroviruses, including endogenous (e.g. HERV-W) and exogenous viruses such
93 as RD114 in cat and Simian Retrovirus (SRV) in Old World monkeys (*5*, *10–12*). These viruses
94 are collectively referred to as RD114 and Type D (RDR) retroviruses (*12*). Thus, we hypothesized
95 that SUPYN might restrict RDR retroviruses by virtue of its binding to ASCT2 .
96

97 To obtain a more detailed view of *SUPYN* expression and regulation in adult tissues and during
98 human development, we analyzed publicly available bulk and single-cell RNA-seq, ATAC-seq,
99 DNase-seq, and ChIP-seq datasets generated from adult tissues, preimplantation embryos, human
100 embryonic stem cell lines (hESC), *in vitro* trophoblast (TB) differentiation models, and placenta
101 explants isolated at multiple stages of pregnancy **(Table S1).** Apart from testis and cerebellum,
102 *SUPYN* expression is generally low (<1 $\log_2$ (TPM+1)) or absent in adult tissues (**fig S2**).
103 Conversely, *SUPYN* is expressed at high levels in preimplantation embryos from the 8-cell stage

104 of development **(Fig 2a, fig S5a)**. Accordingly, the *SUPYN* promoter region, including the long
105 terminal repeat (LTR) of the ancestral HERVH48 provirus, is marked by several peaks of
106 accessible chromatin from the 8-cell to blastocyst stages **(fig S5b)**. In hESCs, *SUPYN* RNA is also
107 abundant (fig 2a) and its promoter region is marked by histone modifications characteristic of
108 transcriptionally active chromatin (H3K4me1, H3K27ac) and bound by core pluripotency (OCT4,
109 NANOG, KLF4, SMAD1) and self-renewal (SRF, OTX2) transcription factors **(Fig 2b)**. During
110 hESC to TB differentiation, pluripotency factors (NANOG, OCT4) occupying the *SUPYN*
111 promoter region are replaced by TB-specific transcription factors (TFAP2A, GATA3); and active
112 chromatin marks (H3K27ac, H3K4me3, H3K9ac) are maintained across all analyzed TB lineages
113 **(Fig 2b)**. These observations indicate *SUPYN* expression persists through the TB differentiation
114 process. We next mined single-cell RNA-seq data generated from placenta at multiple
115 developmental stages. After classifying cell clusters based on known markers, we found *SUPYN*
116 and *ASCT2* expression was relatively high in cytotrophoblasts (CTB) and extra-villous
117 trophoblasts (EVTB), but also detectable in syncytiotrophoblasts (STB) **(Fig 2c-f, fig S6)**. To
118 confirm these transcriptomic observations, we performed immunostaining of SUPYN in
119 preimplantation embryos as well as in second and third trimester placentas. In early embryos,
120 SUPYN is primarily detectable in the trophectoderm, which will give rise to the placenta, and in
121 some OCT4-expressing cells, which mark pluripotent stem cells of blastocysts **(Fig 2g, fig S7)**. In
122 the placenta, SUPYN is widely expressed in STB and potentially CTB of second trimester
123 placental villi **(Fig 2h, fig S8).** The combined transcriptome, genome regulatory data and
124 immunohistochemical staining firmly establish that *SUPYN* is expressed throughout human fetal
125 development (**Fig 2a, 2e, fig S6-8)**. These analyses also indicate that *SUPYN* transcription
126 consistently initiates from its ancestral LTR promoter **(Fig 2b, fig S5b, S13)**.

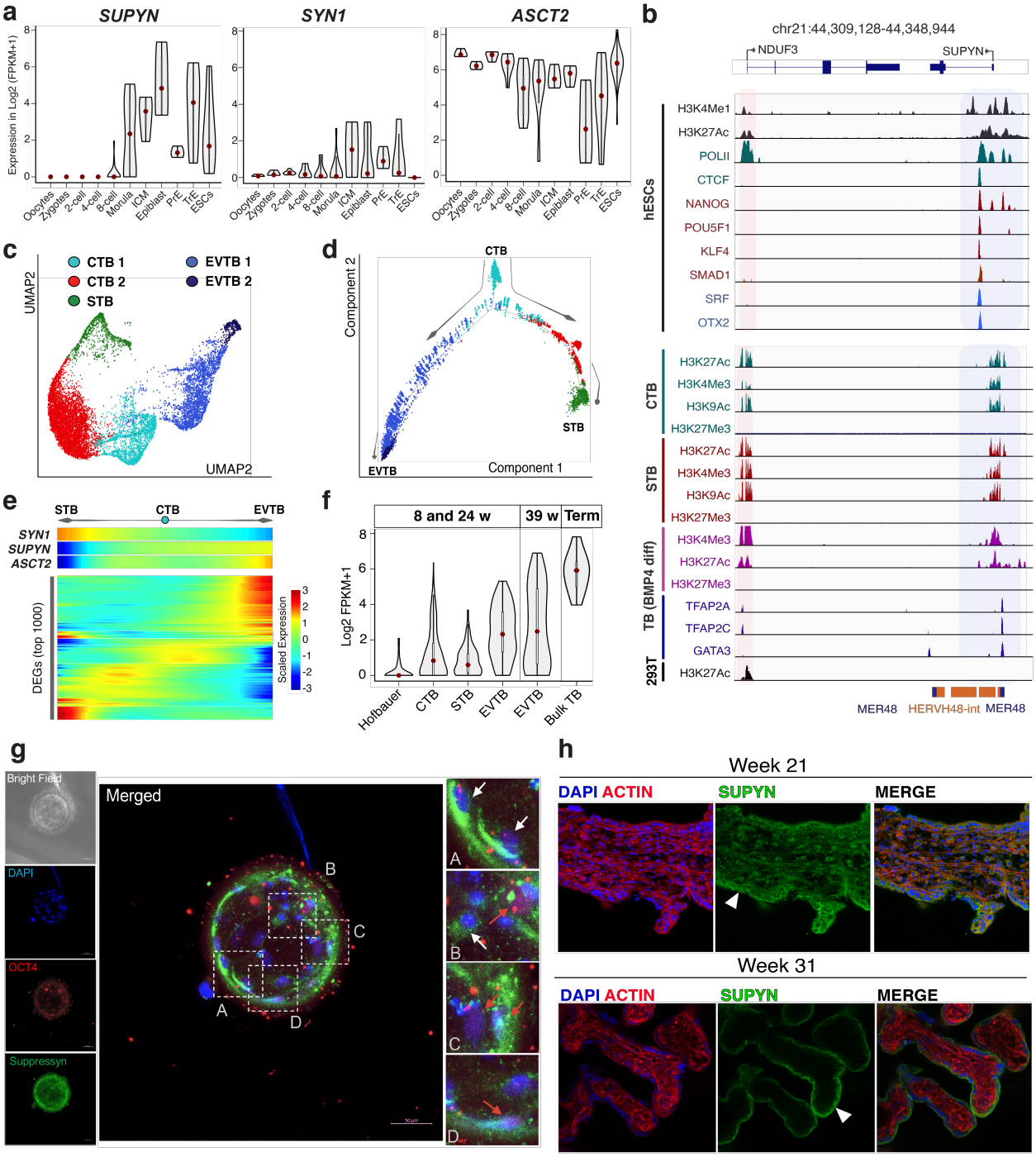


Figure 2: Pluripotency and placentation regulatory factors drive SUPYN expression during fetal development.

(a) Violin plots summarizing *SUPYN*, *SYN1* and *ASCT2* expression in human preimplantation embryo and ESC single-cell RNA-seq data. **(b)** Genome browser view of the regulatory elements surrounding the *SUPYN* locus in hESCs, cyto- (CTB), syncytio- (STB), BMP4-differentiated trophoblasts (TB), and 293T cells. ChIP-seq profiles are shown for indicated transcription factors

135 and histone modifications with shaded area highlighting regions of active chromatin. **(c)** UMAP
136 visualization of TB cell clusters, shown in figure S6. **(d)** Monocle2 pseudotime analysis of cell
137 clusters in (c) illustrates the developmental trajectory of CTBs that give rise to STB and EVTB
138 respectively. Color codes in (c) and (d) denote cell identity. **(e)** Heatmap represents the top 1000
139 differentially expressed genes (row) of single cells (column) and are sorted according to
140 pseudotime analyses in (c) and (d). Cells are ordered according to the pseudotime progression of
141 CTB (middle) to STB (left) and EVTB (right). SYN1, SUPYN, and ASCT2 were fetched from the
142 heatmap below. **(f)** Violin plots denote single-cell *SUPYN* expression in multiple placental-cell
143 lineages throughout indicated pregnancy stages. **(g,h)** Confocal microscopy of human
144 preimplantation embryos (g) and placental villus explants (h) stained for SUPYN (green), OCT4
145 (red), ACTIN (red) and DAPI (blue). In (g), Trophectorderm (white arrows) and OCT4-expressing
146 (red arrows) cells are highlighted in subpanels (A-C). In (h), STBs are marked by arrowheads.
147

148 The expression profile of SUPYN is consistent with the hypothesis that it could protect the
149 developing embryo and germline against RDR infection. To test this hypothesis, we first examined
150 whether human placenta-derived Jar and JEG3 cell lines, and H1 hESC, which robustly express
151 SUPYN and ASCT2 **(Fig 1a, 2a; fig S9, S10)** (*10*, *13*) are resistant to RDR env-mediated infection.
152 We generated HIV-EGFP viral particles pseudotyped with either the feline RD114 env (HIV-
153 RD114) or the glycoprotein G of vesicular stomatitis virus (HIV-VSVg), which uses the LDL
154 receptor (*14*), to monitor the level of infection in cell culture based on EGFP expression **(Fig 3a,**
155 **fig S11)**. These experiments revealed that Jar, JEG3, and H1 cells were susceptible to HIV-VSVg,
156 as previously reported (*15*, *16*), but highly resistant to HIV-RD114 infection **(Fig 3b, c).** By
157 contrast, 293T cells, which do not express *SUPYN* **(fig S5a)**, were similarly susceptible to infection
158 by HIV-RD114 and HIV-VSVg **(Fig 3b, c)**. To test whether SUPYN contributes to the HIV-
159 RD114 resistance phenotype, we repeated these infection experiments in Jar cells engineered to
160 stably express short hairpin RNAs (shRNA) depleting ~80% of *SUPYN* mRNA **(fig S10a).**
161 Depletion of *SUPYN* in Jar cells resulted in a significant increase in susceptibility to HIV-RD114
162 infection but not HIV-VSVg **(Fig 3d**). To confirm that SUPYN expression confers RD114
163 resistance and control for potential off-target effects of *SUPYN*-targeting shRNAs, we transfected

164 Jar-shSUPYN cells with two shRNA-resistant, HA-tagged SUPYN rescue constructs and
165 examined their susceptibility to HIV-RD114 infection. Briefly, the shRNA-targeted SUPYN
166 signal peptide sequence was replaced with a luciferase (SUPYN-lucSP) or modified signal peptide
167 sequence (SUPYN-rescSP) that disrupts shRNA-binding but retains codonidentity (see Methods).
168 Transfection with either SUPYN-rescSP or SUPYN-lucSP restored a significant level of resistance
169 to HIV-RD114 infection **(Fig 3e).** Western blot analysis of transfected cell lysates showed
170 SUPYN-rescSP was more abundantly expressed than SUPYN-lucSP **(Fig 3f)**, which may account
171 for the stronger HIV-RD114 resistance conferred by the former **(Fig 3e).** These results suggest
172 that SUPYN expression is at least in part responsible for preventing Jar cells from RD114env-
173 mediated infection.
174
175 To test if SUPYN expression alone is sufficient to confer protection against RD114env-mediated
176 infection, we transfected 293T cells with unmodified or HA-tagged SUPYN overexpression
177 constructs and subsequently infected with HIV-RD114 and HIV-VSVg respectively. As a positive
178 control, we also transfected 293T cells with an unmodified RD114env overexpression construct,
179 which is predicted to confer resistance to HIV-RD114. Expression of either SUPYN or RD114env
180 resulted in ~80% reduction in the level of HIV-RD114 infection **(Fig 3g; fig S12a)** but had no
181 significant effect on HIV-VSVg infectivity **(fig S12b)**. Taken together, our knockdown and
182 overexpression experiments indicate SUPYN expression is sufficient to confer resistance to
183 RD114env-mediated infection.
184
185 Our RD114env-specific resistance phenotype **(Fig 3g, fig S12a, b)** suggests SUPYN restricts viral
186 entry via a mechanism of receptor interference. If so, this protective effect should extend to

187 infection mediated by other RDRenv (*11*, *17*, *18*). To test this, we generated HIV-GFP reporter
188 virions pseudotyped with Squirrel Monkey Retrovirus env (HIV-SMRVenv) (*11*) **(Fig 3h)** and
189 infected 293T cells previously transfected with SUPYN, SMRVenv (as a positive control) or an
190 empty vector. Cells expressing SUPYN or SMRVenv showed an ~80% reduction of HIV-
191 SMRVenv infected cells **(Fig 3h)**. Thus, SUPYN can restrict infection mediated by multiple
192 RDRenv.
193

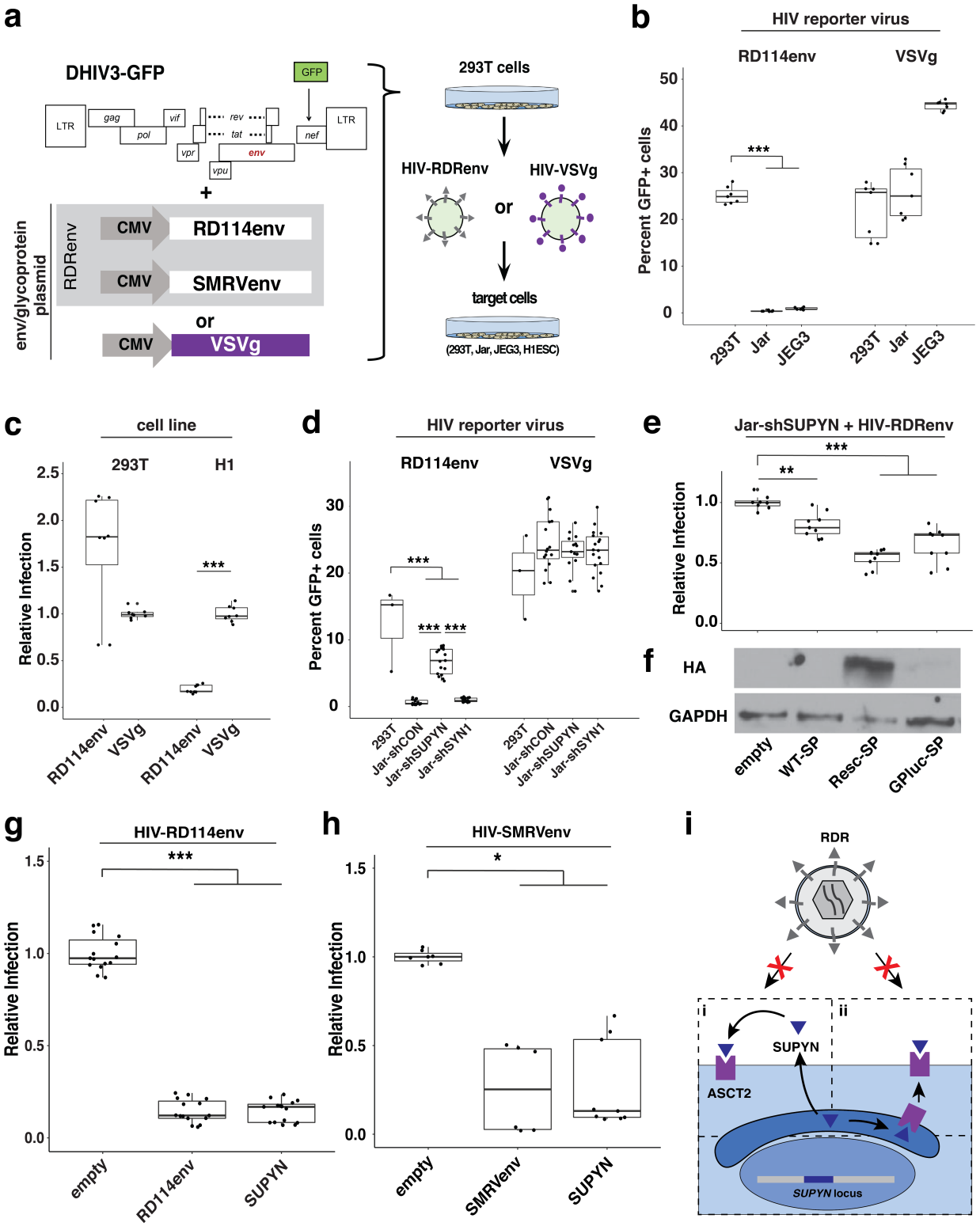


Figure 3: SUPYN confers resistance to RD114 env-mediated infection

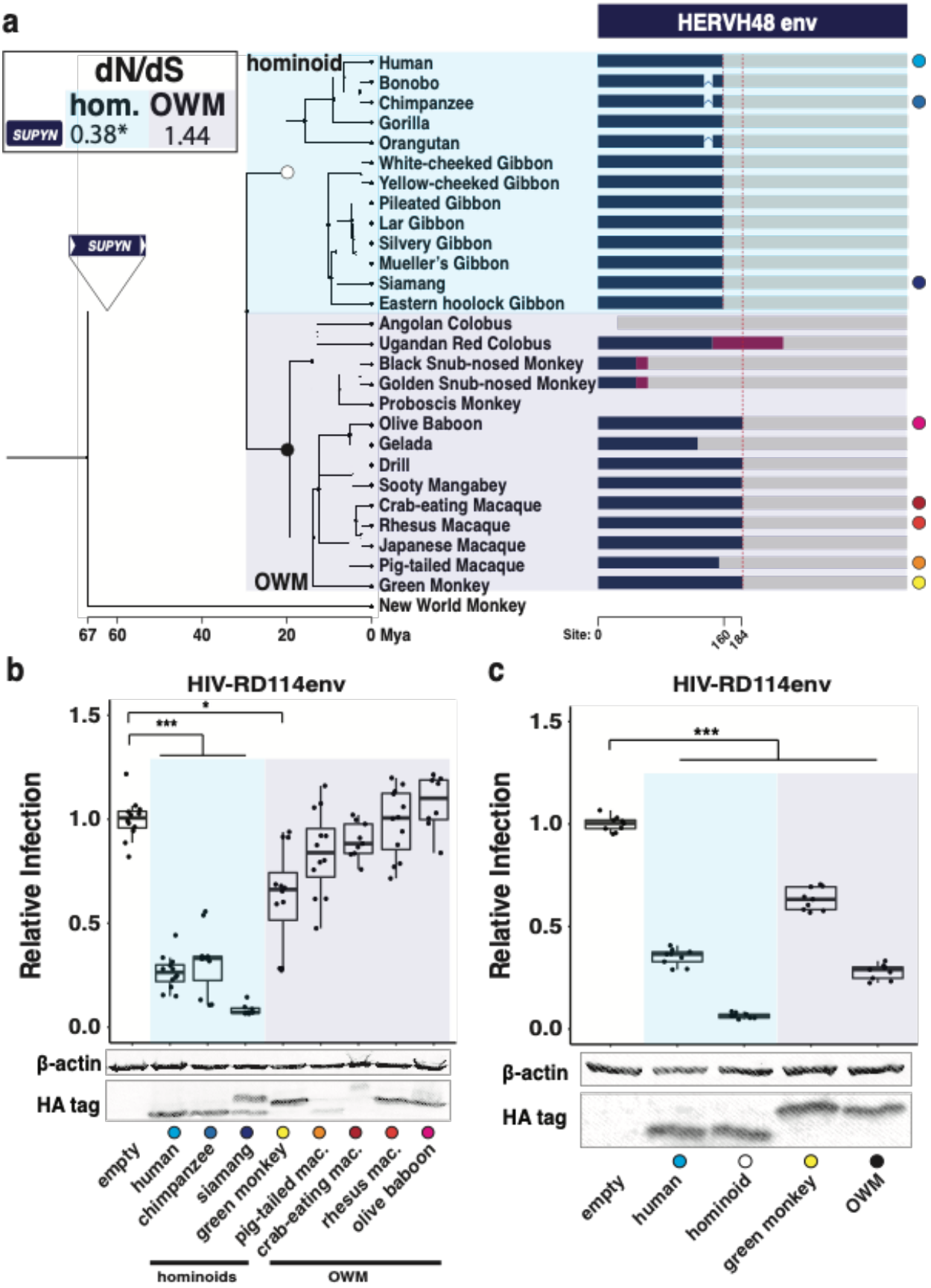
(a) HIV-glycoprotein reporter virus production and infection assay approach (see Methods). **(b)** Proportion of GFP+ 293T, JEG3, and Jar cells infected with indicated reporter viruses. **(c)** Relative infection rate of 293T and H1-ESCs was normalized to mean proportion of HIV-VSVg-infected

199 cells. **(d)** Proportion of GFP+ 293T and shSUPYN-Jar cells infected with indicated reporter virus.
200 **(e)** Relative infection rates of shSUPYN-Jar cells transfected with HA-tagged wild-type (WT-SP),
201 rescue signal peptide (Resc-SP), or luciferase signal peptide (GPluc-SP) encoding SUPYN
202 overexpression constructs. ($n \geq 3$ with ≥ 2 technical replicates; ***adj. $p < 0.001$; **adj. $p < 0.01$;
203 ANOVA with Tukey HSD). **(f)** Western Blot analysis (αHA, αGAPDH) of Jar-shSUPYN cell
204 lysates transfected with indicated SUPYN overexpression constructs. **(g,h)** Relative infection rates
205 of 293T cells transfected with unmodified SUPYN (g, h), RD114env (g), or SMRVenv
206 overexpression plasmids (h) and subsequently infected with indicated reporter viruses. **(i)** Model
207 of SUPYN-dependent RDR infection restriction. SUPYN likely binds ASCT2 following secretion
208 (i) or within the secretory compartment (ii).
209

210 Another prediction of RDR restriction via receptor interference is that it should be a property of
211 env proteins recognizing ASCT2, such as SUPYN, but not those binding other cellular receptors.
212 Consistent with this prediction and previous work with overexpressed SYN1 (*19*), expressing HA-
213 tagged SUPYN and SYN1 strongly restricted HIV-RD114, while HA-tagged env from
214 amphotrophic murine leukemia virus (aMLV) or human endogenous retrovirus H (HERVH),
215 neither of which are expected to interact with ASCT2 (*20–22*), had no effect on HIV-RD114 nor
216 HIV-VSVg infection in 293T cells **(fig S12a, b)**. Importantly, all tested env proteins were
217 expressed at comparable levels **(fig S12c).** Furthermore, we observed that SUPYN overexpression
218 did not significantly alter the level of ASCT2 protein expression in 293T cells **(Fig S12c).** This
219 result suggests that if SUPYN acts by receptor interference, its interaction with ASCT2 does not
220 result in receptor degradation, which is consistent with some instances of receptor interference
221 (*23–25*). In agreement with previous observations (*5*), we noted that *SUPYN* knockdown in Jar
222 cells resulted in the selective loss of a putative non-glycosylated isoform of ASCT2 **(fig S10b)**.
223 We speculate that the presence of SUPYN-dependent non-glycosylated ASCT2 may be the result
224 of SUPYN sterically interfering with the glycosylation machinery within the secretory pathway.
225 Collectively, these observations converge on the model that SUPYN restricts against RDR
226 infection through receptor interference by two possible mechanisms that are not mutually

227 exclusive: SUPYN may bind ASCT2 within the secretory pathway or extracellularly following
228 secretion **(Fig 3i)**.
229
230 To gain insights into the evolutionary emergence of *SUPYN* antiviral activity, we used comparative
231 genomics to investigate the origin and functional constraint acting on *SUPYN*. We found that the
232 HERVH48 provirus from which *SUPYN* is derived is present at orthologous position across the
233 genomes of all available hominoids and most Old World monkeys (OWM), but is absent in other
234 primate lineages **(Fig 4a, fig S13).** Thus, the provirus that gave rise to *SUPYN* inserted in the
235 common ancestor of catarrhine primates ~27-38 million years ago (*26*) **(Fig 4a)**. We next
236 examined whether *SUPYN* ORFs have evolved under functional constraint during primate
237 evolution. The *SUPYN* ORF is almost perfectly conserved in length across hominoids, but not in
238 OWM where some species display frameshifting and truncating mutations **(Fig 4a, fig S14, S15)**,
239 suggesting *SUPYN* evolved under different evolutionary regimes in hominoids and OWMs. To test
240 this idea, we analyzed the ratio (ω) of nonsynonymous (dN) to synonymous (dS) substitution rates
241 using codeml (*27*), which provides a measure of selective constraint acting on codons. Log-
242 likelihood ratio tests comparing models of neutral evolution with selection indicate *SUPYN*
243 evolved under purifying selection in hominoids ($\omega = 0.38$; p = 1.47E-02) but did not depart from
244 neutral evolution in OWMs ($\omega = 1.44$; p = 0.29) **(Fig 4a).** These results indicate that *SUPYN*
245 antiviral activity may be conserved across hominoids but not in OWM. To assess this, we
246 transfected 293T cells with HA-tagged overexpression constructs for the orthologous *SUPYN*
247 sequences of two hominoid species (chimpanzee, siamang), and five OWM species (African green
248 monkey, pigtailed macaque, crab-eating macaque, rhesus macaque, olive baboon) and challenged
249 these cells with HIV-RD114. Both chimpanzee and siamang SUPYN proteins displayed antiviral

250 activity with potency comparable to or greater than human SUPYN, respectively **(Fig 4b).** By
251 contrast, of the five OWM, only African green monkey SUPYN exhibited a modest but significant
252 level of antiviral activity **(Fig 4b, c).** The lack of restriction activity for some of the OWM proteins
253 may be attributed to their relatively low expression level in these human cells **(Fig 4b)** and/or their
254 inability to bind the human ASCT2 receptor due to *SUPYN* sequence divergence **(fig S14, S15)**.
255 To further trace the evolutionary origins of SUPYN antiviral activity, we expressed ancestral
256 *SUPYN* sequences phylogenetically reconstructed for the hominoid and OWM ancestors **(Fig S15)**
257 and assayed their antiviral activity in 293T cells. Both ancestral proteins were expressed at levels
258 comparable to human and green monkey SUPYN and exhibited strong restriction activity **(Fig 4c).**
259 These data indicate that SUPYN antiviral activity against RDRenv-mediated infection is an
260 ancestral trait preserved over ~20 million years of hominoid evolution, but apparently lost in
261 several OWM lineages.
262


a
HERVH48 env
dN/dS
hom. OWM
SUPYN 0.38* 1.44
SUPYN
hominoid
Human
Bonobo
Chimpanzee
Gorilla
Orangutan
White-cheeked Gibbon
Yellow-cheeked Gibbon
Pileated Gibbon
Lar Gibbon
Silvery Gibbon
Mueller's Gibbon
Siamang
Eastern hoolock Gibbon
Angolan Colobus
Ugandan Red Colobus
Black Snub-nosed Monkey
Golden Snub-nosed Monkey
Proboscis Monkey
Olive Baboon
Gelada
Drill
Sooty Mangabey
Crab-eating Macaque
Rhesus Macaque
Japanese Macaque
Pig-tailed Macaque
Green Monkey
New World Monkey
OWM
67 60 40 20 0 Mya
Site: 0 160 184
b
HIV-RD114env
Relative Infection
1.5
1.0
0.5
0.0
*
***
β-actin
HA tag
empty
human
chimpanzee
siamang
green monkey
pig-tailed mac.
crab-eating mac.
rhesus mac.
olive baboon
hominoids
OWM
c
HIV-RD114env
Relative Infection
1.5
1.0
0.5
0.0
***
β-actin
HA tag
empty
human
hominoid
green monkey
OWM

264 **Figure 4: SUPYN emerged in a Catarrhine ancestor and has conserved antiviral activity in**
265 **Hominoids.**
266 **(a)** Consensus primate phylogeny with cartoon representation of *SUPYN* ORFs (blue box).
267 Magenta boxes represent frameshifts in *SUPYN* ORFs. Red dashed lines denote conserved
268 premature stop codon positions. Grey bars represent degraded HERVH48env sequence. Labeled
269 triangles denote ancestral lineage where HERVH48env was acquired. Colored circles indicate
270 species used in **b** and **c.** *SUPYN* dN/dS value is shown in box (*$p < 0.05$; LRT). **(b, c)** Relative
271 infection rates and Western Blot of 293T cells transfected with primate (**b**) or ancestral (**c**)
272 SUPYN-HA constructs and subsequently infected with HIV-RD114env are shown. Relative
273 infection rates were determined by normalizing GFP+ counts to empty vector. ($n \geq 3$ with ≥ 2
274 technical replicates ***adj. $p < 0.001$; *adj. $p < 0.05$; ANOVA with Tukey HSD)
275

276 In this study, we show that the human genome harbors a large pool of novel and previously
277 annotated envORF sequences (*6*, *7*) that are dynamically expressed, particularly in the germline,
278 early embryonic development, and in response to immune stimulation and infection. Because
279 envORFs are predicted to encode proteins with the ability to bind receptors used by their ancestral
280 retroviruses and possibly many contemporary exogenous viruses, we propose that they constitute
281 a reservoir of potential antiviral factors. As a proof of principle, we established that one of these
282 proteins, *SUPYN,* is both necessary and sufficient to confer resistance to RDRenv-pseudotyped
283 reporter virus infection in human cell culture. We further determined that SUPYN expression in
284 the human preimplantation embryo and throughout placental development is driven by a tightly
285 regulated promoter derived from the LTR of its ancestral provirus. These results suggest that the
286 complex cis-regulatory activities encoded by retroviral LTRs facilitate ERV gene cooption (*1*, *2*)
287 The prominent expression of envORF, including SUPYN, in the developing germ line suggests
288 that they may have repeatedly acted as a barrier against the vertical propagation of the cognate
289 endogenous retroviruses by reinfection (*28*). This may explain why most HERV families appear
290 to have propagated in the genome primarily via retrotransposition rather than reinfection (*29*).
291 Under this model, we predict the adaptive benefits of envORFs to be generally transient, unless
292 they have broader antiviral activity or gain additional cellular function. This conjecture may

293 explain why *SUPYN* has been preserved by natural selection throughout hominoid evolution; not
294 only may *SUPYN* have shielded the early embryo and nascent germline from the persistent threat
295 of RDRs but also against the adverse effects of SYN1-mediated infections and syncytalization of
296 the developing placenta (*5*, *10*, *30*). In conclusion, this study provides evidence that truncated
297 envelope-derived peptides encoded in the human genome can evolve and maintain antiviral
298 activity. It is possible that our genome holds many other retrovirus-derived proteins with
299 protective effects against viral infection.
300

301 **Methods**

302 **RNA-seq analyses**

303 We mined published single cell transcriptome datasets (scRNA-seq) of human pre-implantation
304 embryos isolated at developmental stages ranging from oocyte to blastocyst (*31*), primordial germ
305 cells (*32*), human placenta (*33–35*), neuronal differentiation (*36*), haematopoietic stem cells (*37*),
306 pancreas (*38*), prefrontal cortex (*39*). Prior to analyzing the transcriptomic datasets, we first
307 recompiled the reference genome annotations using known gene (genecode V19) (*40*), and
308 *envORF* sequences to generate a reference transcriptome in fasta format. To guide the
309 transcriptome assembly, we converted this updated genome annotation to gtf format, which was
310 subsequently utilized for our expression quantifications. Reads were mapped to the curated human
311 genome (hg19) with STAR (*41*) using the following settings *--alignIntronMin 20 --*
312 *alignIntronMax 1000000 --chimSegmentMin 15 --chimJunctionOverhangMin 15 --*
313 *outFilterMultimapNmax 20*. Only uniquely mapped reads were considered for expression
314 calculations. Gene level counts were obtained using *featureCounts* (*42*) run with RefSeq
315 annotations. Gene expression levels were calculated at Transcript Per Million (TPM) from counts

316 mapped over the entire gene (defined as any transcript located between Transcription Start Site
317 (TSS) and Transcription End Site (TES)). Only genes and cells that met the following criteria were
318 included in this analysis: (1) each cell must express at least 5,000 genes; (2) each gene must be
319 expressed in at least 1% of cells; (3) each gene must be expressed with log2 TPM >1. We clustered
320 cells meeting these criteria using the default parameters of the Seurat (*43*) package (v3.1.1)
321 implemented in R (v3.6.0). Seurat applies the most variable genes to get top principal components
322 that are used to discriminate cell clusters in tSNE or UMAP plots. We set Seurat to use 10 principal
323 components in this cluster analysis. For the placental scRNAseq data **(Fig 2, Fig S5)**, the 2000
324 most differentially expressed genes were used to define cell clusters. Major clusters corresponding
325 to CTB, STB, EVTB, macrophages, and stromal cells were identified based on the expression of
326 known marker genes. Monocle2 (*44*) was used to perform single-cell trajectory analysis and cell
327 ordering along an artificial temporal continuum. The top 500 differentially expressed genes were
328 used to distinguish between CTB, STB and EVTB cell populations. The transcriptome from each
329 single cell represents a pseudo-time point along an artificial time vector that denotes the
330 progression of CTB to STB or EVTB respectively.
331 Data generated on the 10X Genomics scRNA-seq platforms were processed in the following way.
332 Normalized counts and cell-type annotations were used as provided by the original publications.
333 Seurat was used for filtering, normalization, and cell-type identification. The following data
334 processing steps were performed: (1) Cells were filtered based on the criteria that individual cells
335 must have between 1,000 and 5,000 expressed genes with a count ≥1; (2) cells with more than 5%
336 of counts mapping to mitochondrial genes were filtered out; (3) data was normalized by dividing
337 uniquely mapping read counts (defined by Seurat as unique molecular identified (UMI)) for each
338 gene by the total number of counts in each cell and multiplying by 10,000. These normalized

339 values were then natural-log transformed. Cell types were defined by using the top 2000 variable
340 features expressed across all samples. Clustering was performed using the "FindClusters" function
341 with largely default parameters; except resolution was set to 0.1 and the first 20 PCA dimensions
342 were used in the construction of the shared-nearest neighbor (SNN) graph and the generation of 2-
343 dimensional embeddings for data visualization using UMAP. Cell types were assigned based on
344 the annotations provided by the original publication.

345 Bulk RNAseq datasets generated from the GTEx consortium (*45*), placenta (*46*), 293T (*47*, *48*),
346 and immune (*49*, *50*) cells were processed as described above. Briefly, reads were mapped with
347 STAR and uniquely mapped reads were counted with featureCounts. We restricted the
348 visualization to only *envORFs* that were expressed (Log2 TPM > 1) in at least one of the analyzed
349 or compared samples (see the codes submitted on GitHub).

350

351 **ChIP-seq, DNAse-seq and ATAC-seq data analysis**

352 Various ChIP-seq datasets representing histone modifications and transcription factors in human
353 embryonic stem cells and their differentiation were retrieved from (*51*, *52*). We obtained the
354 H3K27Ac for CTB to STB primary culture (*53*), H3K4Me1 for trophoblasts (*54*), H3K4Me3,
355 H3K27Me3 for differentiated trophoblasts (*55*), and GATA2/3, TFAP2A/C (*55*) ChIP-seq datasets
356 in raw fastq format. DNAse-seq and ATAC-seq datasets were retrieved from references (*56*) and
357 (*57*) respectively.

358 Reads from the above-described datasets were aligned to the hg19 human reference genome using
359 Bowtie2 (*58*) run in *--very-sensitive-local* mode. All reads with MAPQ < 10 and PCR duplicates
360 were removed using Picard and *samtools* (*59*). Peaks were called by MACS2 (*60*)

361 (https://github.com/macs3-project/MACS) with the parameters in narrow mode for TFs and broad
362 mode for histone modifications keeping FDR < 1%. ENCODE-defined blacklisted regions (*61*)
363 were excluded from called peaks. We then intersected these peak sets with repeat elements from
364 hg19 repeat-masked coordinates using bedtools *intersectBed* (*62*) with a 50% overlap. To
365 visualize over Refseq genes (hg19) using IGV (*63*), the raw signals of ChIP-seq were obtained
366 from MACS2, using the parameters: *-g hs -q 0.01 -B*. The conservation track was visualized
367 through UCSC genome browser (*13*) under net/chain alignment of given non-human primates
368 (NHPs) and merged beneath the IGV tracks.
369
370 **Cell culture**
371 293T cells were cultured in DMEM (GIBCO, 11995065) containing 10% Fetal Bovine Serum
372 (FBS) (GIBCO, 10438026). Jar cells (provided by Carolyn Coyne) were cultured in RPMI
373 (GIBCO, 11875093) containing 10% FBS. JEG3 cells were cultured in MEM (GIBCO, 11095080)
374 containing 10% FBS. Culture medium for these cell lines was supplemented with sodium pyruvate
375 (GIBCO, 11360070), glutamax (GIBCO, 35050061), and Penicillin Streptomycin (GIBCO,
376 15140122) according to manufacturer specifications. H1-ESCs (*64*) (obtained from WiCell) were
377 grown on Matrigel (Corning, 356277) coated plates in MTESR+ (Stemcell, 05825) growth-media
378 and sub-cultured using Accutase (Innovative Cell Technologies, AT-104) and MTESR+
379 supplemented with CloneR (Stemcell, 05888). All cell lines were cultured at 37°C and 5% CO_2.
380
381 **Human embryos**
382 Prior to starting the project (documentation available upon request), all the procedures to obtain
383 human pre-implantation embryos were approved by local regulatory authorities and the Spanish

384 National Embryo steering committee. After the signing of informed consent documents by parents,
385 all human pre-implantation embryos were donated and fully anonymized. The embryos used in this
386 study were cryopreserved sibling embryos after successful In Vitro Fertilization pregnancies.
387
388 **Culturing of human pre-implantation embryos.**
389 Human pre-implantation embryos at several stages of development (from +2 to +6 days post-
390 fertilization) were thawed and allowed to grow. Only embryos reaching the blastocyst stage and
391 showing viability in most individual cells were selected for further analyses. Cryopreserved human
392 embryos were thawed using thawing media from Vitrolife (G-1 v.5 and G-2 v.5 media), and were
393 classified based on morphological criteria using inverted Hoffman optical microscopy. More than
394 50% of thawed embryos showed signs of cell division.
395
396 **Vector cloning**
397 DHIV3-GFP, phCMV-RD114env, psi(-)-amphoMLV plasmids were provided by Vicente
398 Planelles (University of Utah). pCGCG-SMRVenv plasmid was provided by Welkin Johnson
399 (Boston University). psPAX2 and pVSVg plasmids were provided by John Lis (Cornell
400 University). The following cloning approaches were performed using primers and constructs
401 described in Supplementary Table 3. HERVH1env ORF was PCR-amplified using Q5 polymerase
402 (NEB, M0491L) from HeLa and 293T genomic DNA respectively and cloned into a TOPO vector
403 (ThermoFisher, 450245).
404 To generate stable *SYN1* and *SUPYN* knock-down cell lines, pHIV lentiviral constructs containing
405 shRNAs targeting *SYN1* and *SUPYN* respectively were cloned using the following strategy. The
406 shRNA encoded in pHIV7-U6-shW3, generously provided by Lars Aagaard (Aarhus University),

407 targets *SYN1* (*65*). *SUPYN*-targeting shRNAs were designed using siRNA sequences employed by
408 Jun Sugimoto[4] as a template. pHIV7 lentiviral constructs were cloned using the pHIV7-U6-shW3
409 plasmid (*65*) as a template. pHIV7-U6-shSup-cer, pHIV7-U6-shSup-puro, pHIV7-U6-shC-cer,
410 pHIV7-U6-shC-puro, pHIV7-U6-shSyn1-cer, pHIV7-U6-shSyn1-puro were generated using a
411 Gibson assembly approach. To replace the native GFP marker of pHIV7-U6-shW3 with a Cerulean
412 reporter or puromycin resistance marker, we digested pHIV7-U6-shW3 with NheI (NEB, R3131S)
413 and KpnI (NEB, R3142S). This digest resulted in the production of three DNA fragments: pHIV7
414 backbone, GFP-, and WPRE-containing fragments. We separately PCR amplified each selection
415 marker and WPRE containing pHIV7 fragment. InFusion cloning was then used to ligate the
416 digested pHIV7 backbone to the Cerulean or puromycin cassette and WPRE containing PCR
417 product. shRNAs were cloned into the pHIV7-Cerulean/puromycin transfer construct previously
418 digested with NotI (NEB, R0189S) and NheI. U6-promoter containing shRNA cassettes and the
419 CMV promoter driving marker cassette expression were PCR amplified and subsequently InFusion
420 cloned into the NotI/NheI digested pHIV7-cerulean/puromycin backbone.
421 All pHCMVenv and SUPYN expression constructs, described in this study, were generated as
422 follows: HA-tagged and untagged ORFs with pHCMV homologous overhanging sequence were
423 either PCR amplified using Q5 polymerase (NEB, M0491S) or synthesized (IDT) (see **Table S2**),
424 and cloned into EcoRI (NEB, R3101T) digested pHCMV backbones using the InFusion cloning
425 kit (Takara Bio, 638920). To generate siRNA-resistant SUPYN rescue constructs, we replaced the
426 native signal peptide sequence[4] (which is targeted by siRNAs used in this study) with (1) a *Gaussia*
427 *princeps* luciferase SP (SUPYN-lucSP) (*66*, *67*) and (2) a shSUPYN resistant SUPYN rescue
428 construct (SUPYN-rescSP) in which the codons were modified to retain the codon identity but
429 disrupt siRNA binding.

430

431 **Antibodies**

432 All antibodies used in this study are commercially available. α-GAPDH (D4C6R, D16H11), α-
433 βactin (D6A8), α-HA (C29F4), α-ASCT2 (V501) primary antibodies were purchased from Cell
434 Signaling Technology. α-SUPYN and αOCT4 primary antibodies were purchased from Phoenix
435 Pharma (H-059-052) and Santa Cruz Biotechnology respectively. α-Mouse (#7076) and α-Rabbit
436 (#7074) HRP conjugated secondary antibodies were purchased from Cell Signaling Technology.
437 IRDye secondary antibodies were purchased from Licor (925-32211, 925-68072, 925-32210, 925-
438 68073). Alexa-fluor conjugated secondary antibody was purchased from Invitrogen.

439

440 **Western Blot**

441 Whole cell extracts from cultured cell lines were prepared using 1x GLO lysis buffer (Promega,
442 E266A). One third volume of 4x Laemli buffer was added to one volume whole cell extract
443 samples, then incubated at 95°C for 5 minutes, and sonicated for 15 minutes at 4°C (amplitude
444 100; pulse interval 15 seconds on, 15 seconds off). Approximately 30 ug of protein were separated
445 by SDS-PAGE (BioRad, 1610175), transferred to PVDF membrane (BioRad, 1620177), blocked
446 according to antibody manufacturers specification, and incubated overnight in appropriate primary
447 antibody then incubated in IRDye or peroxidase conjugated goat anti-mouse or anti-rabbit
448 secondary antibodies for 1 hour at room temperature. Protein was then detected using ECL reagent
449 (BioRad, 1705061) or the Licor Odyssey imaging system.

450

451 **IF microscopy**

452 *For placenta:* Human second trimester placental tissue that resulted from elective terminations
453 was obtained from the University of Pittsburgh Health Sciences Tissue Bank through an honest
454 broker system after approval from the University of Pittsburgh Institutional Review Board (IRB)
455 and in accordance with the University of Pittsburgh's tissue procurement guidelines. Tissue was
456 excluded in cases of fetal anomalies or aneuploidy. Third trimester placental tissue was obtained
457 through the Magee Obstetric Maternal & Infant (MOMI) Database and Biobank after approval
458 from the University of Pittsburgh IRB. Women who had previously consented for tissue donation
459 and underwent cesarean delivery were included. Placental tissues were fixed in 4% PFA (in 1x
460 PBS) for 30 minutes, permeabilized with 0.25% Triton X-100 for 30 minutes (on a rocker), washed
461 with 1x PBS and then incubated with primary anti-Suppressyn antibody at 1:200 in 1x PBS for 2-
462 4 hours at room temperature. These samples were incubated with Alexa-fluor conjugated
463 secondary antibody (Invitrogen) diluted 1:1000 and counterstained with actin. DAPI was included
464 in our PBS and then mounted in Vectashield mounting medium with DAPI (Vector Laboratories,
465 H-1200).
466 *For embryo:* Pre-implantation embryos at the blastocyst stage were washed with 1x PBS
467 (Invitrogen) and fixed for 15 minutes using fresh 4% (w/v) paraformaldehyde (Sigma) at room
468 temperature. After fixation, embryos were washed twice on a petri dish using 1x PBS containing
469 0.1% (v/v) Triton X-100 (Sigma). Next, embryos were permeabilized on a lo-bind 2ml Eppendorf
470 tube using 1x PBS containing 0.5% (v/v) Triton X-100 for 24 hours at 4ºC. After permeabilization,
471 embryos were incubated for 24 hours at 4ºC in blocking solution [1x PBS containing 0.1% (v/v)
472 Triton X-100, 1% (v/v) Tissue Culture Grade DMSO (Sigma), and 5% (v/v) Normal Goat Serum
473 (Abcam)]. Next, embryos were incubated with the primary antibodies [1:200 anti-Rabbit
474 Suppressyn (Phoenix Pharmaceuticals, Inc), and 1:200 anti-Goat OCT4 (Santa Cruz

475 Biotechnology, Inc)] in fresh blocking solution for 24 hours at 4ºC. After 30 minutes at room
476 temperature, embryos were washed twice for 5 minutes using 1x PBS containing 0.1% (v/v) Triton
477 X-100, and subsequently incubated with secondary antibodies in blocking solution containing
478 DAPI (Thermo Scientific) [1:1000 anti-Rabbit Alexa 488 (Life Technologies), and 1:1000 anti-
479 Goat Alexa 555 (Life Technologies)] for 24 hours at 4ºC in the dark. Next, embryos were washed
480 twice for 5 minutes using 1x PBS containing 0.1% (v/v) Triton X-100 and mounted in a µ-Slide 8
481 well (Ibidi®) using 1xPBS and avoiding the generation of air bubbles.
482 A confocal Zeiss LSM 710 device (Genyo, Spain) was used to process the slides, using a 63X
483 objective, zoom 0.6x, with a 1024x1024 pixels resolution (corresponding to 224.70 µm x 224.70
484 µm). Fluorescence channels were recorded independently to avoid artifacts, and microscope
485 images shown are maximum projections of several images (2-3 or more). As negative controls,
486 embryos were also incubated with only the secondary antibody.
487
488 **Virus production**
489 Low passage 293T cells were used to produce lentiviral particles. DHIV3-GFP and env-
490 expression plasmids were co-transfected at a mass ratio of 2:1 using lipofectamine 2000
491 (ThermoFisher, 11668030). shRNA encoding lentiviral particles were produced by co-transfecting
492 pHIV7, psPAX2, pVSVg according to BROAD institute lentiviral production protocol
493 (https://portals.broadinstitute.org/gpp/public/resources/protocols) using Lipofectamine 2000.
494 Growth media was replaced on transfected cells after overnight incubation. At 72 hours post-
495 transfection, virus containing supernatant was harvested, centrifuged to remove cell debris, filtered
496 through a 0.45 um pore filter, and stored at -80°C.
497

498 **Infection Assays**

499 293T cells were transfected with env-overexpression constructs using Lipofectamine 2000 and
500 incubated 24 hours. Transfected cells were infected with reporter virus by applying virus (HIV-
501 RD114env, HIV-VSVg, HIV-SMRVenv) stocks in the presence of polybrene (Santa Cruz Bio, sc-
502 134220) at a final concentration of 4 ug/mL. After 6-8 hours, virus stock was replaced with fresh
503 growth media. Infected cells were maintained for 72 hours, replacing media when necessary, and
504 harvested with trypsin. Detached cells were suspended in fresh growth media, strained and
505 analyzed by flow cytometry. For the H1-ESC infection experiment, relative infection rates were
506 calculated by normalizing the percent GFP+, HIV-RD114env infected cells to the percent GFP+
507 HIV-VSVg infected cells. For env/SUPYN overexpression experiments, relative infection rates
508 were calculated by normalizing the percent GFP+ env/SUPYN-transfected cells to the percent
509 GFP+ empty vector transfected cells. ANOVA with Tukey HSD tests were implemented in R
510 (v3.6.3).

511

512 **Placental cell shRNA transduction**

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

518

519 **RT-qPCR**

520 RNA was isolated from cultured cells using the RNeasy Mini Kit (Qiagen, 74104) and an on
521 column dsDNAse digestion (Qiagen, 79254) was performed. 1-3 ug of total RNA were used to
522 generate cDNA with the maxima cDNA synthesis with dsDNAse kit (ThermoFisher, K1681).
523 qPCR reactions were performed using the LC480 Instrument with Sybr Green PCR master mix
524 (Roche, 04707516001) according to manufacturer's protocol and using primers indicated in
525 Supplementary Table 2. Gene expression was then quantified using the ΔΔCT method (*68*). 18S
526 expression was used as a reference housekeeping gene. Wilcox rank sum tests were performed
527 using R (v3.6.3).
528
529 **Envelope evolutionary sequence analyses**
530 Orthologous *SUPYN*, *SYN1*, and *SYN2* sequences were extracted from the 30-species MULTIZ
531 alignment (*13*) and formatted for sequence alignment using the phast package (*69*). These and
532 additional syntenic *SUPYN* and *SYN2* open reading frame sequences were validated/identified by
533 BLASTn (*70*) search with default settings of publicly available Catarrhine primate genomes
534 (ncbi.nih.gov). Mariam Okhovat of the Carbone Lab (Oregon Health and Science University)
535 generously provided BAM files containing read alignment information for *SUPYN*, *SYN1*, and
536 *SYN2* generated from whole genome sequencing of *Hoolock leuconedys* (Hoolock Gibbon),
537 *Symphalangus syndactylus* (Siamang), *Hylobates muelleri* (Müller's Gibbon), *Hylobates lar* (Lar
538 Gibbon), *Hylobates moloch* (Silvery Gibbon), *Hylobates pileatus* (Pileated Gibbon), and
539 *Nomascus gabriellae* (Yellow-cheeked Gibbon) (*71*). Where multiple individuals were sequenced,
540 a consensus sequence was generated using samtools (*59*) and JalView (*72*).
541 To perform *dN/dS* analyses, orthologous env sequences (>90bp length) encoding the mature
542 sequence downstream of the signal peptide cleavage site, were aligned using MEGA7 (*73*) and

543 manually converted to PHYLIP format. A Newick tree was generated based on this alignment
544 using the maximum likelihood algorithm implemented in MEGA7. *Codeml*, implemented in the
545 PAML package, was run to calculate *dN/dS* values and log likelihood (LnL) scores generated under
546 models M0, M1, M2, M7 and M8 (*27*). Chi-square tests comparing LnL scores generated under
547 models of neutral evolution and selection were performed.
548 We used two approaches to reconstruct ancestral hominoid and OWM *SUPYN* sequences. First,
549 we reconstructed ancestral *SUPYN* sequences using the majority rule consensus sequence
550 (calculated in JalView) of the hominoid and OWM clade respectively. At positions where
551 nucleotide identity was ambiguous, the dominant nucleotide identity in the neighboring clade was
552 used as a tiebreaker. These sequences were used for our infection assays shown in **Fig 4h.** We
553 also employed a maximum likelihood approach using the *baseml* program, implemented in PAML
554 (*27*). We reconstructed ancestral *SUPYN* sequences using the hominoid species, shown in **Fig 1**,
555 and the 6 OWM monkeys with the most complete *SUPYN*-coding open reading frame (olive
556 baboon, drill, crab-eating macaque, rhesus macaque, japanese macaque, green monkey) as our
557 input sequences. Because PAML requires a Newick tree as an input, the MEGAX maximum
558 likelihood algorithm was used to generate a Newick tree with the above described *SUPYN*
559 sequences (*74–76*). *Baseml* was run using models 3-7 (F84, HKY85, T92, TN93, REV). As shown
560 in **fig S16**, both the consensus- and maximum likelihood reconstructions were identical for the
561 OWM *SUPYN* sequences. The consensus-based hominoid sequence reconstruction differed from
562 our maximum likelihood-based reconstruction by two amino acids. These two positions are
563 unlikely to affect the function of the resulting protein because these sites are identical to siamang
564 *SUPYN*, which restricts RD114env-mediated infection **(Fig 4a, b)**.
565

566 **Code availability**

567 Code used to process and plot the high throughput datasets are available at
568 https://github.com/Manu-1512/Envelopes_Suppressyn

569

570 **Acknowledgements**

571 This work was funded by Cornell University and by the National Institutes of Health to CF (R01
572 GM112972, R35 GM122550). J.L.G.P´s lab acknowledge support from CICE-FEDER-P12-CTS-
573 2256, The Wellcome Trust-University of Edinburgh Institutional Strategic Support Fund (ISFF2),
574 the European Research Council (ERC-Consolidator ERC-STG-2012-309433), an International
575 Early Career Scientist grant from the Howard Hughes Medical Institute (IECS-55007420), and
576 from a private donation to the lab from Ms Francisca Serrano (*Trading y Bolsa para Torpes*,
577 Granada, Spain). We thank Dr. Vicente Planelles for sharing reporter virus plasmids and training;
578 Dr. Welkin Johnson for sharing his SMRVenv expression plasmid; Dr. Lars Aagaard for sharing
579 shRNA transduction constructs; Dr. John Lis for providing lentiviral packaging construct; Dr.
580 Amnon Koren and Rita Rebello for providing embryonic stem cell cultures and technical support
581 respectively**;** Dr Lucia Carbone and Dr. Mariam Okhovat for sharing gibbon genome sequences.
582 We thank Dr. Ray Malfavon-Borja for producing an initial list of envelope open reading frames in
583 the human genome. We thank Maia Clare for her contribution to evolutionary sequence analyses.
584 We thank Dr. Akiko Iwasaki for her critical reading of this manuscript. We thank members of the
585 Feschotte lab for helpful advice and discussion throughout the project.

586

587 **Author contributions**

588 CF conceived of this project. JAF and CF designed and developed this project. JAF designed and
589 conducted all experiments, evolutionary sequence analyses, and analyzed all experimental data.
590 MS performed all gene expression and regulation analyses. HBC and RAK helped perform
591 infection assays and evolutionary sequence analyses. CBC performed placental stains. JLC and
592 JLGP obtained, cultured, and prepared human embryos for microscopy. MBB performed embryo
593 stains and imaging. JAF, CF and MS wrote this manuscript.
594

595 **Competing interest declaration**

596 The authors declare no competing interests.
597

598 **References**

599

600 1. W. E. Johnson, Origins and evolutionary consequences of ancient endogenous retroviruses.
601 *Nature reviews. Microbiology*. **72**, 5955 (2019).

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

604 3. H. W. Virgin, The Virome in Mammalian Physiology and Disease. *Cell*. **157**, 142–150 (2014).

605 4. R. Malfavon-Borja, C. Feschotte, Fighting Fire with Fire: Endogenous Retrovirus Envelopes
606 as Restriction Factors. *Journal of Virology*. **89**, 4047–4050 (2015).

607 5. J. Sugimoto, D. J. Schust, T. Kinjo, Y. Aoki, Y. Jinno, Y. Kudo, Suppressyn localization and
608 dynamic expression patterns in primary human tissues support a physiologic role in human
609 placentation. *Scientific Reports*. **9**, 19502–12 (2019).

610 6. C. Lavialle, G. Cornelis, A. Dupressoir, C. Esnault, O. Heidmann, C. Vernochet, T.
611 Heidmann, Paleovirology of "syncytins", retroviral env genes exapted for a role in placentation.
612 *Philosophical Transactions of the Royal Society B: Biological Sciences*. **368**, 20120507–
613 20120507 (2013).

614 7. S. Nakagawa, M. U. Takahashi, gEVE: a genome-based endogenous viral element database
615 provides comprehensive viral protein-coding sequences in mammalian genomes. *Database J*
616 *Biological Databases Curation*. **2016**, baw087 (2016).

617 8. A. C. van der Kuyl, HIV infection and HERV expression: a review. *Retrovirology*. **9**, 6
618 (2012).

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

621 10. J. Sugimoto, M. Sugimoto, H. Bernstein, Y. Jinno, D. Schust, A novel human endogenous
622 retroviral protein inhibits cell-cell fusion. *Scientific Reports*. **3**, 1462–8 (2013).

623 11. A. Sinha, W. E. Johnson, ScienceDirect Retroviruses of the RDR superinfection interference
624 group: ancient origins and broad host distribution of a promiscuous Env gene. *Current Opinion*
625 *in Virology*. **25**, 105–112 (2017).

626 12. M. Scalise, L. Pochini, L. Console, M. A. Losso, C. Indiveri, The Human SLC1A5 (ASCT2)
627 Amino Acid Transporter: From Function to Structure and Role in Cell Biology. *Frontiers in cell*
628 *and developmental biology*. **6**, 96 (2018).

629 13. M. Haeussler, A. S. Zweig, C. Tyner, M. L. Speir, K. R. Rosenbloom, B. J. Raney, C. M.
630 Lee, B. T. Lee, A. S. Hinrichs, J. N. Gonzalez, D. Gibson, M. Diekhans, H. Clawson, J. Casper,
631 G. P. Barber, D. Haussler, R. M. Kuhn, W. J. Kent, The UCSC Genome Browser database: 2019
632 update. *Nucleic acids research*. **47**, D853–D858 (2019).

633 14. D. Finkelshtein, A. Werman, D. Novick, S. Barak, M. Rubinstein, LDL receptor and its
634 family members serve as the cellular receptors for vesicular stomatitis virus. *Proceedings of the*
635 *National Academy of Sciences of the United States of America*. **110**, 7306–7311 (2013).

636 15. M. Dottori, C. Tay, S. M. Hughes, Neural development in human embryonic stem cells-
637 applications of lentiviral vectors. *Journal of cellular biochemistry*. **112**, 1955–1962 (2011).

638 16. M. Sakata, H. Tani, M. Anraku, M. Kataoka, N. Nagata, F. Seki, M. Tahara, N. Otsuki, K.
639 Okamoto, M. Takeda, Y. Mori, Analysis of VSV pseudotype virus infection mediated by rubella
640 virus envelope proteins. *Scientific Reports*. **7**, 11607 (2017).

641 17. W. E. Johnson, Endogenous Retroviruses in the Genomics Era. *Annual review of virology*. **2**,
642 135–159 (2015).

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

645 19. V. G. Ponferrada, B. S. Mauck, D. P. Wooley, The envelope glycoprotein of human
646 endogenous retrovirus HERV-W induces cellular resistance to spleen necrosis virus. *Arch Virol*.
647 **148**, 659–675 (2003).

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

651 21. M. van Zeijl, S. V. Johann, E. Closs, J. Cunningham, R. Eddy, T. B. Shows, B. O'Hara, A
652 human amphotropic retrovirus receptor is a second member of the gibbon ape leukemia virus
653 receptor family. *Proceedings of the National Academy of Sciences*. **91**, 1168–1172 (1994).

654 22. D. G. Miller, R. H. Edwards, A. D. Miller, Cloning of the cellular receptor for amphotropic
655 murine retroviruses reveals homology to that for gibbon ape leukemia virus. *Proceedings of the*
656 *National Academy of Sciences*. **91**, 78–82 (1994).

657 23. M. Liu, M. V. Eiden, The receptors for gibbon ape leukemia virus and amphotropic murine
658 leukemia virus are not downregulated in productively infected cells. *Retrovirology*. **8**, 53–14
659 (2011).

660 24. Z. Jobbagy, S. Garfield, L. Baptiste, M. V. Eiden, W. B. Anderson, Subcellular redistribution
661 of Pit-2 P(i) transporter/amphotropic leukemia virus (A-MuLV) receptor in A-MuLV-infected
662 NIH 3T3 fibroblasts: involvement in superinfection interference. *Journal of Virology*. **74**, 2847–
663 2854 (2000).

664 25. J. W. Kim, J. M. Cunningham, N-linked glycosylation of the receptor for murine ecotropic
665 retroviruses is altered in virus-infected cells. *Journal of Biological Chemistry*. **268**, 16316–16320
666 (1993).

667 26. P. Perelman, W. E. Johnson, C. Roos, H. N. Seuánez, J. E. Horvath, M. A. M. Moreira, B.
668 Kessing, J. Pontius, M. Roelke, Y. Rumpler, M. P. C. Schneider, A. Silva, S. J. O'Brien, J.
669 Pecon-Slattery, A molecular phylogeny of living primates. *PLoS Genetics*. **7**, e1001342 (2011).

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

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

674 29. G. Magiorkinis, R. J. Gifford, A. Katzourakis, J. D. Ranter, R. Belshaw, Env-less
675 endogenous retroviruses are genomic superspreaders. *Proceedings of the National Academy of*
676 *Sciences of the United States of America*. **109**, 7385–7390 (2012).

677 30. Y. Tang, B. O. Woodward, L. Pastor, A. M. George, O. Petrechko, F. J. Nouvet, D. W. Haas,
678 G. Jiang, J. E. K. Hildreth, Endogenous Retroviral Envelope Syncytin Induces HIV-1 Spreading
679 and Establishes HIV Reservoirs in Placenta. *Cell reports*. **30**, 4528-4539.e4 (2020).

680 31. L. Yan, M. Yang, H. Guo, L. Yang, J. Wu, R. Li, P. Liu, Y. Lian, X. Zheng, J. Yan, J.
681 Huang, M. Li, X. Wu, L. Wen, K. Lao, R. Li, J. Qiao, F. Tang, Single-cell RNA-Seq profiling of

682 human preimplantation embryos and embryonic stem cells. *Nature structural & molecular*
683 *biology*. **20**, 1131–1139 (2013).

684 32. F. Guo, L. Yan, H. Guo, L. Li, B. Hu, Y. Zhao, J. Yong, Y. Hu, X. Wang, Y. Wei, W. Wang,
685 R. Li, J. Yan, X. Zhi, Y. Zhang, H. Jin, W. Zhang, Y. Hou, P. Zhu, J. Li, L. Zhang, S. Liu, Y.
686 Ren, X. Zhu, L. Wen, Y. Q. Gao, F. Tang, J. Qiao, The Transcriptome and DNA Methylome
687 Landscapes of Human Primordial Germ Cells. *Cell*. **161**, 1437–1452 (2015).

688 33. Y. Liu, X. Fan, R. Wang, X. Lu, Y.-L. Dang, H. Wang, H.-Y. Lin, C. Zhu, H. Ge, J. C.
689 Cross, H. Wang, Single-cell RNA-seq reveals the diversity of trophoblast subtypes and patterns
690 of differentiation in the human placenta. *Cell research*. **28**, 819–832 (2018).

691 34. M. Pavličev, G. P. Wagner, A. R. Chavan, K. Owens, J. Maziarz, C. Dunn-Fletcher, S. G.
692 Kallapur, L. Muglia, H. Jones, Single-cell transcriptomics of the human placenta: inferring the
693 cell communication network of the maternal-fetal interface. *Genome Research*. **27**, 349–361
694 (2017).

695 35. R. Vento-Tormo, M. Efremova, R. A. Botting, M. Y. Turco, M. Vento-Tormo, K. B. Meyer,
696 J.-E. Park, E. Stephenson, K. Polański, A. Goncalves, L. Gardner, S. Holmqvist, J. Henriksson,
697 A. Zou, A. M. Sharkey, B. Millar, B. Innes, L. Wood, A. Wilbrey-Clark, R. P. Payne, M. A.
698 Ivarsson, S. Lisgo, A. Filby, D. H. Rowitch, J. N. Bulmer, G. J. Wright, M. J. T. Stubbington, M.
699 Haniffa, A. Moffett, S. A. Teichmann, Single-cell reconstruction of the early maternal-fetal
700 interface in humans. *Nature*. **563**, 347–353 (2018).

701 36. J. L. Close, Z. Yao, B. P. Levi, J. A. Miller, T. E. Bakken, V. Menon, J. T. Ting, A. Wall, A.-
702 R. Krostag, E. R. Thomsen, A. M. Nelson, J. K. Mich, R. D. Hodge, S. I. Shehata, I. A. Glass, S.
703 Bort, N. V. Shapovalova, N. K. Ngo, J. S. Grimley, J. W. Phillips, C. L. Thompson, S.
704 Ramanathan, E. Lein, Single-Cell Profiling of an In Vitro Model of Human Interneuron
705 Development Reveals Temporal Dynamics of Cell Type Production and Maturation. *Neuron*. **93**,
706 1035-1048.e5 (2017).

707 37. L. Velten, S. F. Haas, S. Raffel, S. Blaszkiewicz, S. Islam, B. P. Hennig, C. Hirche, C. Lutz,
708 E. C. Buss, D. Nowak, T. Boch, W.-K. Hofmann, A. D. Ho, W. Huber, A. Trumpp, M. A. G.
709 Essers, L. M. Steinmetz, Human haematopoietic stem cell lineage commitment is a continuous
710 process. *Nature cell biology*. **19**, 271–281 (2017).

711 38. M. Enge, H. E. Arda, M. Mignardi, J. Beausang, R. Bottino, S. K. Kim, S. R. Quake, Single-
712 Cell Analysis of Human Pancreas Reveals Transcriptional Signatures of Aging and Somatic
713 Mutation Patterns. *Cell*. **171**, 321-330.e14 (2017).

714 39. S. Darmanis, S. A. Sloan, Y. Zhang, M. Enge, C. Caneda, L. M. Shuer, M. G. H. Gephart, B.
715 A. Barres, S. R. Quake, A survey of human brain transcriptome diversity at the single cell level.
716 *Proceedings of the National Academy of Sciences of the United States of America*. **112**, 7285–
717 7290 (2015).

718 40. A. Frankish, M. Diekhans, A.-M. Ferreira, R. Johnson, I. Jungreis, J. Loveland, J. M. Mudge,
719 C. Sisu, J. Wright, J. Armstrong, I. Barnes, A. Berry, A. Bignell, S. Carbonell Sala, J. Chrast, F.
720 Cunningham, T. Di Domenico, S. Donaldson, I. T. Fiddes, C. García Girón, J. M. Gonzalez, T.
721 Grego, M. Hardy, T. Hourlier, T. Hunt, O. G. Izuogu, J. Lagarde, F. J. Martin, L. Martínez, S.
722 Mohanan, P. Muir, F. C. P. Navarro, A. Parker, B. Pei, F. Pozo, M. Ruffier, B. M. Schmitt, E.
723 Stapleton, M.-M. Suner, I. Sycheva, B. Uszczynska-Ratajczak, J. Xu, A. Yates, D. Zerbino, Y.
724 Zhang, B. Aken, J. S. Choudhary, M. Gerstein, R. Guigó, T. J. P. Hubbard, M. Kellis, B. Paten,
725 A. Reymond, M. L. Tress, P. Flicek, GENCODE reference annotation for the human and mouse
726 genomes. *Nucleic Acids Res*. **47**, gky955- (2018).

727 41. A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson,
728 T. R. Gingeras, STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)*.
729 **29**, 15–21 (2013).

730 42. Y. Liao, G. K. Smyth, W. Shi, featureCounts: an efficient general purpose program for
731 assigning sequence reads to genomic features. *Bioinformatics (Oxford, England)*. **30**, 923–930
732 (2014).

733 43. T. Stuart, A. Butler, P. Hoffman, C. Hafemeister, E. Papalexi, W. M. Mauck, Y. Hao, M.
734 Stoeckius, P. Smibert, R. Satija, Comprehensive Integration of Single-Cell Data. *Cell*. **177**, 1888-
735 1902.e21 (2019).

736 44. X. Qiu, A. Hill, J. Packer, D. Lin, Y.-A. Ma, C. Trapnell, Single-cell mRNA quantification
737 and differential analysis with Census. *Nature methods*. **14**, 309–315 (2017).

738 45. Gte. Consortium, Genetic effects on gene expression across human tissues. *Nature*. **550**,
739 204–213 (2017).

740 46. J. Zadora, M. Singh, F. Herse, L. Przybyl, N. Haase, M. Golic, H. W. Yung, B. Huppertz, J.
741 E. Cartwright, G. Whitley, G. M. Johnsen, G. Levi, A. Isbruch, H. Schulz, F. C. Luft, D. N.
742 Müller, A. C. Staff, L. D. Hurst, R. Dechend, Z. Izsvák, Disturbed Placental Imprinting in
743 Preeclampsia Leads to Altered Expression of DLX5, a Human-Specific Early Trophoblast
744 Marker. *Circulation*. **136**, 1824–1839 (2017).

745 47. M. D. Antonio, D. Weghorn, A. D. Antonio-Chronowska, F. Coulet, K. M. Olson, C.
746 DeBoever, F. Drees, A. Arias, H. Alakus, A. L. Richardson, R. B. Schwab, E. K. Farley, S. R.
747 Sunyaev, K. A. Frazer, Identifying DNase I hypersensitive sites as driver distal regulatory
748 elements in breast cancer. *Nature communications*. **8**, 436 (2017).

749 48. H. Chung, J. J. A. Calis, X. Wu, T. Sun, Y. Yu, S. L. Sarbanes, V. L. D. Thi, A. R. Shilvock,
750 H.-H. Hoffmann, B. R. Rosenberg, C. M. Rice, Human ADAR1 Prevents Endogenous RNA
751 from Triggering Translational Shutdown. *Cell*. **172**, 811-824.e14 (2018).

752 49. D. Calderon, M. L. T. Nguyen, A. Mezger, A. Kathiria, F. Müller, V. Nguyen, N. Lescano,
753 B. Wu, J. Trombetta, J. V. Ribado, D. A. Knowles, Z. Gao, F. Blaeschke, A. V. Parent, T. D.
754 Burt, M. S. Anderson, L. A. Criswell, W. J. Greenleaf, A. Marson, J. K. Pritchard, Landscape of

755 stimulation-responsive chromatin across diverse human immune cells. *Nature Publishing Group*.
756 **51**, 1494–1505 (2019).

757 50. I. L. Shytaj, B. Lucic, M. Forcato, C. Penzo, J. Billingsley, V. Laketa, S. Bosinger, M.
758 Stanic, F. Gregoretti, L. Antonelli, G. Oliva, C. K. Frese, A. Trifunovic, B. Galy, C. Eibl, G.
759 Silvestri, S. Bicciato, A. Savarino, M. Lusic, Alterations of redox and iron metabolism
760 accompany the development of HIV latency. *Embo J*. **39**, e102209 (2020).

761 51. A. M. Tsankov, H. Gu, V. Akopian, M. J. Ziller, J. Donaghey, I. Amit, A. Gnirke, A.
762 Meissner, Transcription factor binding dynamics during human ES cell differentiation. *Nature*.
763 **518**, 344–349 (2015).

764 52. T. S. Barakat, F. Halbritter, M. Zhang, A. F. Rendeiro, E. Perenthaler, C. Bock, I. Chambers,
765 Functional Dissection of the Enhancer Repertoire in Human Embryonic Stem Cells. *Cell stem*
766 *cell*. **23**, 276-288.e8 (2018).

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

770 54. C. E. Dunn-Fletcher, L. M. Muglia, M. Pavličev, G. Wolf, M.-A. Sun, Y.-C. Hu, E.
771 Huffman, S. Tumukuntala, K. Thiele, A. Mukherjee, S. Zoubovsky, X. Zhang, K. A. Swaggart,
772 K. Y. B. Lamm, H. Jones, T. S. Macfarlan, L. J. Muglia, Anthropoid primate-specific retroviral
773 element THE1B controls expression of CRH in placenta and alters gestation length. *PLoS*
774 *biology*. **16**, e2006337 (2018).

775 55. C. Krendl, D. Shaposhnikov, V. Rishko, C. Ori, C. Ziegenhain, S. Sass, L. Simon, N. S.
776 Müller, T. Straub, K. E. Brooks, S. L. Chavez, W. Enard, F. J. Theis, M. Drukker, GATA2/3-
777 TFAP2A/C transcription factor network couples human pluripotent stem cell differentiation to
778 trophectoderm with repression of pluripotency. *Proceedings of the National Academy of Sciences*
779 *of the United States of America*. **114**, E9579–E9588 (2017).

780 56. L. Gao, K. Wu, Z. Liu, X. Yao, S. Yuan, W. Tao, L. Yi, G. Yu, Z. Hou, D. Fan, Y. Tian, J.
781 Liu, Z.-J. Chen, J. Liu, Chromatin Accessibility Landscape in Human Early Embryos and Its
782 Association with Evolution. *Cell*. **173**, 248-259.e15 (2018).

783 57. J. Wu, J. Xu, B. Liu, G. Yao, P. Wang, Z. Lin, B. Huang, X. Wang, T. Li, S. Shi, N. Zhang,
784 F. Duan, J. Ming, X. Zhang, W. Niu, W. Song, H. Jin, Y. Guo, S. Dai, L. Hu, L. Fang, Q. Wang,
785 Y. Li, W. Li, J. Na, W. Xie, Y. Sun, Chromatin analysis in human early development reveals
786 epigenetic transition during ZGA. *Nature*. **557**, 256–260 (2018).

787 58. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nature methods*. **9**,
788 357–359 (2012).

789 59. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R.
790 Durbin, 1000 Genome Project Data Processing Subgroup, The Sequence Alignment/Map format
791 and SAMtools. *Bioinformatics (Oxford, England)*. **25**, 2078–2079 (2009).

792 60. J. M. Gaspar, Improved peak-calling with MACS2. *bioRxiv*, 496521 (2018).

793 61. H. M. Amemiya, A. Kundaje, A. P. Boyle, The ENCODE Blacklist: Identification of
794 Problematic Regions of the Genome. *Scientific Reports*. **9**, 9354 (2019).

795 62. A. R. Quinlan, I. M. Hall, BEDTools: a flexible suite of utilities for comparing genomic
796 features. *Bioinformatics (Oxford, England)*. **26**, 841–842 (2010).

797 63. J. T. Robinson, H. Thorvaldsdóttir, W. Winckler, Integrative genomics viewer. *nature.com*
798 (2011).

799 64. J. A. Thomson, J. Itskovitz-Eldor, S. S. Shapiro, M. A. Waknitz, J. J. Swiergiel, V. S.
800 Marshall, J. M. Jones, Embryonic Stem Cell Lines Derived from Human Blastocysts. *Science*.
801 **282**, 1145–1147 (1998).

802 65. L. Aagaard, B. Bjerregaard, A. L. Kjeldbjerg, F. S. Pedersen, L.-I. Larsson, J. J. Rossi,
803 Silencing of endogenous envelope genes in human choriocarcinoma cells shows that envPb1 is
804 involved in heterotypic cell fusions. *The Journal of general virology*. **93**, 1696–1699 (2012).

805 66. C. Luft, J. Freeman, D. Elliott, N. Al-Tamimi, J. Kriston-Vizi, J. Heintze, I. Lindenschmidt,
806 B. Seed, R. Ketteler, Application of Gaussia luciferase in bicistronic and non-conventional
807 secretion reporter constructs. *BMC biochemistry*. **15**, 14 (2014).

808 67. S. Knappskog, H. Ravneberg, C. Gjerdrum, C. Trösse, B. Stern, I. F. Pryme, The level of
809 synthesis and secretion of Gaussia princeps luciferase in transfected CHO cells is heavily
810 dependent on the choice of signal peptide. *Journal of biotechnology*. **128**, 705–715 (2007).

811 68. T. D. Schmittgen, K. J. Livak, Analyzing real-time PCR data by the comparative CT method.
812 *Nat Protoc*. **3**, 1101–1108 (2008).

813 69. M. J. Hubisz, K. S. Pollard, A. Siepel, PHAST and RPHAST: phylogenetic analysis with
814 space/time models. *Briefings in bioinformatics*. **12**, 41–51 (2011).

815 70. J. Ye, S. McGinnis, T. L. Madden, BLAST: improvements for better sequence analysis.
816 *Nucleic acids research*. **34**, W6-9 (2006).

817 71. M. Okhovat, K. A. Nevonen, B. A. Davis, P. Michener, S. Ward, M. Milhaven, L. Harshman,
818 A. Sohota, J. D. Fernandes, S. R. Salama, R. J. O'Neill, N. Ahituv, K. R. Veeramah, L. Carbone,
819 Co-option of the lineage-specific LAVA retrotransposon in the gibbon genome. *Proceedings of*
820 *the National Academy of Sciences of the United States of America*. **117**, 19328–19338 (2020).

821 72. A. M. Waterhouse, J. B. Procter, D. M. A. Martin, M. Clamp, G. J. Barton, Jalview Version
822 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics (Oxford,*
823 *England)*. **25**, 1189–1191 (2009).

824 73. S. Kumar, G. Stecher, K. Tamura, MEGA7: Molecular Evolutionary Genetics Analysis
825 Version 7.0 for Bigger Datasets. *Molecular Biology and Evolution*. **33**, 1870–1874 (2016).

826 74. G. Stecher, K. Tamura, S. Kumar, Molecular Evolutionary Genetics Analysis (MEGA) for
827 macOS. *Mol Biol Evol*. **37**, 1237–1239 (2020).

828 75. S. Kumar, G. Stecher, M. Li, C. Knyaz, K. Tamura, MEGA X: Molecular Evolutionary
829 Genetics Analysis across Computing Platforms. *Mol Biol Evol*. **35**, 1547–1549 (2018).

830 76. K. Tamura, M. Nei, Estimation of the number of nucleotide substitutions in the control
831 region of mitochondrial DNA in humans and chimpanzees. *Mol Biol Evol*. **10**, 512–526 (1993).

832