HEMO, an ancestral endogenous retroviral envelope protein shed in the blood of pregnant women and expressed in pluripotent stem cells and tumors

Odile Heidmann[a,b,1], Anthony Béguin[a,b], Janio Paternina[a,b], Raphaël Berthier[a,b], Marc Deloger[c], Olivia Bawa[d], and Thierry Heidmann[a,b]

[a]Unité Physiologie et Pathologie Moléculaires des Rétrovirus Endogènes et Infectieux, CNRS UMR 9196, Gustave Roussy, Villejuif, F-94805, France; [b]UMR9196, Université Paris-Sud, Orsay, F-91405, France; [c]Plateforme de Bioinformatique, INSERM US23/CNRS UMS3655, Gustave Roussy, Villejuif, F-94805, France; and [d]Plateforme d'Évaluation Préclinique, Laboratoire de Pathologie Expérimentale, Gustave Roussy, Villejuif, F-94805, France

Edited by John M. Coffin, Tufts University School of Medicine, Boston, MA, and approved June 6, 2017 (received for review February 17, 2017)

Capture of retroviral envelope genes is likely to have played a role in the emergence of placental mammals, with evidence for multiple, reiterated, and independent capture events occurring in mammals, and be responsible for the diversity of present day placental structures. Here, we uncover a full-length endogenous retrovirus envelope protein, dubbed HEMO [human endogenous MER34 (medium-reiteration-frequency-family-34) ORF], with unprecedented characteristics, because it is actively shed in the blood circulation in humans via specific cleavage of the precursor envelope protein upstream of the transmembrane domain. At variance with previously identified retroviral envelope genes, its encoding gene is found to be transcribed from a unique CpG-rich promoter not related to a retroviral LTR, with sites of expression including the placenta as well as other tissues and rather unexpectedly, stem cells as well as reprogrammed induced pluripotent stem cells (iPSCs), where the protein can also be detected. We provide evidence that the associated retroviral capture event most probably occurred >100 Mya before the split of Laurasiatheria and Euarchontoglires, with the identified retroviral envelope gene encoding a full-length protein in all simians under purifying selection and with similar shedding capacity. Finally, a comprehensive screen of the expression of the gene discloses high transcript levels in several tumor tissues, such as germ cell, breast, and ovarian tumors, with in the latter case, evidence for a histotype dependence and specific protein expression in clear-cell carcinoma. Altogether, the identified protein could constitute a "stemness marker" of the normal cell and a possible target for immunotherapeutic approaches in tumors.

HERV | endogenous retrovirus | envelope protein | placenta | development | stem cells | tumors

Endogenous retroviral sequences represent ∼8% of the human genome. These sequences [called human endogenous retroviruses (HERVs)] share strong similarities with present day retroviruses and are the proviral remnants of ancestral germ-line infections by active retroviruses, which have thereafter been transmitted in a Mendelian manner (1–3). The >30,000 proviral copies found in the human genome can be grouped into about 80 distinct families, with most of these elements being nonprotein-coding because of the accumulation of mutations, insertions, deletions, and/or truncations (4, 5). However, some retroviral genes have retained a coding capacity, and some of them have even been diverted by remote primate ancestors for a physiological role. The so-called "*syncytins*," namely *syncytin-1* and *-2* in humans, are retroviral envelope (*env*) genes captured 25 and 40 Mya, respectively, with a full-length protein-coding sequence, a fusogenic activity, and strong placental expression (6–9). These genes have been shown to be involved in placenta formation, with their fusogenic activity contributing to the formation of the syncytiotrophoblast (ST) at the maternofetal interface as a result of the syncytin-mediated cell–cell fusion of the underlying mononucleated cytotrophoblasts (CTs). Syncytins were, thereafter, identified in all placental mammals where they have been searched for, and their unambiguous role in placentation was shown via the generation and characterization of KO mice (10, 11). Syncytins are also present in marsupials, where they are expressed in a short-lived placenta that is very transiently formed (a few days) before the embryo pursues its development in an external pouch (12).

Previous systematic searches for genes encoding endogenous retroviral Env proteins within the human genome have led to the identification of 18 genes with a full-length coding sequence (among which are *syncytin-1* and *-2*) (13, 14). These analyses have been performed using methods based on the search for characteristic motifs carried by retroviral Envs (Fig. 1), which include, from the N terminus to the C terminus, a signal peptide; a furin cleavage site (R-X-R/K-R) between the surface (SU) and transmembrane (TM) subunits, with the latter carrying additional signatures including an immunosuppressive domain (ISD; 17-aa motif), which is also found in most oncoretroviruses; a characteristic C-$(X)_{5-7}$-C motif; and a transmembrane hydrophobic domain anchoring the Env protein in the cell or virion membrane (4, 15).

Less stringent methods based on BLAST searches using large panels of retroviral Env proteins, including the increasing number of newly identified ERV genes from other animals, led us to identify a gene encoding a full-length retroviral Env protein with unprecedented characteristics. This Env protein gene—dubbed HEMO [human endogenous MER34 (medium-reiteration-frequency-family-34) ORF]—is

Significance

Endogenization of retroviruses has occurred multiple times in the course of vertebrate evolution, with the captured retroviral envelope syncytins playing a role in placentation in mammals, including marsupials. Here, we identify an endogenous retroviral envelope protein with unprecedented properties, including a specific cleavage process resulting in the shedding of its extracellular moiety in the human blood circulation. This protein is conserved in all simians—with a homologous protein found in marsupials—with a "stemness" expression in embryonic and reprogrammed stem cells, as well as in the placenta and some human tumors, especially ovarian tumors. This protein could constitute a versatile marker—and possibly an effector—of specific cellular states and being shed, be immunodetected in the blood.

Author contributions: O.H. and T.H. designed research; O.H., A.B., J.P., R.B., M.D., and O.B. performed research; O.H., A.B., J.P., M.D., and T.H. analyzed data; and O.H., A.B., and T.H. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accession nos. MF320351–MF320355).

[1]To whom correspondence should be addressed. Email: odile.heidmann@gustaveroussy.fr.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1702204114/-/DCSupplemental.

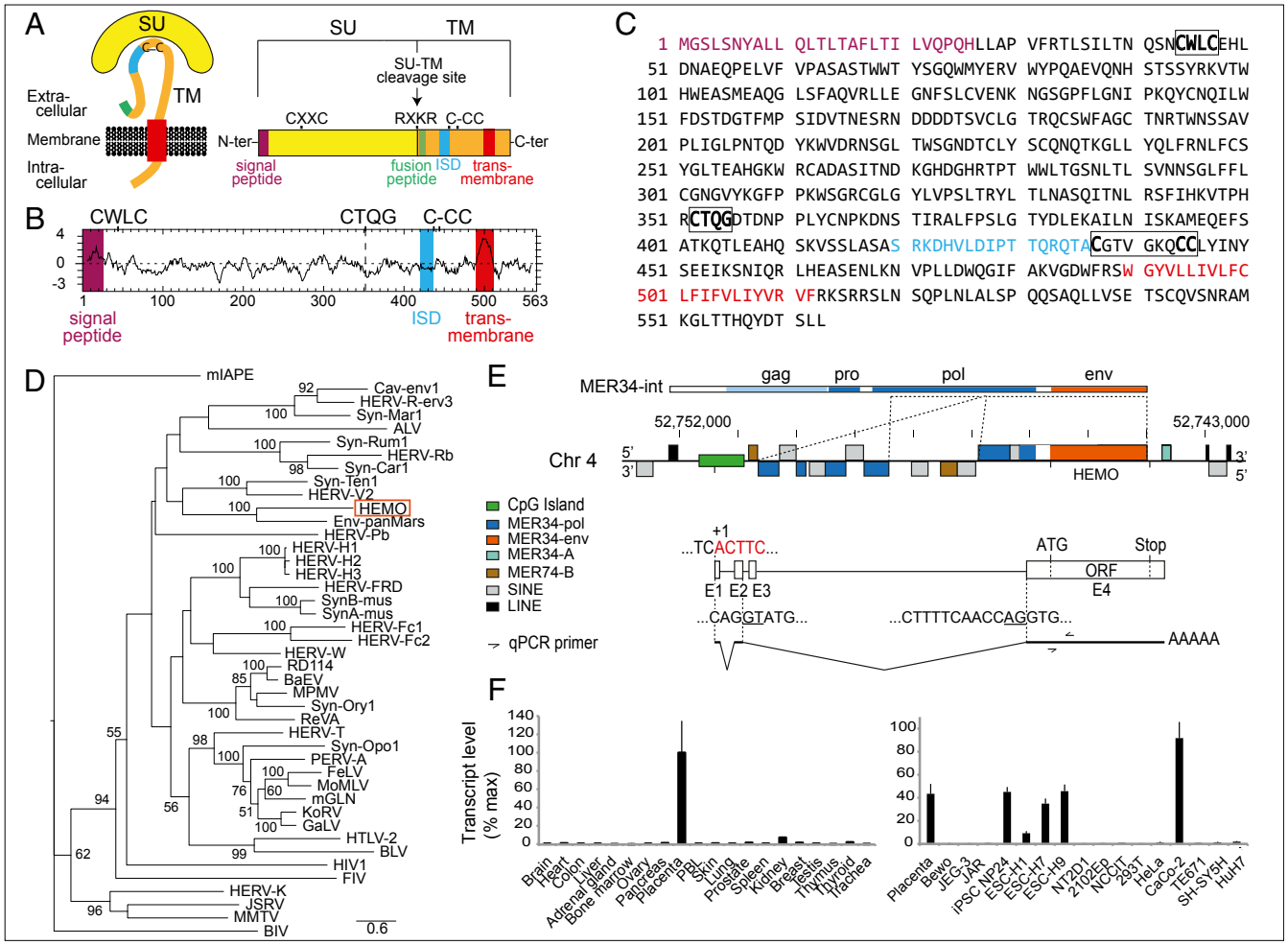

Fig. 1. Characterization of the human HEMO Env retroviral protein and the *HEMO env* gene. (*A*) Schematic representation of a canonical retroviral Env protein delineating the SU and TM subunits. The furin cleavage site (consensus: R-X-R/K-R) between the two subunits, the C-X-X-C motif involved in SU-TM interaction, the hydrophobic signal peptide (purple), the fusion peptide (green), the transmembrane domain (red), and the putative ISD (blue) along with the conserved C-X_{5-7}-CC motif (C-CC) are indicated. Adapted from ref. 38, copyright (2007) National Academy of Sciences. (*B*) Hydrophobicity profile of HEMO Env. The canonical structural features highlighted in *A* are positioned and shown in the color code used in *A*. The mutated furin site (CTQG) is shown as a dotted line. (*C*) Amino acid sequence of the HEMO Env protein with the same color code. (*D*) Retroviral Env protein-based phylogenetic tree with the identified HEMO-Env protein. The maximum likelihood tree was constructed using the full-length SU-TM amino acid sequences from HERV Envs (including an HERV-K consensus), all previously identified *syncytins*, and a series of endogenous and infectious retroviruses. The lengths of the horizontal branches are proportional to the average numbers of amino acid substitutions per site (scale bar in the lower right), and the percentage bootstrap values obtained from 1,000 replicates are indicated at the nodes. (*E*) Schematic representation of the *HEMO* gene locus on chromosome 4 (4q12; with the GRCh38 assembly coordinates of the Genome Reference Consortium). (*Top*) MER34-int consensus (Repbase) with putative gag, pro, and pol retroviral ORFs indicated according to consensus amino acid sequences. Dotted lines delineate parts of the MER34 sequences found in the *HEMO* locus. (*Middle*) The *HEMO* gene locus (10 kb) is located between the *RASL11B* gene (~120 kb 5′) and the *USP46* gene (~120 kb 3′). *HEMO env* ORF is shown as an orange box, and repetitive sequences identified on the Dfam.org website are shown as different colored boxes, with the sense sequences above and antisense sequences below the line. Of note, the gene is part of an MER34 provirus that has kept only degenerate *pol* sequences (mostly in opposite orientation), a truncated putative 3′ LTR (MER34-A), and no 5′ LTR. No other MER34 sequences are found 100 kb apart from the gene. A CpG island (chromosome 4:52750911–52751703), detected by the EMBOSS-newcpgreport software, is indicated as a green box. (*Bottom*) Intron–exon structure predicted from the National Center for Biotechnology Information and RNA transcripts: exons found in placental RNA, as determined by RACE experiments, are indicated with the main E1-E2-E4 spliced *env* subgenomic transcript below. Nucleotide sequences of the start site (ACTTC...; red) and large intron splice sites for the *HEMO env* ORF are depicted; arrows specify qRT-PCR primers (Table S3). (*F*) Real-time qRT-PCR analysis of the *HEMO* transcripts in a panel of 20 human tissues and 16 human cell lines. Transcript levels are expressed as percentage of maximum and were normalized relative to the amount of housekeeping genes (*SI Methods*). Placenta values are the means of 12 samples from first trimester pregnancies, and other tissues are from a commercial panel (Zyagen).

the oldest captured full-length *env* gene identified to date in humans, because it entered the genome of a mammalian ancestor more than 100 Mya. The HEMO protein is released in the human blood circulation via a specific shedding process closely related to that observed for the Ebola filovirus, and it is highly expressed by stem cells and also, by the placenta resulting in an enhanced concentration in the blood of pregnant women. It is also expressed in some human tumors, thus providing a marker for a pathological state as well as, possibly, a target for immunotherapies.

Results

Identification of *HEMO*, an HERV Gene Encoding a Full-Length Env Protein. The most recent human genome sequence release (GRCh38 Genome Reference Consortium Human reference 38, December 2013) was screened for the presence of genes encoding ERV Env proteins by a BLAST search for ORFs (from the Met start codon to the stop codon) > 400 aa using a selected series of 42 Env sequences representative of both infectious retrovirus and ERV families, including all of the previously identified syncytins (*SI*

Methods). It yielded 45 Env-encoding ORFs, which could be, for all except one, grouped by clustalW alignments into already known HERV Env families (among which 24 Env-encoding ORFs for HERV-K, and 20 Env-encoding ORFs belonging to the set of 12 previously described HERV Envs) (Table S1). However, an unrelated *env* gene (*HEMO*) can be identified (Fig. 1) with a 563-aa ORF displaying some—but not all—of the characteristic features of a full-length retroviral Env protein, namely an N-terminal signal peptide; a CWLC motif in the putative SU subunit; and in the TM subunit, an ISD domain, a C-X6-CC motif, and a 23-aa hydrophobic transmembrane domain followed by a C-terminal cytoplasmic tail. Of note, the putative HEMO protein lacks a clearly identified furin cleavage site (CTQG instead of the canonical R-X-R/K-R) as well as an adjacent hydrophobic fusion peptide (Fig. 1*B*). The HEMO sequence was incorporated into the Env phylogenetic tree shown in Fig. 1*D* containing 42 retroviral envelope amino acid sequences used for the genomic screen. Fig. 1*D* shows that the sequence most closely related to the HEMO protein is Env-panMars encoded by a conserved, ancestrally captured retroviral *env* gene found in all marsupials, which has a premature stop codon upstream of the transmembrane domain (12).

Finally, BLAST analysis of the human genome indicates that the *HEMO* gene is part of a very old degenerate multigenic family known as medium reiteration frequency family 34 (MER34; first described in ref. 16). In this family, an internal consensus sequence with a Gag-Pro-Pol-Env retroviral structure (MER34-int) and LTR-MER34 sequences have been described and reported in RepBase (17). Genomic BLAST with the MER34-int consensus sequence could not detect any full-length putative ORFs for the *gag* or *pol* genes. Among the *env* sequences of the MER34 family scattered in the human genome (20 copies with >200-bp homology identified by BLAST) (Table S2), *HEMO* is clearly an outlier (1,692 bp/563 aa), with all of the other sequences containing numerous stop codons, short interspersed nuclear elements (SINE) or long interspersed nuclear elements (LINE) insertions, and no ORF longer than 147 aa.

The *HEMO* Gene Locus and Transcription Profile. The *HEMO* gene is located on chromosome 4q12 between the *RASL11B* and *USP46* genes at about 120 kb from each gene (see Fig. 9). Close examination of the *HEMO env* gene locus (10 kb) by BLAST comparison with the RepBase MER34-int consensus (17) reveals only remnants of the retroviral *pol* gene in a complex scrambled structure (Fig. 1*E*), with part of it being in reverse orientation and further disrupted by numerous SINE insertions. The locus organization indicates low selection pressure for the proviral non-*env* genes, such as often observed in the previously characterized loci harboring captured *envs*.

A quantitative RT-PCR (qRT-PCR) analysis using primers within the identified ORF and RNAs from a panel of human tissues and cell lines (Fig. 1*F*) shows that *HEMO* is expressed at a high level in the placenta. It is also significantly expressed in the kidney but at a lower level. In cell lines, expression of the *HEMO* gene looks heterogeneous, except for its systematic expression in stem cells [embryonic stem cells (ESCs) and induced pluripotent stem cells (iPSCs)]. Quite unexpectedly, there is an absence of detectable transcripts in several placental choriocarcinoma cell lines (BeWo, JEG-3, and JAR) as well as a series of embryonal carcinoma (NT2D1, 2102Ep, and NCCIT) and tumor cell lines (but see the CaCo-2 colon adenocarcinoma).

The structure of the *HEMO env* transcripts was determined by RACE-PCR analysis of Env-encoding transcripts from the placenta. It allowed the identification of multiply-spliced transcripts, with the intron boundaries corresponding to donor/acceptor splice sites predicted from the genomic sequence and as classically observed for retroviral *env* genes, a functional acceptor site located close to the *env* ATG start site. Interestingly, the transcript 3′ end falls within an identifiable MER34 LTR, as expected for a retroviral transcript. However, the transcription start site, located ∼5 kb 5′ to the *env* gene, does not correspond to any identifiable LTR structure. Rather, the sequence associated with the transcript start site is located in a CpG-rich domain (Fig. 1*E* and Fig. S1*A*) and most probably corresponds to a cellular promoter unrelated to any retroviral element. The transcript 5′ end (i.e., tc|ACTTC) falls within a canonical RNA Polymerase II Core Promoter Initiator Motif [yy|ANWYY (18)].

The CpG-rich start site-containing region (CpG island) (reviewed in ref. 19) was studied further for its promoter activity by in vitro transfection assays using luciferase reporter genes. As illustrated in Fig. S1*B*, a 760-bp fragment including the identified start site acts as a strong promoter in this assay (>500-fold compared with none). Lower expression is observed (10- to 50-fold compared with none) in partial deletion mutants and as expected for a CpG promoter, when placed in antisense orientation.

DNA methylation patterns of sequences surrounding the transcription start site within the identified CpG island were analyzed by bisulfite treatment. As shown in Fig. S1*C*, the majority of the CpGs are methylated in the HEMO-negative cell lines (293T and BeWo), whereas they are unmethylated in HEMO-expressing cell lines (iPSC and CaCo-2). To get additional insight into this dependence of the promoter activity on the CpG island methylation pattern, 5-Aza-2′-deoxycytidine (5-Aza-dC) treatment was performed on BeWo and 293T cells at doses ranging from 0.1 to 5 μM (Fig. S1*D*). Transcripts were detectable by qRT-PCR after a 3-d treatment at low dose for BeWo cells (0.1 μM) and higher dose for 293T cells (5 μM). Of note, the high transcript level of the *HEMO* gene in CaCo-2 cells was not further amplified by a similar 5-Aza-dC treatment. Altogether, these results indicate that *HEMO* expression is sensitive to the methylation status of the CpG promoter.

HEMO Protein Synthesis and Structure: Specific Shedding. The capacity of the identified gene to produce an envelope protein was tested by introduction of the *env* ORF into a CMV promoter-driven expression vector and in vitro transient transfection assays. Polyclonal antibodies and mAbs were raised by immunization of mice with a recombinant protein corresponding to a 163-aa fragment of the putative SU moiety of the protein (*SI Methods*). As illustrated by the immunofluorescence assay shown in Fig. 2, *Upper* using the anti-HEMO antibodies, a strong labeling can be observed on permeabilization of the transfected cells (and not of control cells transfected with an empty vector). Furthermore, HEMO proteins can be detected at the cell surface as evidenced by the specific immunofluorescence labeling of the cell membrane of nonpermeabilized transfected HeLa cells in the successive confocal images shown in Fig. 2, *Lower*, consistent with *HEMO* being a retroviral *env* gene.

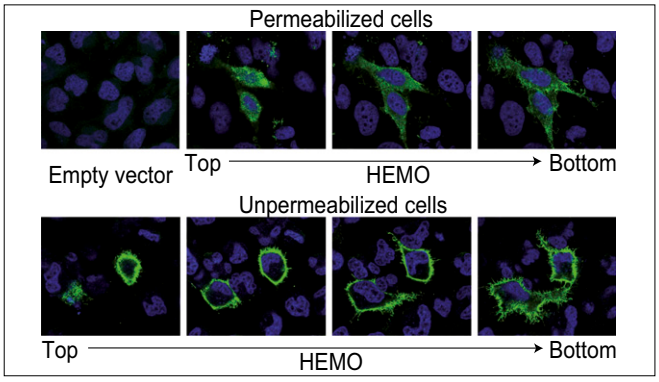

Fig. 2. Immunofluorescence analysis of HEMO protein expression in transfected HeLa cells. Cells (HeLa) were transfected with the phCMV-HEMO expression vector (or an empty vector as a negative control), fixed, permeabilized (*Upper*) or not permeabilized (*Lower*), and stained for HEMO protein expression using a specific anti-HEMO polyclonal antibody (*SI Methods*). (*Upper*) Specific staining of the phCMV-HEMO transfected cells vs. empty vector transfected cells. (*Lower*) Successive confocal images show cell surface localization of the protein.

As illustrated in the Western blot of a whole-cell lysate (Fig. 3*A*, lane 3), transfection with the above HEMO expression vector yielded a strong band with an apparent molecular mass >80 kDa, much larger than expected for the HEMO full-length SU-TM protein (theoretical molecular mass = 61 kDa) but consistent with its glycosylation—as expected for a retroviral protein. Indeed, treatment of the cell extract with peptide *N*-glycosidase F (PNGase F)—to deglycosylate proteins—resolved the >80-kDa band into two bands of lower molecular mass (lane 4): a major band of ∼58 kDa and a fainter one of ∼48 kDa. The major band most probably corresponds to the full-length SU-TM protein (estimated size of 61 kDa), whereas the lower 48-kDa band has a size inconsistent with that of the sole SU subunit (estimated size 37 kDa; see below).

Analysis of the cell supernatants provided an unexpected answer as to the origin of the 48-kDa protein. Indeed, this 48-kDa protein turns out to be the major form in the cell supernatant (Fig. 3*A*, lane 6) (with PNGase F treatment of the supernatant), whereas the larger 58-kDa band observed in the whole-cell extract (Fig. 3*A*, lane 4) (with similar PNGase F treatment) is almost undetectable, as expected for a cell membrane-attached full-length Env protein (Fig. 3*A*, lane 6). This secreted 48-kDa protein is glycosylated, being observed at a much higher molecular mass in the cell supernatant without PNGase F treatment (Fig. 3*A*, lane 5). Altogether, these data strongly suggest that the HEMO protein, which is a transmembrane protein exported at the cell surface, can nevertheless be quantitatively released in the supernatant in the form of a protein—Shed Env—which has a molecular mass that is larger than that of the SU alone. This property, unexpected for a retroviral Env protein, is indeed not observed using the same protocols and expression vectors for syncytin-1 (HERV Env-W) used as a negative control (Fig. 3*A*, lanes 1 and 2).

To go further into the characterization of this shed soluble protein, we purified it from the supernatant of transfected 293T cells (*SI Methods*) and characterized its sequence by using mass spectrometry (MS) for the determination of both its N and C termini. As illustrated in Fig. 3*B* (and Fig. S2, which provides the HEMO protein sequence coverage by MS analysis of trypsin- or chymotrypsin-generated peptides), it turns out that the shed protein is truncated at its C terminus, mainly within the ISD domain, with two C-terminal sites identified with a different abundance (namely, Q432 and R433 at a 4:1 ratio). At the N terminus, the HEMO protein begins at position 27 (i.e., 2 aa after the predicted signal peptide cleavage site; using SignalP 4.1 Server software; www.cbs.dtu.dk/services/SignalP/). To confirm the MS size determination of the shed HEMO protein, several mutants were constructed by inserting stop codons before the anchoring transmembrane domain at the indicated positions (marked in Fig. 3 *B* and *C* with asterisks: 433R-stop, 472P-stop, and 489S-stop) or introducing the consensus furin site RTKR in place of the human CTQG [as in the New World monkeys (NWMs); human furin+ construct (H-fur+)] (Fig. S3). Western blot analysis of the supernatant of the HEMO mutant transfected cells then clearly showed that the wild-type (WT) deglycosylated shed HEMO protein migrates as the 433R-stop mutant as expected. In addition, the H-fur+ mutant displays a smaller 37-kDa band, consistent with the size estimated for the deglycosylated SU subunit alone. Of note, the figure also shows that the 489S- and 472P-stop mutant proteins do not comigrate with the shed WT HEMO—but are larger—indicating that, despite the fact that they still contain the shedding sequence, they have not been further processed by the shedding machinery. This absence of shedding is most probably because they are not membrane-associated (as a consequence of the premature stop codon introduced before the transmembrane domain), thus suggesting that anchoring the Env protein at the cell surface is required for an efficient processing by the shedding machinery.

To determine if the shed form of the HEMO protein could be observed under in vivo conditions, placental tissues (which show high transcription levels for the *HEMO* gene) (Fig. 1*F*) were recovered from first trimester legal abortions together with the local placental blood (which bathes the placental villi and can be analyzed in parallel), and proteins were extracted and deglycosylated for Western blot analyses. As shown in Fig. 3*A*, lane 7, the small 48-kDa band (and a very faint SU-TM 58-kDa band) can be detected in the placental tissue extract. The 48-kDa band is also detected in the placental blood, most probably corresponding to the protein secreted by the placenta. MS analysis (as above) of the 48-kDa protein in the corresponding gel bands confirmed the relevance of the immunological detection.

The release of a processed HEMO protein is reminiscent of what has been observed for the viral envelope protein of a completely unrelated virus (i.e., the Ebola filovirus), for which it has been further shown that cleavage was mediated by a cell-associated metalloproteinase ADAM protein (20). Accordingly, we tested whether chemical inhibitors of metalloproteinases [including the ADAM and MMP proteins (20–22)] had any effect on HEMO shedding in 293T transfected cells. As illustrated in Fig. 4, the broad-range ADAM and MMP inhibitors Batimastat and Marimastat and the MMP inhibitor GM6001 clearly inhibited HEMO release in the supernatant to various extents and in a dose-dependent manner, with visible accumulation of the nonsecreted form in the cell lysates. These experiments suggest that, in vivo, HEMO shedding could be driven by one or

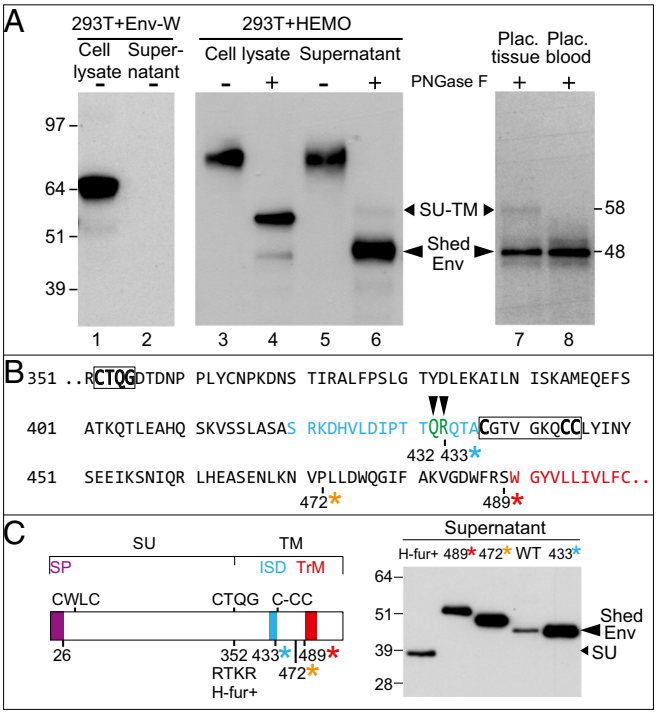

Fig. 3. Characterization of the shed HEMO protein. (*A*) Detection of the shed HEMO Env protein by Western blot analysis. (*Left*) Detection of the syncytin-1 protein with the anti–Env-W polyclonal antibody (58) in the cell lysate of phCMV-Env-W transfected 293T cells. (*Center* and *Right*) Detection of the two forms of the HEMO protein (full-length SU-TM and Shed Env) with the anti-HEMO polyclonal antibody in the cell lysate and supernatant of phCMV-HEMO transfected 293T cells (*Center*) and first trimester placental tissue and placental blood (*Right*; matched representative samples from the same individual); samples were treated (+) or not (–) with PNGase F. (*B*) Detail of the shedding site amino acid sequence indicated by green capital letters. *Positions of the stop codons introduced in the mutants analyzed in *C*. (*C*) Migration pattern of the mutant HEMO forms analyzed as in *A*. (*Left*) Schematic representation of the HEMO protein with the same color code as in Fig. 1 and the stop codons of the generated mutants positioned together with that of the mutant with a reconstituted furin site (H-fur+; with an RTKR furin site). (*Right*) Supernatant of 293T cells transfected with the expression vectors for the WT and the mutant HEMO plasmids analyzed after PNGase F treatment, SDS gel electrophoresis, and Western blot as in *A*.

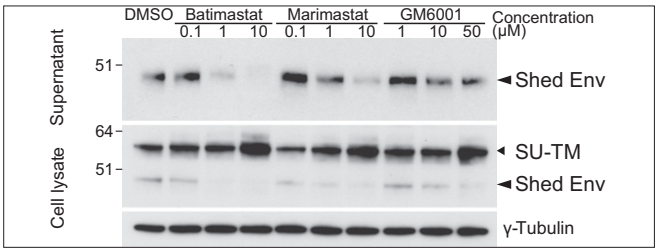

Fig. 4. Inhibition of HEMO release in the supernatant of transfected cells. Western blot analysis of cell lysate and supernatant of 293T cells transfected with the phCMV-HEMO expression vector using the polyclonal anti-HEMO antibody (*SI Methods*). Cells were treated for 3 d with the indicated doses of the ADAM and MMP chemical inhibitors Batismastat, Marismastat, and GM6001 or DMSO alone. Anti–γ-tubulin antibody was used as a control of cell lysate protein loading. The full-length HEMO protein (SU-TM) and the secreted form (Shed Env) are indicated by arrowheads.

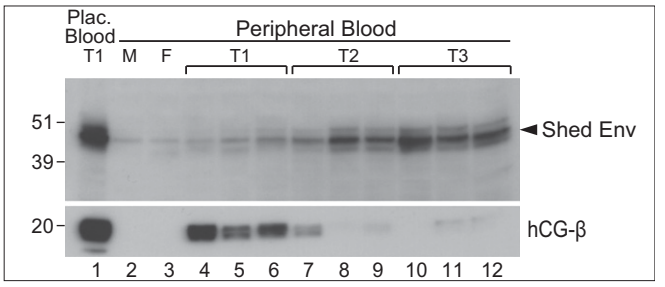

Fig. 5. Release of the HEMO protein in the peripheral blood during pregnancy. Western blot analysis of purified blood samples with the polyclonal anti-HEMO antibody (*Upper*) and anti–hCG-beta antibody (*Lower*). The shed HEMO protein is detected in the placental blood from first trimester pregnancy (T1) and from peripheral blood of men (M), nonpregnant women (F), and pregnant women from the first (T1), second (T2), and third (T3) trimesters. Each lane corresponds to distinct individuals. Bands observed at both slightly higher and lower molecular masses might correspond to minor alternatively processed/shed forms of the HEMO protein.

several metalloproteinases known to be present notably in placental cells (23–25).

HEMO Expression in Vivo: HEMO Release in the Blood Circulation of Pregnant Women. The combined results of the qRT-PCRs on the panel of human tissues shown in Fig. 1*F* and the shedding of the protein shown in Fig. 3 led us to hypothesize that HEMO could be detected in the blood circulation, especially in pregnant women. Sera were, therefore, collected and assayed for the presence of shed HEMO by Western blotting. Sera were treated with wheat germ agglutinin (WGA) to isolate glycosylated proteins, which were then deglycosylated. As illustrated in Fig. 5, *Lower*, the hCG-beta protein, which is a well-known early biomarker of pregnancy (26), shows undetectable levels in the peripheral blood of men and nonpregnant women (Fig. 5, lanes 2 and 3), whereas a very high level is observed for women in the first trimester of pregnancy (20-kDa band) (Fig. 5, lanes 4–6, T1), with a decrease at later stages (Fig. 5, lanes 7–12, T2 and T3). Remarkably, the deglycosylated shed HEMO form (48 kDa; previously identified in the placental blood) (Fig. 3*A*, lane 8 and Fig. 5, *Upper*, lane 1) can also be detected in the peripheral blood of pregnant women, beginning at a faint level in first trimester pregnancies (Fig. 5, *Upper*, lanes 4–12). As pregnancy proceeds, the level of HEMO protein increases very significantly, consistent with the large increase in placental mass during pregnancy. HEMO concentration at the peak can be estimated to be in the 1- to 10-nM range [by comparative Western blot analysis of serial dilutions of a purified recombinant shed HEMO protein; i.e., about one to two logs below that for hCG at the peak (T1) and for additional comparison, about the same as that for alpha-fetoprotein in the blood of pregnant women at the peak (T2)]. Of note, a faint level of shed HEMO protein can also be observed in the blood of men and nonpregnant women (Fig. 5, *Upper*, lanes 2 and 3), consistent with its nonnegligible expression in other organs, such as the kidney (qRT-PCR results in Fig. 1*F*).

Identification of HEMO-Producing Cells in the Placenta. The human placenta is of the hemochorial type and characterized by the presence of fetal villi in direct contact with—and bathed by—the maternal placental blood (Fig. 6*A*) (9). These villi arise from the chorionic membrane—of fetal origin—and have an inner mononucleated cytotrophoblast layer (CT) underlying the surface syncytial layer, the syncytiotrophoblast (ST) (reviewed in refs. 27 and 28). The placenta invades the maternal uterine part, with anchoring villi characterized by invasive extravillous trophoblasts (EVTs).

To localize precisely HEMO expression in the placenta, immunohistochemistry experiments were then performed on sections of first trimester placental tissues from abortion cases. As illustrated in Fig. 6 *B* and *C*, specific staining was obtained with the monoclonal anti-HEMO antibody and not with a control isotype as shown in Fig. 6*B* (4× magnification). In the three enlargements (60× magnification) shown in Fig. 6*C*, strong staining is observed in the trophoblast cells, including the villous CTs and the EVTs, suggesting that HEMO is indeed produced by these cells. More diffuse staining can be observed in the ST layer (Fig. 6*C3*), which is generated by CT fusion and involved in the exchanges between fetal and maternal blood.

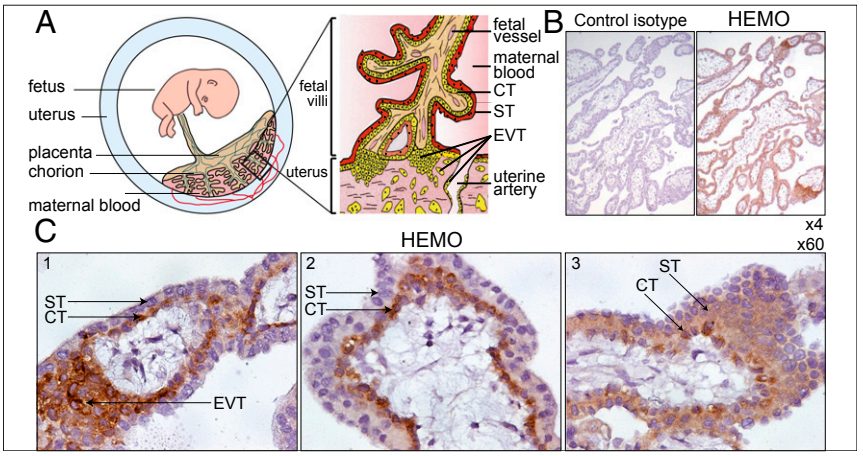

Fig. 6. Immunohistochemical detection of the HEMO protein in first trimester human placenta. (*A*) Schematic representation of the fetoplacental unit with an enlarged anchored villus bathed by maternal blood and displaying the ST layer, the underlying mononucleated CT, and the invading EVT. Adapted from ref. 59. (*B*) Serial sections of multiple placental villi stained with a control IgG2a mouse isotype (*Left*) or the anti-HEMO mAb 2F7 (*Right*). (Magnification: 4×.) (*C*) Enlarged views of placental villi showing staining of CTs (*1–3*), with strong staining of EVTs (*1*) and diffuse staining of STs (*3*). (Magnification: 60×.)

Altogether, the immunohistochemical analyses of the placenta carried out with the above anti-HEMO antibody show strong labeling essentially at the trophoblast level and are consistent with the observed shedding of HEMO in the mother's blood (Fig. 5).

Profile of *HEMO* Expression in Development. To get insight into the possible involvement of *HEMO* in embryonic development, we further analyzed by data mining a series of human RNA sequencing (RNA-Seq) experiments deposited at the Sequence Read Archive National Center for Biotechnology Information platform corresponding to different stages of development (29–32). Extraction of the expression profiles of a set of human genes was performed, and the results are illustrated in Fig. 7*A* for the *HEMO*, the *syncytin 1* (*env-W*), and the *syncytin 2* (*env-FRD*) *env* genes as well as specific genes expressed in either the placenta (*GCM1*) or stem cells (*OCT4*/*POU5F1*). For each gene of interest, read counts were verified to be equally distributed over the coding sequence (*SI Methods*). All three *env* genes (together with the placental GCM1-specific gene) are found in the RNA-Seq samples of placental tissues as expected (Fig. 7*A*, *Left*). Fig. 7*A*, *Center* clearly shows that *HEMO* has a wide expression profile, being expressed early in embryonic development starting at the eight-cell stage up to the late blastocyst stage and being permanently expressed in the derived ESCs from passage 0 up to passage 10. The *HEMO* gene RNA-Seq expression profile found in stem cells confirms the qRT-PCR results shown in Fig. 1*F* and is clearly different from what is observed for the two human syncytin genes: *env-W*, which is expressed very early in development, is completely down-regulated in the human stem cells, and *env-FRD* remains almost undetectable (33). Finally, RNA-Seq expression of *HEMO* was analyzed in the reprogramming experiments of differentiated somatic cells into iPSCs described in ref. 32, and hits reported in Fig. 7*A*, *Right* highlight the specific reprogramming of the *HEMO* gene—not observed with *env-W* and *env-FRD*—which parallels the expected profile of expression of the OCT4/POU5F1 transcription factor. Of note, as illustrated in Fig. 7*B* at the protein level, we could verify by Western blot analysis of iPSCs in culture that the *HEMO* gene expression unraveled above also results in the shedding of HEMO proteins, with a 48-kDa band detected in the iPSC supernatants.

Conclusively, the *HEMO* gene displays a specific pattern of expression—that includes ESCs—a feature possibly linked to the "capture" of a specific CpG-rich promoter of non-LTR origin, with the bona fide production of HEMO in the form of a soluble protein from at least trophoblast and stem cells.

HEMO Expression in Tumors. To get insight into the possible expression of the *HEMO* gene in human tumors, we performed an in silico analysis of microarray data using the dataset E-MTAB-62 elaborated in ref. 34, which includes 1,033 samples from normal tissues and 2,315 samples from neoplasm tissues obtained from various ArrayExpress (AE) and Gene Expression Omnibus (GEO) studies (*SI Methods*). In normal tissues, as expected from the qRT-PCR analysis in Fig. 1*F*, significant levels of expression were essentially observed in placental tissues and to a limited extent, the kidney (Fig. 8*A*). In several tumors, as illustrated in Fig. 8*B*, heterogeneity was detected among samples from the same organ (represented by the outliers plotted as black dots in Fig. 8*B*), with in some cases, evidence for high-level expression of the *HEMO* gene (for instance, in germ-line, liver, lung, or breast tumors, with the most salient heterogeneity being observed for ovary tumors). In the latter case, additional search for annotation data related to various histological types of ovarian carcinoma (35, 36) led us to correlate the highest values with specific tumor histotypes, mainly clear cell carcinoma. To enlarge this dataset, ovary tumor samples from five other GEO databases were collected and further normalized (*SI Methods*) together with E-MTAB-62, giving a total of 479 tumor samples. As shown in Fig. 8*C*, higher expression values of the gene are observed for clear cell carcinomas (60 samples) and to a lesser extent, endometrioid cancer (96 samples) samples. No clear-cut up-regulation of the *HEMO* gene is observed in the serous cancer histotype (289 samples; albeit with some heterogeneity) and the mucinous histotype (34 samples).

In agreement with these transcription data, immunohistochemistry analyses of normal vs. clear cell carcinoma ovarian tissues using the anti-HEMO mAb disclose a highly specific staining of the tumoral clear cells compared with the control isotype staining (Fig. 8*D*).

***HEMO* Insertion Date and Conservation Across Mammalian Genomes.** A strong hint for a physiological role of a captured gene is its

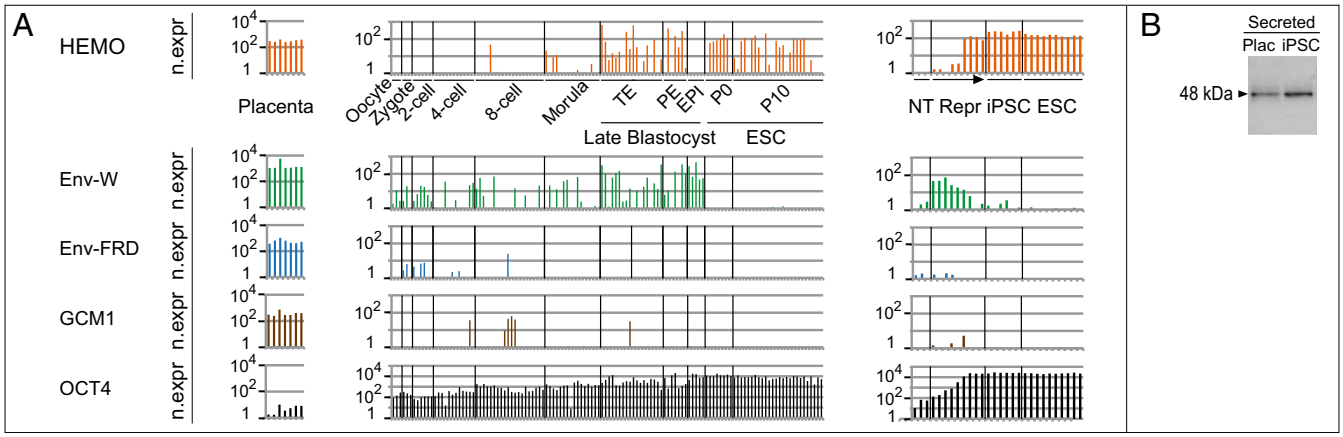

Fig. 7. Expression of the *HEMO* gene during development by in silico RNA-Seq analysis. (*A*) In silico analysis of three panels of RNA-Seq data for *HEMO*, *syncytin-1* (*env-W*) and *-2* (*env-FRD*), *GCM1* (Glial Cells Missing homolog 1, a specific placenta-expressed gene), and *OCT4* (highly expressed in stem cells). RNA-Seq raw data were screened with the coding part of each gene, and hits were reported in log scale per kilobase of screened sequence and after normalization with two housekeeping genes, RPLP0 and RPS6 (*SI Methods*). (*Left*) Panel of seven samples of normal placental tissues from distinct individuals at the same stage of pregnancy (29). (*Center*) Panel of 124 single-cell RNA-Seq of human preimplantation embryos and ESCs at the indicated stages of development or cell passage (30) (similar patterns were obtained from data in ref. 31, which covered the oocyte to morula stages). (*Right*) Panel of 28 RNA-Seq samples from the reprogramming (Repr.; from day 4 to day 21) of human CD34+ cells (NT) to iPSCs (six subclones) and from independent human ESC lines (32). EPI, epiblast; n. expr, normalized expression; P0, Passage 0; P10, Passage 10; PE, primitive endoderm; TE, trophectoderm. (*B*) Western blot analysis of WGA-purified placental blood (first trimester pregnancy) and WGA-purified supernatant of confluent iPSC cloneN (grown an extra 36 h without serum and concentrated 20×). Samples were treated with PNGase F. The shed HEMO form is detected using the polyclonal anti-HEMO antibody.

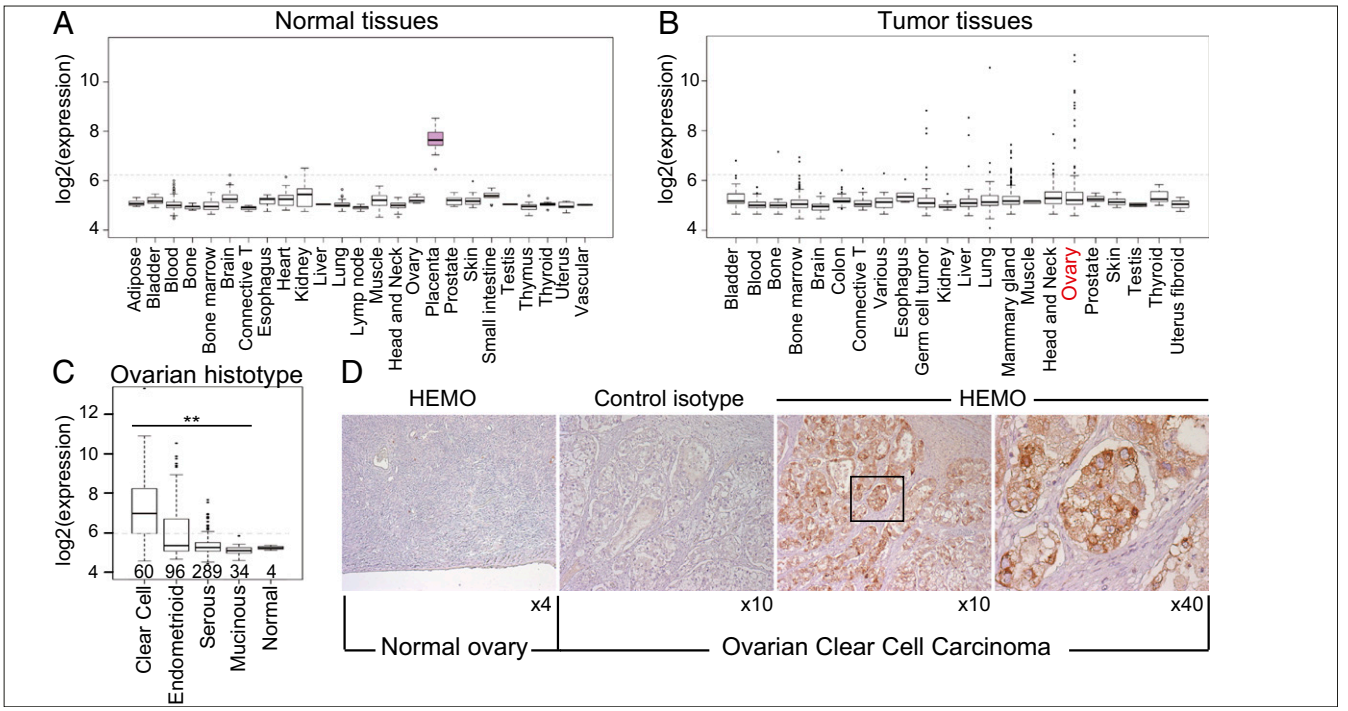

Fig. 8. Microarray analysis of *HEMO* expression within normal tissues and tumor samples. (*A* and *B*) Box plot representations of normalized values obtained for *HEMO* gene expression extracted from the E-MTAB62 dataset (on a logarithmic scale). Original tissue categories were adjusted to group together samples from the same biological source, keeping the major groups described by the authors: normal tissues (*A*) and tumor samples (*B*). (*C*) Box plot representation of normalized values obtained from an enlarged ovarian tumor sample extracted as raw.CEL files from various AE and GEO studies (*SI Methods*). Values for normal ovarian tissues were included, as control, in the normalization process. Tumoral ovarian histotypes correspond to 60 clear cell carcinoma, 96 endometrioid, 34 mucinous, and 289 serous tumoral samples (Wilcoxon's rank sum test). $^{**}P < 0.01$. (*D*) Immunohistochemical analysis using the 2F7 mAb specific for the HEMO protein (or a control isotype) of formalin-fixed normal ovarian tissues (column 1) and ovarian clear cell carcinoma (column 2–4; at two magnifications).

conservation in evolution and the nature of the selection to which it is subjected. Accordingly, we performed an extensive search for the *HEMO* gene in eutherian mammals by both in silico screening and PCR cloning and sequencing and further extended it to marsupials (the phylogenetic tree in Fig. 1*D* shows homology of HEMO with Env-panMars; see below). These analyses also aimed at the determination of the *HEMO* date of insertion into the genome of a mammalian ancestor, the determination of the coding capacity of the identified genes in the various species, and the determination of the presence of a shed HEMO protein after introduction of the cloned gene into an expression vector and transfection of 293T cells. The overall data are summarized in Fig. 9. We performed an in silico analysis of syntenic loci (coordinates listed in Table S4) by using the MultiPipMaker synteny building tool between the *RASL11B* and *USP46* genes conserved in all mammalian genomes (each found at about 120 kb from the human *HEMO* gene). Focus on the 15-kb *HEMO* region (Fig. 9*A*) shows that the *HEMO* gene entered the genome of mammals before the radiation of Laurasiatherians (e.g., ruminants, carnivora) and Euarchontoglires (e.g., primates, rodents, lagomorphes), i.e., between 100 and 120 Mya (37), being found in neither Afrotherians (Elephant, Tenrec) nor Xenarthrans (Armadillo). It also allowed the identification of the orthologous *HEMO* gene in primates (and as a very degenerate sequence in rodents) and among Laurasiatherians in several species including the dog, cat, horse, and cow. Closer analysis further discloses that the *HEMO* gene (coordinates and sequences in Table S5 and Dataset S1) has been conserved as a full-length protein-coding sequence in all simians (Fig. S3) and unexpectedly, the cat (Fig. S4). The identified full-length *HEMO* ORFs show high similarities, ranging from 84 to 99% amino acid identities (Fig. 9*B*, lower triangle), and signs of purifying selection, with nonsynonymous to synonymous ratios (dN/dS) between all pairs of species lower than unity (mean value 0.46), except for very close species (e.g., human/chimpanzee), for which the number of mutations is not high enough to provide significant dN/dS values. For example, dN/dS values of 0.29–0.42 are observed between great apes and Old World monkeys (OWMs) (Fig. 9*B*, upper triangle) as expected for a host gene. These low values further fall within those that can be determined for the similarly captured primate syncytin-1 (0.85) and syncytin-2 (0.28) functional *env*-derived genes.

To test the conservation of the specific shedding property observed in humans, a series of simian *HEMO* genes were cloned, introduced into the phCMV expression vector, and tested by transfection of 293T cells as described above. As shown in Fig. 9*C*, the *HEMO* genes from all of the tested species encode a protein that can be detected with the human HEMO antibodies (yet with a lower intensity for the distant NWMs), with in all cases, evidence for protein shedding in the cell supernatant. Even in the NWM branch, where the HEMO protein has retained a functional furin site (Fig. S3), a shed form of the protein is released in the supernatant together with a smaller SU form. The smaller size observed for the spider monkey protein is consistent with a small 10-aa deletion in the 5′ part of the gene (amino acids 182–191) (Fig. S3). Accordingly, it seems that the shedding of the HEMO protein is a very well-conserved property among simians, a feature that, together with the purifying selection applying to this gene, is a hint for a possible role of this secreted protein, notably in pregnant females. Of note, the domains 3′ to the shed protein form (Fig. S3) are much less conserved at the sequence level among simians, except for the transmembrane anchoring domain that seems to be required for shedding of the HEMO protein at the cell membrane (Fig. S3).

A Related *HEMO* Gene in Marsupials. To determine whether *HEMO*-like sequences could be present in some species where the orthologous gene could not be identified, a less stringent BLAST

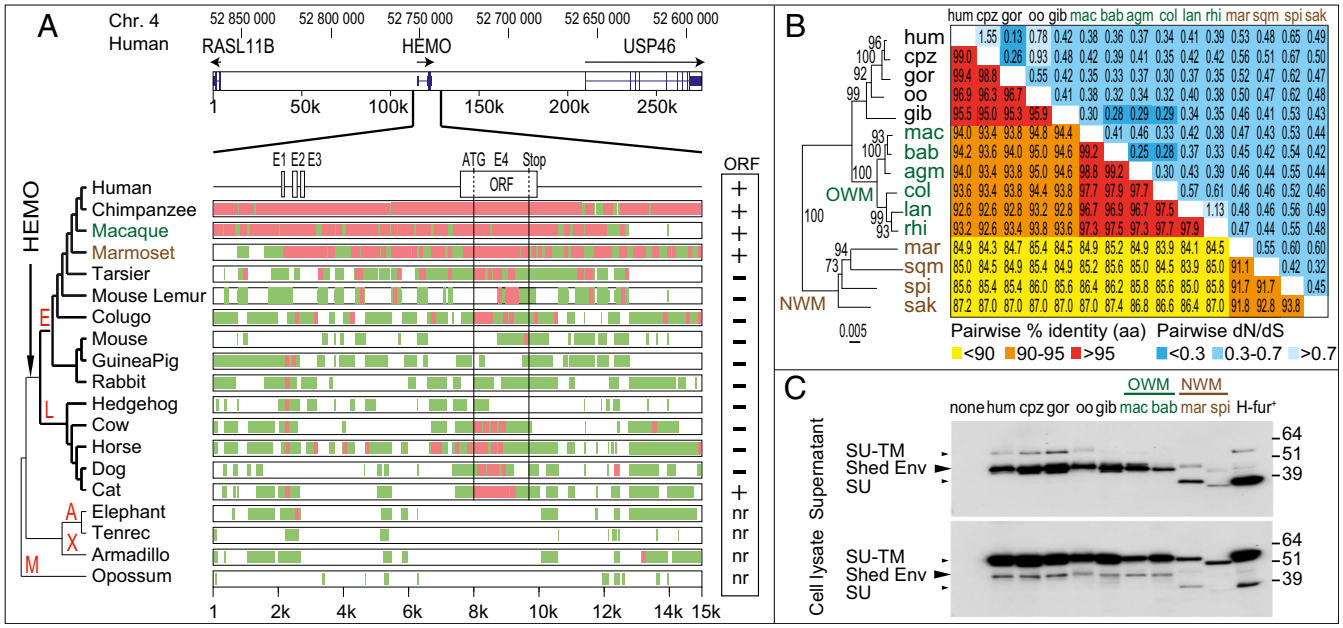

Fig. 9. Sequence conservation and purifying selection of the *HEMO* gene in simians. (*A*) Syntenic conservation of the *HEMO* locus in mammalian species. The genomic locus of the *HEMO* gene on human chromosome 4 along with the surrounding *RASL11B* and *USP46* genes (275 kb apart; genomic coordinates listed in Table S4) was recovered from the UCSC Genome Browser together with the syntenic loci of the indicated mammals from five major clades [Euarchontoglires (E), Laurasiatherians (L), Afrotherians (A), Xenarthres (X), and Marsupials M)]; exons and sense of transcription (arrows) are indicated. Exons of the *HEMO* gene (E1–E4) are shown on an enlarged view of the 15-kb *HEMO* locus together with the homology of the syntenic loci (analyzed using the Multi-PipMaker alignment-building tool). Regions with significant homology as defined by the BLASTZ software (60) are shown as green boxes, and highly conserved regions (more than 100 bp without a gap displaying at least 70% identity) are shown as red boxes. Sequences with (+) or without (–) a full-length HEMO ORF are indicated on the right. nr, not relevant. (*B*) Purifying selection in simians. *HEMO*-based maximum likelihood phylogenetic tree was determined using nucleotide alignment of the *HEMO* genes (listed in Table S5 and Dataset S1). The horizontal branch length and scale indicate the percentage of nucleotide substitutions. Percentage bootstrap values obtained from 1,000 replicates are indicated at the nodes. Double-entry table for the pairwise percentage of amino acid sequence identity (lower triangle) and the pairwise value of dN/dS (upper triangle) between the *HEMO* gene from the various simian species listed on the phylogenetic tree to the left and listed in the same order in abbreviated form at the top. A color code is provided for both series of values. (*C*) Conservation of HEMO shedding in simians illustrated by Western blot analysis of 293T cells transfected with expression vectors for the indicated simian *HEMO* genes or the human *HEMO* mutant with a consensus furin site (H-fur+). Cell lysates and supernatants were harvested and treated with PNGase F before Western blot analysis with the polyclonal anti-HEMO antibody. The entire SU-TM HEMO protein is the main form observed in cell lysates, whereas the shed and the free SU form (for the NWM genes with a furin site and the H-fur+ mutant) are mainly observed in the supernatants. agm, African green monkey; bab, baboon; col, Angolan black-and-white colobus; cpz, chimpanzee; gib, gibbon; gor, gorilla; hum, human; lan, langur; mac, macaque; mar, marmoset; NWM, New World monkey; oo, orangutan; OWM, Old World monkey; rhi, golden snub-nosed monkey; sak, saki monkey; spi, spider monkey; sqm, squirrel monkey.

search was performed, which provided hits in Marsupials—but still in neither Afrotherians nor Xenarthrans. Of note, the closest *env* gene identified is a conserved marsupial *env* gene that we had previously identified to be present in all marsupials (12), namely *env-panMars* (the phylogenetic tree is shown in Fig. 1*D*). Amino acid sequence comparison of this conserved marsupial envelope protein with HEMO indicates only 20–30% similarity, but alignment of simian, cat, and marsupial (from opossum, wallaby, and Tasmanian devil) sequences (Fig. S4*A*) shows significant identity regions, all along the extracellular domains. The Env-panMars sequences correspond to truncated *env* because of a stop codon upstream of the transmembrane domain. The encoded proteins are, therefore, expected to be soluble proteins. As illustrated in Fig. S4*B* with HA-tagged Env-panMars proteins, the opossum and wallaby Env proteins are indeed released in the supernatant of cells transfected with the corresponding expression vectors. In the supernatant from wallaby transfected cells, a 15-kDa faint band can also be observed, which probably corresponds to the HA-tagged TM subunit produced after partial cleavage at a degenerate furin site (FHKR). No similar band is observed for the opossum [sequence at the furin site (VHKP)].

Furthermore, RACE-PCR experiments performed on wallaby RNA transcripts from ovary (Fig. S4*C*) locate the transcription start site within a CpG-rich region, with multiply-spliced RNAs in the promoter region as observed for the *HEMO* gene. In the case of the opossum, RNA-Seq data compiled in UCSC (Fig. S4*C*) show similar organization (with almost identical transcription start site located in a homologous CpG island and the use of the same E3 exon). Altogether, these data suggest that both the simian and marsupial *env* genes have a common retroviral ancestor. However, because of the long evolutionary distance between Marsupials and Eutherian mammals, which results in poor synteny data, no convincing evidence could be further obtained indicating that the marsupial gene would be the ortholog of HEMO (Fig. 9*A*) or of any of the noncoding copy in Table S2.

Discussion

Here, we have identified an endogenous retroviral envelope gene, *HEMO*, with a full-length protein-coding sequence conserved in simians, including humans, and an unprecedented characteristic feature for a retroviral envelope because it is shed and released in the extracellular medium, being found at a detectable level in the blood of pregnant women. Several retroviral envelope gene "captures" have been reported among most mammalian species, and in a number of cases, these genes were shown to be *syncytins* (i.e., genes playing a role in placentation), with the canonical immunosuppressive and fusogenic properties inherited from their ancestral retroviral progenitors being involved in a physiological function of benefit to the host (38) (reviewed in refs. 9 and 39). The presently identified *HEMO* gene shares some of the properties of syncytins but is different, because it is shed in the extracellular environment with no evidence for fusogenic activity. In addition, its pattern of expression is not strictly restricted to the placenta—although it is the organ

where its expression is highest. However, its conservation in evolution with characteristic features of a coopted gene (i.e., evidence for purifying selection) together with the identification of a closely related retroviral *env* gene captured and conserved in the remote marsupial clade (which diverged from eutherian mammals more than 150 Mya) sharing with *HEMO* a CpG-rich promoter and the capacity of its protein product to be released in the extracellular medium [in that case because of a stop codon located just upstream of the transmembrane domain of the TM subunit (12)] constitute a strong hint for a potential physiological role in simians (see below).

The identified retroviral *env* gene belongs to a poorly characterized and moderately reiterated ERV family, namely the MER34 family, with only highly degenerated elements (5, 16, 17). Analysis of the structure of the genomic locus where *HEMO* can be identified only reveals traces of an ancestral provirus, with a highly rearranged gene organization. Of note, an LTR structure is only barely detectable 3′ to *HEMO*, and the 5′ LTR is no longer present. Actually, RACE-PCR analysis of the *HEMO* transcripts reveals a transcription start site within a CpG-rich domain unrelated to an LTR but clearly possessing a promoter activity as shown by transfection of reporter plasmids—with the promoter in both orientations—in cells in culture. This unusual promoter is most probably responsible for the specific pattern of expression of the *HEMO* gene, which is found to be active in a series of stem cells ex vivo as well as in vivo very early in the developing embryo. The encoded protein itself has some unusual features, because it no longer possesses a furin cleavage site (although a functional one can still be shown for the *HEMO* ortholog present within the NWM genome) and more importantly, because it is specifically cleaved at the cell membrane via a metalloproteinase-mediated processing that results in the shedding of its ectodomain into the extracellular medium —observed for all simians, including NWMs. Shedding is a process that has not been reported previously for a retroviral envelope, although such a process is used by the cellular machinery for a series of cellular genes (e.g., Notch, TNF-alpha) involved, for instance, in signaling, cell mobility, and migration (reviewed in refs. 21, 22, and 40). Of note, a closely related molecular event also takes place in the case of the Ebola filovirus envelope protein, which is, in part, shed in the cell medium by a specific ADAM-mediated cleavage upstream of the transmembrane domain (20, 41). In that case as well, the shed protein is detected in the blood and anticipated to play a critical role in the associated pathology either by exerting a decoy effect on anti-Env antibodies or even through direct immune activation and increased vascular permeability in the infected individuals (42). The presently observed shedding of the HEMO retroviral envelope protein de facto makes a link between unrelated viruses (e.g., a filovirus and a retrovirus), with a possible convergent evolution for the triggering of a systemic effect via a shedding process.

A still unresolved question concerns the possible role of HEMO in human physiology and/or pathology. Because of (*i*) the high level of purifying selection acting on the gene in simians; (*ii*) the conservation in marsupials of a gene transcribed from a similar promoter type and encoding a protein closely related in both sequence and mature protein extracellular localization (both proteins are released in the supernatant by shedding for HEMO and because of absence of a transmembrane-anchoring domain for the marsupial protein); (*iii*) the rather uncommon profile of expression in development; and (*iv*) the massive shedding by the placenta of the protein into the blood, it can be anticipated that HEMO fulfils a role, most probably in pregnancy. Among the possible roles that can be hypothesized, a protective effect against infection by—still to be identified—viruses and/or retroviruses would be relevant. Such protective effects could be mediated by classical "interference" via the sequestration of the receptor for the incoming virus, an effect that could be further enhanced by the release of the HEMO protein in the blood circulation and direct targeting of such receptors (reviewed in refs. 43 and 44). Alternately, HEMO might possess a cytokine- or hormone-like activity, with a possible role in pregnancy still to be uncovered. An effect of *HEMO* in development should also be considered, taking into consideration that its expression is observed as early as at the eight-cell stage and persists at all of the subsequent embryonic stages. Of note, other ERVs—including HERV-H and HERV-K—have related profiles of expression, and abundant HERV-H RNA was recently shown to be a marker of cell "stemness" in humans and possibly play a role—via transcriptional effects and/or specific ERV-driven transcripts—in the maintenance of pluripotency in human stem cells (45–49) (reviewed in refs. 50–52). In the case of *HEMO*, which unambiguously encodes a retroviral envelope protein that can further be detected, its expression might not only be a stemness marker as for the above multicopy ERVs, but its encoded protein might also constitute—as the OCT4, SOX2, KLF4, or MYC "reprogramming" factors (53)—a molecular effector of pluripotency per se. Finally, we could unravel *HEMO* gene expression in a series of human tumors and show HEMO protein expression in ovarian tumors. Additional immunological analyses based on a large number of tumors and control tissues will have to be performed to definitely correlate HEMO protein expression with specific tumor histotypes (54, 55) (reviewed in ref. 56 for other retroviral Env expressed in tumors) and assess whether this protein can be considered as a reliable marker of a given tumoral state and tentatively, a possible target for immunotherapeutic approaches.

Experiments are now in progress to identify the cellular interacting partners of the HEMO protein, with the hope that their identification will allow a definite characterization of HEMO functions in vivo in both normal development and the onset of pathological processes.

Methods

Biological Samples. First trimester human placenta tissues were obtained from legal elective terminations of pregnancy (gestational age 8–12 wk) with parent's written informed consent from the Department of Obstetrics and Gynecology at the Cochin Hospital. Blood samples from pregnant (11–18 wk of amenorrhea) and nonpregnant (before ovulation induction hormonal therapy) women were from Labo Eylau with MTA protocol MTA2015-45. Male blood samples were from the Etablissement Français du Sang with agreement 15EFS018.

Ovary tissue samples were from the Biological Resources Centre and the Department of Laboratory Medicine and Pathology of the Gustave Roussy Institute (Research Agreement RT09916).

RNAs from human ESCs (H1, H7, and H9) were from U1170-INSERM of the Gustave Roussy Institute. iPSCs (reprogrammed CD34+ human cells at passage 24) and their supernatants were from the iPSC Platform of the Gustave Roussy Institute. The source of nonhuman primate genomic DNA is in ref. 57 [except for gibbon (from the European Collection of Authenticated Cell Cultures) and spider monkey (from Coriell Institute)], and the source of wallaby RNA is in ref. 12.

Ethics Statement. All human samples were obtained with written informed consent. Experiments were approved by the Ethics Committee of the Gustave Roussy Institute. This study was carried out in strict accordance with the French and European laws and regulations regarding Animal Experimentation (Directive 86/609/EEC regarding the protection of animals used for experimental and other scientific purposes).

Other methods are as in ref. 12, and they are detailed in *SI Methods*.

ACKNOWLEDGMENTS. We thank David Cornu and Laïla Sago (Paris-Saclay Proteomic Platform, Institut Biologie Integrative Cellule, Gif sur Yvette) and Emilie Cochet and Vasily Ogryzko (Proteomic Platform of the Gustave Roussy Institute) for MS analyses; Larissa Lordier (Gustave Roussy iPSC Platform) for the gift of iPSCs and supernatant; Antonio di Stefano and Nathalie Balayn (UMR 1170, Gustave Roussy) for the gift of ESCs RNAs; and Mélanie Polrot (Gustave Roussy Animal Care Facilities) for assistance in mice handling. We also thank Jean-Michel Teboul (Port-Royal Maternity Hospital) for providing placental tissues; Jean-Yves Scoazec, Catherine Genestie, and Christine Machavoine for the gift of tumor tissues (Gustave Roussy, Centre de Ressources Biologiques); and Martine Bacry (Laboratoire Eylau) for serum samples. We thank Philippe Dessen, Guillaume Meurice, and Bastien Job (Bioinformatic Platform, Gustave Roussy) for helpful bioinformatics assistance and discussion; Anne Dupressoir and Jérome Salmon for fruitful discussions; and Christian Lavialle for critical reading of the manuscript. This work was supported by the CNRS and grants to T.H. from the Ligue Nationale contre Le Cancer and the Agence Nationale de la Recherche (ANR "RETRO-PLACENTA").

1. Feschotte C, Gilbert C (2012) Endogenous viruses: Insights into viral evolution and impact on host biology. *Nat Rev Genet* 13:283–296.
2. Dewannieux M, Heidmann T (2013) Endogenous retroviruses: Acquisition, amplification and taming of genome invaders. *Curr Opin Virol* 3:646–656.
3. Mager DL, Stoye JP (2015) Mammalian endogenous retroviruses. *Microbiol Spectr* 3:MDNA3-0009-2014.
4. de Parseval N, Heidmann T (2005) Human endogenous retroviruses: From infectious elements to human genes. *Cytogenet Genome Res* 110:318–332.
5. Vargiu L, et al. (2016) Classification and characterization of human endogenous retroviruses; mosaic forms are common. *Retrovirology* 13:7.
6. Mi S, et al. (2000) Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. *Nature* 403:785–789.
7. Blond JL, et al. (2000) 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.
8. Blaise S, de Parseval N, Bénit L, Heidmann T (2003) Genomewide screening for fusogenic human endogenous retrovirus envelopes identifies syncytin 2, a gene conserved on primate evolution. *Proc Natl Acad Sci USA* 100:13013–13018.
9. Lavialle C, et al. (2013) Paleovirology of 'syncytins', retroviral env genes exapted for a role in placentation. *Philos Trans R Soc Lond B Biol Sci* 368:20120507.
10. Dupressoir A, et al. (2009) Syncytin-A knockout mice demonstrate the critical role in placentation of a fusogenic, endogenous retrovirus-derived, envelope gene. *Proc Natl Acad Sci USA* 106:12127–12132.
11. Dupressoir A, et al. (2011) A pair of co-opted retroviral envelope syncytin genes is required for formation of the two-layered murine placental syncytiotrophoblast. *Proc Natl Acad Sci USA* 108:E1164–E1173.
12. Cornelis G, et al. (2015) Retroviral envelope gene captures and syncytin exaptation for placentation in marsupials. *Proc Natl Acad Sci USA* 112:E487–E496.
13. de Parseval N, Lazar V, Casella JF, Benit L, Heidmann T (2003) 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.
14. Villesen P, Aagaard L, Wiuf C, Pedersen FS (2004) Identification of endogenous retroviral reading frames in the human genome. *Retrovirology* 1:32.
15. Henzy JE, Johnson WE (2013) Pushing the endogenous envelope. *Philos Trans R Soc Lond B Biol Sci* 368:20120506.
16. Tóth G, Jurka J (1994) Repetitive DNA in and around translocation breakpoints of the Philadelphia chromosome. *Gene* 140:285–288.
17. Jurka J, et al. (2005) Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* 110:462–467.
18. Smale ST, Kadonaga JT (2003) The RNA polymerase II core promoter. *Annu Rev Biochem* 72:449–479.
19. Deaton AM, Bird A (2011) CpG islands and the regulation of transcription. *Genes Dev* 25:1010–1022.
20. Dolnik O, et al. (2004) Ectodomain shedding of the glycoprotein GP of Ebola virus. *EMBO J* 23:2175–2184.
21. Okazaki I, Nabeshima K (2012) Introduction: MMPs, ADAMs/ADAMTSs research products to achieve big dream. *Anticancer Agents Med Chem* 12:688–706.
22. Weber S, Saftig P (2012) Ectodomain shedding and ADAMs in development. *Development* 139:3693–3709.
23. Pollheimer J, Fock V, Knöfler M (2014) Review: The ADAM metalloproteinases - novel regulators of trophoblast invasion? *Placenta* 35:S57–S63.
24. Aghababaei M, Beristain AG (2015) The Elsevier Trophoblast Research Award Lecture: Importance of metzincin proteases in trophoblast biology and placental development: A focus on ADAM12. *Placenta* 36:S11–S19.
25. Majali-Martinez A, et al. (2016) Placental membrane-type metalloproteinases (MT-MMPs): Key players in pregnancy. *Cell Adhes Migr* 10:136–146.
26. Cole LA (2009) New discoveries on the biology and detection of human chorionic gonadotropin. *Reprod Biol Endocrinol* 7:8.
27. Bischof P, Irminger-Finger I (2005) The human cytotrophoblastic cell, a mononuclear chameleon. *Int J Biochem Cell Biol* 37:1–16.
28. Maltepe E, Fisher SJ (2015) Placenta: The forgotten organ. *Annu Rev Cell Dev Biol* 31: 523–552.
29. Uhlén M, et al. (2015) Proteomics. Tissue-based map of the human proteome. *Science* 347:1260419.
30. Yan L, et al. (2013) Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. *Nat Struct Mol Biol* 20:1131–1139.
31. Xue Z, et al. (2013) Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. *Nature* 500:593–597.
32. Friedli M, et al. (2014) Loss of transcriptional control over endogenous retroelements during reprogramming to pluripotency. *Genome Res* 24:1251–1259.
33. Soygur B, Moore H (2016) Expression of Syncytin 1 (HERV-W), in the preimplantation human blastocyst, embryonic stem cells and trophoblast cells derived *in vitro*. *Hum Reprod* 31:1455–1461.
34. Lukk M, et al. (2010) A global map of human gene expression. *Nat Biotechnol* 28: 322–324.
35. Cho KR, Shih IeM (2009) Ovarian cancer. *Annu Rev Pathol* 4:287–313.
36. Kurman RJ, Shih IeM (2016) The dualistic model of ovarian carcinogenesis: Revisited, revised, and expanded. *Am J Pathol* 186:733–747.
37. Meredith RW, et al. (2011) Impacts of the Cretaceous Terrestrial Revolution and KPg extinction on mammal diversification. *Science* 334:521–524.
38. Mangeney M, et al. (2007) Placental syncytins: Genetic disjunction between the fusogenic and immunosuppressive activity of retroviral envelope proteins. *Proc Natl Acad Sci USA* 104:20534–20539.
39. Denner J (2016) Expression and function of endogenous retroviruses in the placenta. *APMIS* 124:31–43.
40. Jones JC, Rustagi S, Dempsey PJ (2016) ADAM Proteases and Gastrointestinal Function. *Annu Rev Physiol* 78:243–276.
41. Dolnik O, et al. (2015) Shedding of ebola virus surface glycoprotein is a mechanism of self-regulation of cellular cytotoxicity and has a direct effect on virus infectivity. *J Infect Dis* 212:S322–S328.
42. Escudero-Pérez B, Volchkova VA, Dolnik O, Lawrence P, Volchkov VE (2014) Shed GP of Ebola virus triggers immune activation and increased vascular permeability. *PLoS Pathog* 10:e1004509.
43. Nethe M, Berkhout B, van der Kuyl AC (2005) Retroviral superinfection resistance. *Retrovirology* 2:52.
44. Malfavon-Borja R, Feschotte C (2015) Fighting fire with fire: Endogenous retrovirus envelopes as restriction factors. *J Virol* 89:4047–4050.
45. Macfarlan TS, et al. (2012) Embryonic stem cell potency fluctuates with endogenous retrovirus activity. *Nature* 487:57–63.
46. Santoni FA, Guerra J, Luban J (2012) HERV-H RNA is abundant in human embryonic stem cells and a precise marker for pluripotency. *Retrovirology* 9:111.
47. Fort A, et al.; FANTOM Consortium (2014) Deep transcriptome profiling of mammalian stem cells supports a regulatory role for retrotransposons in pluripotency maintenance. *Nat Genet* 46:558–566.
48. Grow EJ, et al. (2015) Intrinsic retroviral reactivation in human preimplantation embryos and pluripotent cells. *Nature* 522:221–225.
49. Göke J, et al. (2015) Dynamic transcription of distinct classes of endogenous retroviral elements marks specific populations of early human embryonic cells. *Cell Stem Cell* 16: 135–141.
50. Schlesinger S, Goff SP (2015) Retroviral transcriptional regulation and embryonic stem cells: War and peace. *Mol Cell Biol* 35:770–777.
51. Friedli M, Trono D (2015) The developmental control of transposable elements and the evolution of higher species. *Annu Rev Cell Dev Biol* 31:429–451.
52. Izsvák Z, Wang J, Singh M, Mager DL, Hurst LD (2016) Pluripotency and the endogenous retrovirus HERVH: Conflict or serendipity? *BioEssays* 38:109–117.
53. Shi Y, Inoue H, Wu JC, Yamanaka S (2017) Induced pluripotent stem cell technology: A decade of progress. *Nat Rev Drug Discov* 16:115–130.
54. Wang-Johanning F, et al. (2007) Expression of multiple human endogenous retrovirus surface envelope proteins in ovarian cancer. *Int J Cancer* 120:81–90.
55. Wang-Johanning F, et al. (2014) Human endogenous retrovirus type K antibodies and mRNA as serum biomarkers of early-stage breast cancer. *Int J Cancer* 134:587–595.
56. Strissel PL, et al. (2012) Reactivation of codogenic endogenous retroviral (ERV) envelope genes in human endometrial carcinoma and prestages: Emergence of new molecular targets. *Oncotarget* 3:1204–1219.
57. Esnault C, Cornelis G, Heidmann O, Heidmann T (2013) Differential evolutionary fate of an ancestral primate endogenous retrovirus envelope gene, the EnvV syncytin, captured for a function in placentation. *PLoS Genet* 9:e1003400.
58. Malassiné A, et al. (2007) Expression of the fusogenic HERV-FRD Env glycoprotein (syncytin 2) in human placenta is restricted to villous cytotrophoblastic cells. *Placenta* 28:185–191.
59. Malassiné A, et al. (2008) Human endogenous retrovirus-FRD envelope protein (syncytin 2) expression in normal and trisomy 21-affected placenta. *Retrovirology* 5:6.
60. Schwartz S, et al. (2003) Human-mouse alignments with BLASTZ. *Genome Res* 13: 103–107.