Origins and evolutionary consequences of ancient endogenous retroviruses

Welkin E. Johnson

Abstract | Retroviruses infect a broad range of vertebrate hosts that includes amphibians, reptiles, fish, birds and mammals. In addition, a typical vertebrate genome contains thousands of loci composed of ancient retroviral sequences known as endogenous retroviruses (ERVs). ERVs are molecular remnants of ancient retroviruses and proof that the ongoing relationship between retroviruses and their vertebrate hosts began hundreds of millions of years ago. The long-term impact of retroviruses on vertebrate evolution is twofold: first, as with other viruses, retroviruses act as agents of selection, driving the evolution of host genes that block viral infection or that mitigate pathogenesis, and second, through the phenomenon of endogenization, retroviruses contribute an abundance of genetic novelty to host genomes, including unique protein-coding genes and *cis*-acting regulatory elements. This Review describes ERV origins, their diversity and their relationships to retroviruses and discusses the potential for ERVs to reveal virus–host interactions on evolutionary timescales. It also describes some of the many examples of cellular functions, including protein-coding genes and regulatory elements, that have evolved from ERVs.

Endogenous retrovirus
(ERV). Heritable retrovirus-derived sequence elements found in the genomes of most or all vertebrates; ERVs usually originate as proviruses integrated into germline DNA.

Loss
Refers to the case when an allelic variant of a locus disappears from the population over time.

Fixation
Refers to the case in which an allelic variant of a locus achieves a frequency of 100% in the population, thereby displacing all other alleles at that locus.

Random genetic drift
Refers to the change in frequency of an allele over time owing to random chance (in the absence of selection).

Biology Department, Boston College, Chestnut Hill, MA, USA.

e-mail: welkin.johnson@bc.edu

https://doi.org/10.1038/s41579-019-0189-2

Retrovirus virions contain RNA copies of the viral genome. Upon entry into a target cell, these are reverse transcribed into a double-stranded DNA molecule and integrated into the genomic DNA of the host cell. The resulting provirus contains the promoters and regulatory elements required for transcription of viral RNA and encodes all the structural proteins and enzymes necessary for assembling progeny virions. Retroviruses typically infect somatic tissues; however, as a retrovirus spreads in a host population, there is an unknown but finite probability that integration may occur in germline cells or in the precursors of germline cells, resulting in production of host gametes carrying proviruses as novel insertions. Upon entering the host gene pool in this way, a provirus is known as an endogenous retrovirus (ERV) and is fated for either loss or fixation depending on the vagaries of random genetic drift and natural selection (FIG. 1). An ERV may also increase in copy number by various post-endogenization mechanisms. Thus, ERVs are genetic loci whose ultimate origins trace back to exogenously replicating retroviruses, regardless of whether they retain the capacity to express infectious virions. Indeed, the vast majority of ERVs are defective for viral gene expression as a consequence of mutations accumulated across thousands to millions of years of vertebrate evolution.

Endogenization is not an essential property of any known retrovirus, and germline insertion is probably very rare relative to infection of somatic tissues. Importantly, the ability to replicate and spread in germline cells is not a prerequisite for endogenization. Only the early stages of the retroviral life cycle (entry, reverse transcription and integration) are necessary for provirus biogenesis, and all viral components essential for completing these steps are provided by the incoming virion — neither de novo viral genome synthesis nor expression of viral genes is required to produce an integrated provirus. Nonetheless, over the span of millions of years, the genomes of vertebrates have accumulated thousands and, in some cases, hundreds of thousands of ERV loci. This vast molecular archive of ancient, extinct retroviruses has captured the attention of virologists and evolutionary biologists interested in the impact of viruses on the evolution of their vertebrate hosts[1–7]. In addition, because they are found in virtually all vertebrate genomes, ERVs may be expressed in many commonly used cell lines, tissues and model organisms, potentially compromising interpretation of experimental results, contaminating preparations of biological and pharmacological reagents and vaccines[8,9], complicating the use of animal organs for xenotransplantation[10] and, perhaps, contributing to human disease[11,12]. Moreover, ERV expression can be induced by a variety of conditions, including infection with viruses such as HIV or exposure to epigenetic modifying drugs[13], and studies in cell culture and laboratory mice have documented the potential for recombination, either between ERVs or between ERVs and exogenous retroviruses, to produce viral strains with novel biological and pathogenic properties[14–17]. This Review describes ERVs

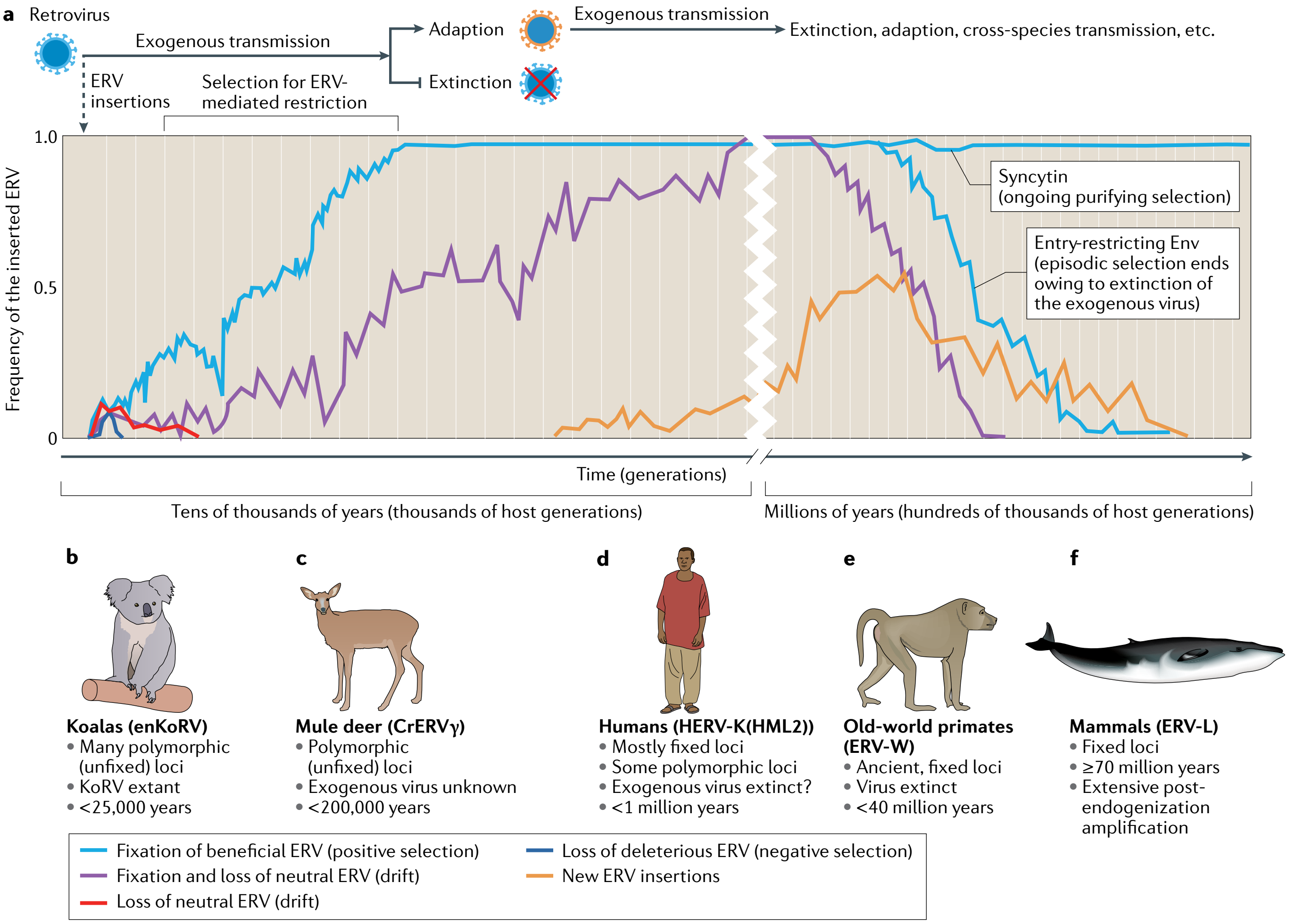

a
Retrovirus
Exogenous transmission
Adaption
Extinction
Exogenous transmission
Extinction, adaption, cross-species transmission, etc.
ERV insertions
Selection for ERV-mediated restriction
1.0
0.5
0
Frequency of the inserted ERV
Syncytin (ongoing purifying selection)
Entry-restricting Env (episodic selection ends owing to extinction of the exogenous virus)
Time (generations)
Tens of thousands of years (thousands of host generations)
Millions of years (hundreds of thousands of host generations)
b
c
d
e
f
Koalas (enKoRV)
• Many polymorphic (unfixed) loci
• KoRV extant
• <25,000 years
Mule deer (CrERVγ)
• Polymorphic (unfixed) loci
• Exogenous virus unknown
• <200,000 years
Humans (HERV-K(HML2))
• Mostly fixed loci
• Some polymorphic loci
• Exogenous virus extinct?
• <1 million years
Old-world primates (ERV-W)
• Ancient, fixed loci
• Virus extinct
• <40 million years
Mammals (ERV-L)
• Fixed loci
• ≥70 million years
• Extensive post-endogenization amplification
Fixation of beneficial ERV (positive selection)
Fixation and loss of neutral ERV (drift)
Loss of neutral ERV (drift)
Loss of deleterious ERV (negative selection)
New ERV insertions

Fig. 1 | **Random genetic drift, natural selection and the early stages of endogenous retrovirus evolution in a host population.** Hypothetical evolutionary stages of endogenous retrovirus (ERV) loci, depicted as changes in allele frequencies, are shown (part **a**). At the time of insertion, the exogenous retrovirus is still extant and continuing to spread in the host population (virion and arrows at the top). The graph includes the change in frequencies of four different ERV insertions in a hypothetical host population consisting of ~500–1,000 breeding individuals and a generation time of ~10–20 years. Each locus begins with two alleles: the major (high frequency) allele being the uninterrupted chromosomal site (not shown) and the minor (low frequency) allele being the same chromosomal site containing the ERV insertion (shown as coloured lines). In the example, the four ERV insertions are already in the population with frequencies <10%. ERV insertions that have strong negative effects on host fitness are unlikely to have persisted in the population and are not shown. An ERV that is only mildly deleterious may initially increase in frequency by chance but is subject to negative selection and is most likely to be lost within a few generations (dark blue line). A neutral ERV is most likely to be lost by drift (red line), although there is a finite probability that a neutral ERV will instead drift to fixation (purple line), replacing the uninterrupted chromosomal site as the major allele. An ERV that confers a strong selective advantage (light blue line) may increase in frequency and achieve fixation more rapidly than a neutral ERV. If the selectively advantageous ERV encodes an essential, developmental function such as a syncytin (see the main text), the locus will be preserved by long-term purifying selection (upper light blue line). Alternatively, an ERV that encodes a restriction that inhibits infection by the corresponding exogenous virus may contribute to the extinction of the virus, after which it is no longer subject to selection and decreases in frequency as it is replaced by inactive or defective alleles (lower light blue line). Alternatively, the viral lineage may adapt through receptor switching, increasing the number of genetically susceptible individuals and allowing new invasions of the germ line (orange lines). It is noteworthy that ERV sequences may persist in animal genomes for millions of years beyond the extinction of the original exogenous retrovirus. In this regard, ERVs are often described as molecular ‘fossils’ left by ancient viruses. The identification and study of natural outbred populations at different stages of endogenization could help to illuminate details of the endogenization process and its impact on host evolution. Shown are examples of ERVs at different stages in the evolutionary process in natural populations; these correspond roughly to the graph in the upper panel. Australian koalas (part **b**) harbour an actively spreading gammaretrovirus (koala retrovirus; KoRV) and have large numbers of unfixed KoRV-related ERV (enKoRV), suggesting that the virus may still be actively invading the germ line in these animals. North American mule deer (part **c**) harbour multiple copies of an endogenous gammaretrovirus (cervid endogenous gammaretrovirus; CrERVγ), estimated to have inserted in the germ line within the past 200,000 years. Many of the CrERVγ loci are still polymorphic (unfixed), consistent with a relatively recent endogenization event. HERV-K(HML2) elements (part **d**) in the human genome include a majority fixed ERV but also a significant minority of unfixed ERV, some with intact ORFs and identical LTRs, suggestive of evolutionarily recent genome invasion in modern humans. Thus far, exogenous forms of human endogenous retrovirus K HML-2 (HERV-K(HML-2)) have not been reported, and the virus may have gone extinct. The family of related ERV known as ERV-W includes a large number of homologous ERV loci shared by multiple species, indicating that endogenization began long ago in the ancestor or ancestors of modern old-world primates (part **e**). The ERV-L family elements (part **f**) began invading the mammalian lineage germ line over 70 million years ago and have subsequently undergone post-endogenization amplification in various lineages, including those leading to mice and to humans.

Long-terminal repeats (LTRs). Direct identical repeats found at the 5′ and 3′ ends of a DNA provirus generated during reverse transcription of the retroviral RNA genome.

and their relationship to exogenous retroviruses, highlights the ways in which ERVs aid our understanding of the origins and evolution of retroviruses, discusses advances in the reconstitution and functional characterization of ancient ERV genes and provides a virological perspective on the contributions of ERVs to cellular functions.

Diversity of endogenous retroviruses

All retroviruses have a similar genome structure (FIG. 2). Reverse transcription and integration result in a provirus of approximately 5–10 kb, comprising identical long-terminal repeats (LTRs) with the viral genes arrayed between them. LTRs contain the primary promoter and regulatory elements for provirus expression, as

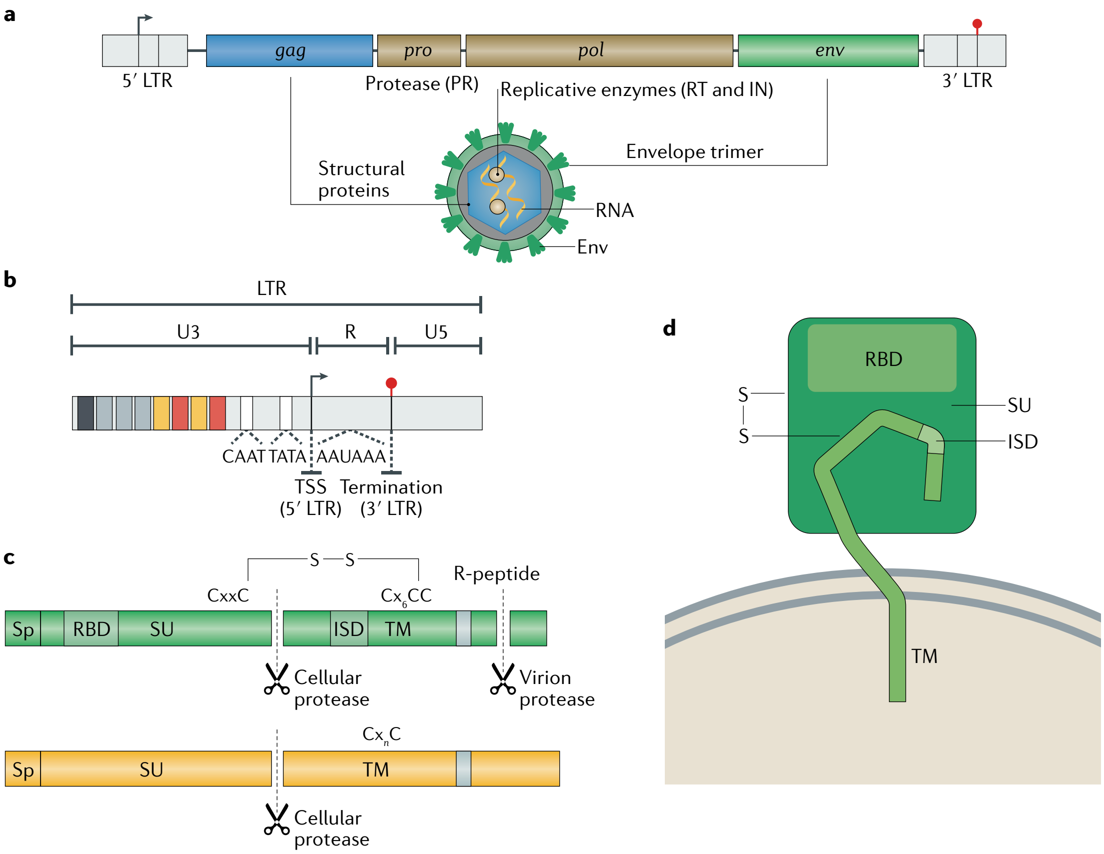


Fig. 2 | **Features of a typical DNA provirus. a** | A typical provirus consists of two identical long-terminal repeats (LTRs) bracketing the four canonical viral genes, *gag*, *pro*, *pol* and *env*. These genes encode the structural proteins that make up the viral capsid core, the virion protease, the replicative enzymes and the Env glycoprotein, respectively. The number and location of accessory genes vary between different genera and even species of retrovirus (not depicted). The start and stop sites for full-length and spliced viral mRNAs are shown as a small arrow in the 5′ LTR and a red marker in the 3′ LTR, respectively. **b** | The LTRs comprise many of the *cis*-acting regulatory elements that control proviral gene expression (coloured boxes). These include the core promoter and transcription-factor-binding sites, for example, the CAAT box and TATA box, enhancers, repressors and polyadenylation signals. LTRs are divided into three segments — U3, R and U5. The R (repeat) segments comprise the very 5′ and 3′ ends of the RNA genome; U5 and U3 are present in single copy in the RNA genome but are duplicated during reverse transcription such that both LTRs have copies. U3 typically contains the core promoter elements as well as many of the transcription-factor-binding sites. The U3–R junction in the 5′ LTR demarcates the transcription start site (TSS, small arrow) of viral RNA genomes and mRNAs. In the 3′ LTR, the R–U5 junction marks the termination of the viral RNA (red marker). **c** | A majority of retroviral Env glycoproteins exist as one of two types[33], which are distinguished by the presence (gamma-type) or absence (beta-type) of an intersubunit disulfide bond that covalently links the surface unit (SU) and transmembrane (TM) domains. The presence of the bond can be predicted on the basis of the presence of the appropriate CxxC and CX$_n$CC motifs in the SU and TM domains, respectively. By contrast, beta-type Envs lack the CxxC motif in the SU domain and have a CX$_n$C motif in the TM domain. Gamma-type Envs are also distinguished by a highly conserved classical immunosuppressive domain (ISD) in the TM domain and a modular domain arrangement with the receptor-binding domain (RBD) mostly confined to the amino-terminal portion of the SU domain. The receptor-binding determinants of a beta-type Env often involve discontinuous elements spread throughout the primary amino acid sequence. Finally, gamma-type Envs are also known to have a carboxy-terminal R peptide that must be cleaved during virion maturation to activate the fusion capacity of the Env complexes on virions. **d** | The cartoon depicts the arrangement of the SU and TM domains of a gamma-type Env. Env spikes comprise trimers of three SU–TM multimers (only one is shown). The SU domain contains the RBD and partially covers the metastable TM domain. The TM domain spans the membrane (double grey line) and anchors the entire complex in the surface of the virion or producer cell. Beta-type Env spikes have a similar arrangement but lack the intersubunit disulfide bond (S–S). IN, integrase; RT, reverse transcriptase.

well as the *cis*-acting motifs required for integration. A common set of genes includes *gag*, which encodes the structural proteins that make up the virion core; *pro*, which encodes the viral protease; *pol*, which encodes the viral replicative enzymes reverse transcriptase (RT) and integrase (IN); and *env*, which encodes the glycoprotein complex that governs receptor-mediated fusion and entry. Retroviruses vary considerably in the number and genomic position of noncanonical accessory genes.

LTRs consist of three regions: from 5′ to 3′, these are U3, R and U5 (FIG. 2b). R is repeated at both ends of the viral RNA, whereas U5 and U3 are present as one copy each. The process of reverse transcription duplicates U5 and U3 to produce identical LTRs at both ends of the DNA provirus. The U3–R junction corresponds to the transcription start site (TSS) in the 5′ LTR, whereas the R–U5 junction corresponds to the 3′ end of the proviral transcripts in the 3′ LTR. U3 contains various motifs that interact with the regulatory milieu of the host cell and governs provirus expression; its length varies between different retroviruses (~190–1,200 bases) and comprises a dense and highly variable cluster of enhancer and promoter elements[18,19]. The variations in U3 of different retroviruses reflect differences in cellular or tissue tropism and host range.

ERVs originate as integrated proviruses and can range from complete proviruses to highly fragmented remnants of proviruses. Even where substantial portions of *gag*, *pro*, *pol* and *env* remain, these are often inactive owing to the accumulation of substitutions, deletions and insertions. The degree of sequence degradation correlates approximately with the age of the provirus (that is, the amount of time that has passed since germline insertion). A majority of ERVs exist as solo-LTRs produced by homologous recombination between the 5′ and 3′ LTRs. Solo-LTR formation deletes all internal sequences, including the viral genes[20]. LTRs are the most variable sequences in the retroviral genome, and there is little or no resemblance between the LTRs of retroviruses from different genera[18,19]. Consequently, annotating solo-LTRs in genome assemblies often depends on an association with a known retrovirus or previously characterized ERV, although query-independent identification of LTRs has been reported[18,19].

The presence of ERV sequences in genomic DNA was first confirmed more than 50 years ago[21]. In the years that followed, ERV loci were detected and characterized first by hybridization methods and later by cloning or PCR and were found in the genomes of a wide range of vertebrate species. Whole-genome sequencing and related computational tools accelerated the discovery and phylogenetic analysis of ERV loci, permitting detailed comparisons to extant retroviruses. ERV loci within a genome can be clustered into groups of related elements on the basis of sequence[5,22,23]. These groups may reflect multiple germline insertions by the same species of retrovirus but can also result from different post-endogenization amplification mechanisms[24,25]. These include activation and expression of an ERV locus resulting in particles that reinfect germline cells and insert new copies of the element; infection or retrotransposition in *trans*, whereby ERV transcripts are packaged, copied and integrated by another virus or transposable element; or expansions of chromosomal DNA segments that contain ERVs (for example, segmental duplications).

Reverse transcriptase amino acid sequences are highly conserved and readily aligned across the entire taxonomic range of known reverse-transcribing viruses and retrotransposons and are useful for reconstructing deep phylogenetic relationships. ERVs are easily incorporated into such analyses, either directly or after in silico reconstruction of RT-coding sequences. RT-based phylogenies of the family *Retroviridae* contain three major branches, and taxa comprising these branches are sometimes referred to as class I, II or III[5,23,26]. As more vertebrate genomes are assembled, incorporating larger numbers of ERVs has not drastically changed the overall topology — most retroviral and ERV RT sequences analysed to date cluster within the three main branches[5,27]. Retrovirus phylogenies can also be based on the conserved ectodomain of the transmembrane (TM) subunit of the viral envelope glycoprotein (Env), and discrepancies between RT and TM phylogenies can reveal lineages that originated by recombination between distantly related retroviruses[28]. The number of unique ERV lineages extracted from genome data now exceeds the number of distinct retroviruses that have been classified by the International Committee on Taxonomy of Viruses[29]. Incorporating these lineages into retroviral taxonomy will likely require creating additional genera within *Retroviridae*, some of which may consist mostly or exclusively of extinct retrovirus species[30].

ERVs and retroviruses interleave in RT-based phylogenies, indicating that the phenomenon of endogenization is not unique to any particular type of retrovirus. However, there are notable differences in the degree to which different types of ERV are represented in vertebrate genomes. For example, ERVs related to gammaretroviruses are abundant in the genomes of a wide variety of vertebrate species[31–33], whereas ERVs related to deltaretroviruses have only been identified in the genomes of *Miniopterus* and *Rhinolophus* bats[34,35]. ERVs related to lentiviruses are also rare and, thus far, have only been found in the genomes of a small number of mammalian species, none of which is known to host extant lentiviruses[36–42]. Differences in frequency and distribution of different types of ERV have not been explained but may reflect biological differences that influence the probability of endogenization. For example, a retrovirus that can infect germline cells as the result of broad tissue tropism or by virtue of being specifically adapted to germline cells would have a higher probability of producing heritable proviruses. Conversely, ERVs of viruses whose expression is intrinsically cytotoxic might be selected against and less likely to persist in the germ line.

Insights into ancient retroviruses

ERVs can be exploited to study the natural history of viruses and their hosts, revealing the extent to which vertebrate evolution has been impacted by retroviruses and providing insights relevant to the study of modern viruses. For example, a comparison of human endogenous retrovirus K HML-2 (HERV-K(HML-2)) loci found in human, Neanderthal and Denisovan genomes

CAAT box
A *cis*-acting transcription-factor-binding site frequently found upstream of eukaryotic promoters and in retroviral long-terminal repeats.

Accessory genes
Viral genes that are dispensable for the essential steps of the viral replication cycle but that provide one or more functions that contribute to optimal viral fitness in vivo, such as antagonizing intrinsic and innate immune defences or modifying the metabolic state of the host cell.

Solo-LTRs
Solitary long-terminal repeats (LTRs) lacking any other proviral sequence that usually arise by homologous recombination between the 5′ and 3′ LTRs of an ERV locus.

Retrotransposition
The amplification of a genomic DNA sequence by reverse transcription of an RNA intermediate followed by integration of the new DNA copies.

Segmental duplications
Stretches of initially identical or nearly identical genomic sequences that arise by DNA duplication.

reflects the spread of an ancient betaretrovirus among the ancestors of modern humans[43–45], while the discovery of unfixed, largely intact HERV-K(HML2)-related proviruses in gorillas raises the possibility that some populations may still harbour infectious virus[46]. Ancient ERVs in lemur and rabbit genomes are missing links that help clarify the relationship between divergent species of modern lentiviruses[37]. Similarly, 3D structures of ancient lentiviral capsid proteins have been resolved and compared with the corresponding structures of modern relatives, such as HIV-1 (REF.[47]). Ancient spumavirus-related ERVs suggest that retroviruses may have colonized marine animals of the Palaeozoic (more than 450 million years ago)[48], and ERVs have been used to trace the emergence and spread of a gammaretrovirus during the Oligocene[49]. Reconstructed ERV sequences reveal extensive patterns of cross-species transmission of ancient viruses and broaden the known host ranges of modern viral groups[32,49–52]. ERV analysis has also revealed a striking difference in the rates at which viruses evolve over long versus short timescales, a major problem when applying molecular clock calculations to viral taxa[2,53]. Host populations with young (unfixed) ERVs, such as cervid endogenous gammaretrovirus (CrERVγ) elements in mule deer, help shed light on the earliest stages of the endogenization process[54]. This is particularly true for cases in which the related exogenous agent is still extant and potentially pathogenic, such as koala retrovirus (KoRV) in Australian koalas[55].

A major challenge in such studies is accurate reconstruction of ancestral viral sequences from ERV data. Families of related ERV loci are convenient for generating and fine-tuning ancestral sequences by consensus[49] or for inferring ancestral states by phylogenetic analysis[56–58]. ERVs are also uniquely amenable to molecular clock analysis[59,60], which is useful for estimating integration times[61,62] and for dating the emergence and spread of ancient retroviruses[40,42,49,63,64] (BOX 1).

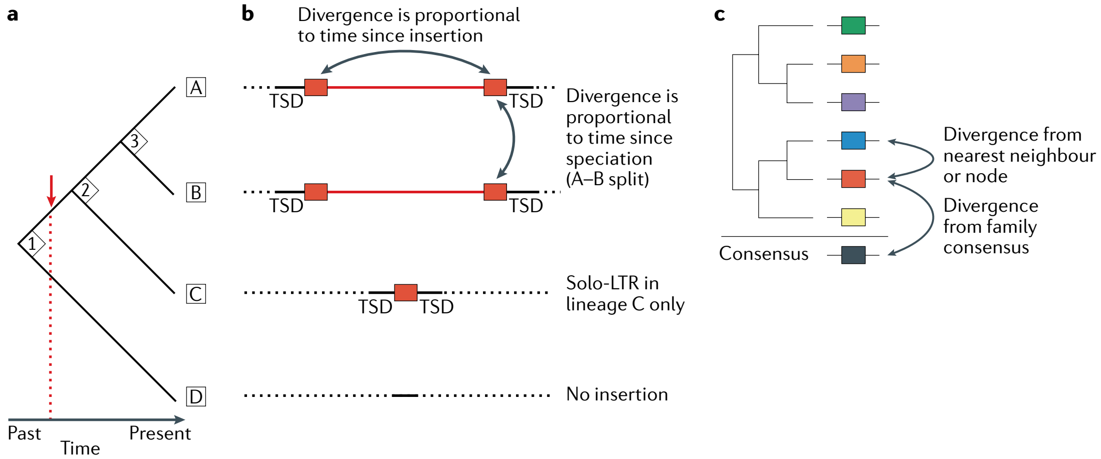
Functional hypotheses can also be tested by biomolecular characterization of reconstituted ERV genes (FIG. 3).

Box 1 | Estimating the ages of endogenous retroviruses and associated ancient retroviruses

Several features of endogenous retroviruses (ERVs) are useful for estimating the ages of ERV families or of individual ERV loci. For example, the distribution of a shared orthologous ERV among the genomes of extant organisms is an indication of its age (see the figure, part **a**). An insertion shared by two or more taxa must have originated in a common ancestor (red arrow and dashed red line), and comparing the distributions of different ERV loci among taxa provides a means for estimating their ages relative to one another. If there is independent evidence for the dates of speciation events (nodes with numbered diamonds), these provide lower bound estimates for the times of ERV insertion. In the example (see the figure, parts **a**,**b**), the solo-long-term repeat (solo-LTR) is found only in taxon C and could have formed any time after the species split at node 2. Loci confined to members of one taxon (not shown) are assumed to reflect insertions that occurred since the most recent common ancestor, although these could also reflect incomplete lineage sorting of insertions that were unfixed at the time of speciation.

It is possible to apply molecular clock analysis by taking advantage of the fact that reverse transcription and integration of the retroviral RNA genome produce a DNA provirus with two identical LTRs, flanked by a short target site duplication (TSD) of 4–6 bp (see the figure, part **b**). In the case of an ERV, the LTR sequences will diverge over time (mostly owing to drift) such that the genetic distance is roughly proportional to age (horizontal curved arrow). If an estimated rate of host sequence evolution (for example, in substitutions per site per year) is available, the 5′ LTR–3′ LTR divergence can also be used to estimate the age of the ERV provirus, for example, in the years before present or millions of years ago[59–61,165]. This also provides a minimum age for the original, horizontally transmitted exogenous retrovirus. As with other types of molecular marker, a molecular clock can also be applied to the interspecies divergence of the shared ERV locus (vertical curved arrow); this gives an estimate of the time of host speciation, which should be less than or equal to the age of the shared ERV.

In cases where both LTRs of an ERV are not available for molecular clock analyses, a variety of other approaches has been used. In the example (see the figure, part **c**), the minimum age of an ERV locus belonging to a multilocus family is estimated by comparing its divergence from either a consensus of closely related family members from the same genome or from a nearest neighbour (the closest related locus selected from among the family members)[62]. Molecular clock calculations can also be applied to ERV loci with structurally distinguishable alleles, for example, when the same locus includes a proviral allele and a solo-LTR allele[37], or to ERV insertions that fall within duplicated segments of a chromosome[36] (not shown).


a
A
B
C
D
1
2
3
Past
Present
Time
b
Divergence is proportional to time since insertion
TSD
TSD
Divergence is proportional to time since speciation (A–B split)
TSD
TSD
TSD
TSD
Solo-LTR in lineage C only
No insertion
c
Consensus
Divergence from nearest neighbour or node
Divergence from family consensus

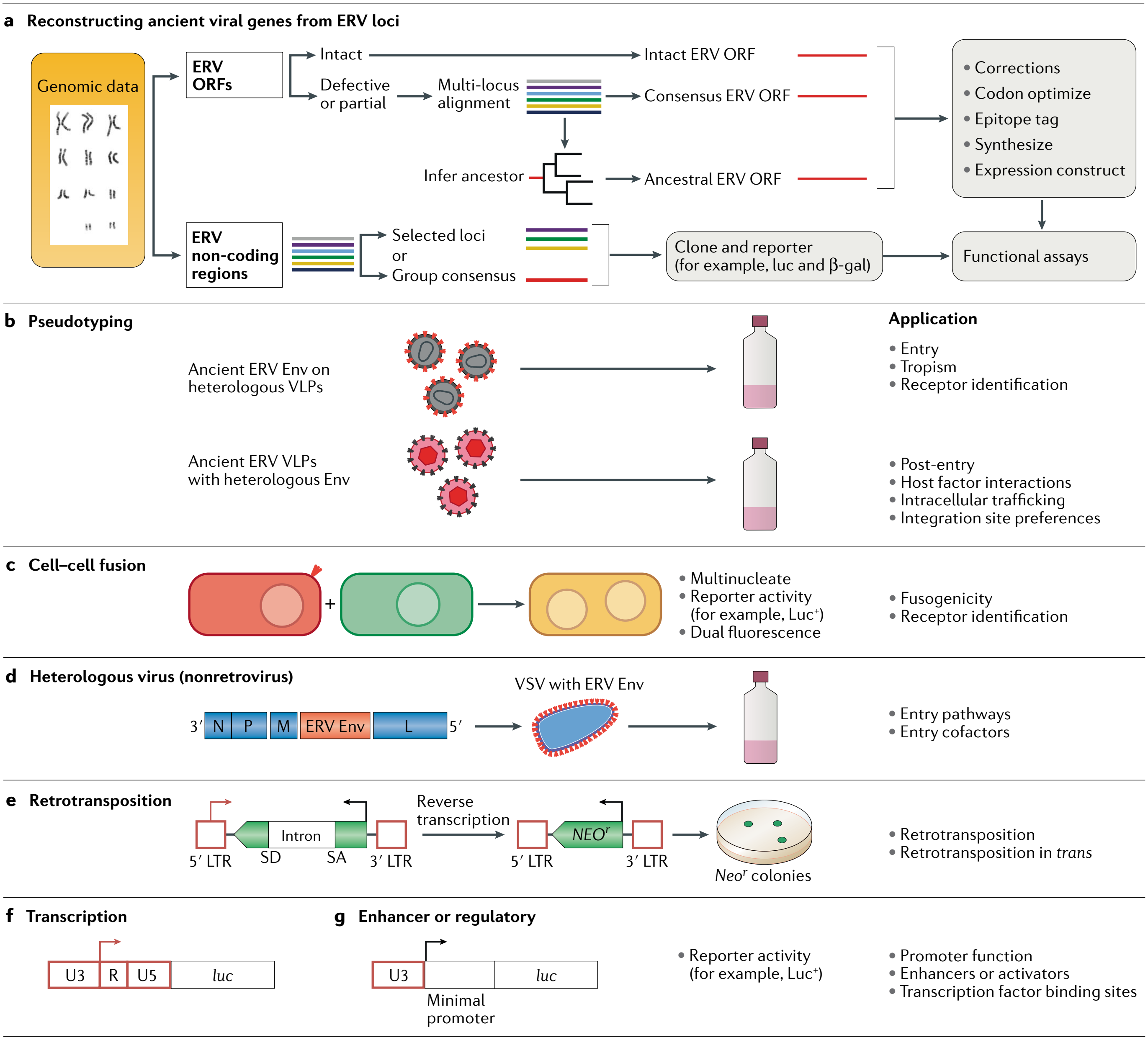


Fig. 3 | **Reconstructing and analysing ancient endogenous retrovirus genes.** **a** | Reconstructing ancient viral genes from endogenous retrovirus (ERV) loci. The first step in functional analysis of endogenous retrovirus genes is accurate reconstitution of the ancestral viral sequence, which must account for bias in the ERV data and for post-endogenization sequence drift. For multilocus families, ERV genes can be reconstituted on the basis of consensus alignments, although greater accuracy may be achieved through ancestral state reconstruction and adjustment for multiple hits, hypermutation and so on. The predicted sequence can then be synthesized and analysed by transfection and protein blots, as well as by capitalizing on a range of biochemical and cellular assays regularly applied to the study of retroviruses. **b** | Pseudotyping. The modularity of retroviral genomes enables the production of infectious virus-like particles (VLPs) by supplying the different virion components encoded on different plasmids. This process is called pseudotyping and is useful for studying viral protein functions under conditions mimicking normal infection and entry. **c** | Cell–cell fusion. Some retroviral envelopes can drive fusion between cells expressing the Env proteins and cells expressing a cognate receptor. The readouts for such assays include visual inspection by microscopy or activation of reporter genes in one cell by a transactivator expressed in the other cell. **d** | Heterologous viruses (non-retroviruses). Remarkably, the gene encoding the entry glycoprotein of the rhabdovirus vesicular stomatitis virus (VSV) can be replaced with retroviral *env* genes to produce infectious VSV with altered tropism according to the receptor specificity of the introduced glycoprotein. The chimeric VSV can then be used for a variety of cell entry assays and as a highly efficient screening tool. **e** | Retrotransposition. A standard but elegant assay for retrotransposition is based on constructing a minimal retroviral genome carrying the necessary *cis*-acting elements and a reporter gene (usually a selectable marker) in antisense orientation (relative to the viral genome) and interrupted by an intron (in the opposite sense orientation). The reporter gene has its own promoter but can only be expressed after a round of transcription, which results in splicing, followed by reverse transcription and integration, which generate a new copy of the provirus in which the reporter gene lacks the intron and can be expressed. The use of a selectable marker permits quantification of retrotransposition and the selection of colonies representing rare retrotransposition events. **f** | Transcription and regulation. Long-terminal repeat (LTR) functions can be assessed using standard assays for promoter functions or for enhancer or repressor functions, such as linking the LTR to a suitable reporter ORF (for example, luciferase (luc) or β-galactosidase (β-gal)) and introducing it by transfection into a suitable cell line.

For example, promoters and regulatory elements can be studied using standard reporter assays[65,66]; retrotransposition can be detected and quantified using a sensitive cell culture assay[58,67–69]; and reconstituted proteins can be studied in the context of infection using pseudotyped particles or by replacing discrete domains in the polyproteins of replication-competent retroviruses with the homologous ERV domains[70–72]. Other viral platforms are also useful. For example, a rhabdovirus was engineered to express an ancient ERV Env in place of its own glycoprotein, creating a tool to delineate the viral entry pathway and to identify the cellular cofactors likely to have been used by the extinct virus[73,74].

Reconstituted ERV proteins have been used to identify the entry receptors for two ancient retroviruses[57,75] and to test the sensitivity of ancient viruses to host defence factors[76]. Reconstituted virus-like particles related to HERV-K(HML2) loci have been used to examine tropism, to test sensitivity to innate immune effectors and to reveal differences between genome-wide integration site preferences (in cell culture) and the distribution of HERV-K(HML2) loci in the human genome[70,77–80].

There are now many examples of ERV loci that have evolved to provide important cellular functions, attracting the attention of researchers from various fields including virology, genome biology, population genetics and evolutionary developmental biology. In this regard, the past 100 years of research on retroviruses have provided a wealth of insight, as well as the various assays described above, that can be used to explore how retroviruses and ERVs have influenced the evolution of vertebrate genes and genomes.

Exaptation of endogenous retroviruses

Gould and Vrba coined the term exaptation to be used when referring to an adaptation that fulfils a new function distinct from its originally selected function[81]. They discussed, among other examples, repetitive DNA, including transposable elements, as a special class of sequences available for exaptation[81]. The idea that transposable elements may have roles in gene regulation was proposed in the mid-1950s by McClintock[82] and was incorporated into an early hypothetical model of gene regulation[83]. ERVs are often categorized as transposable elements and are related to LTR retrotransposons (BOX 2). However, the ultimate origins of ERVs are exogenous retroviruses, whose sequences reflected adaptation to a wide variety of vertebrate hosts and a spectrum of cellular niches. This distinctive natural history may contribute to the exaptive potential of ERVs, connecting the biology of rapidly evolving, exogenous retroviruses to the co-opted functions of their germline counterparts.

Exaptation of Env proteins

Most examples of exaptation of ERV-coding sequences involve *env* genes. The primary viral function of Env glycoproteins is to facilitate entry into host cells, which involves binding to cell surface receptors and driving fusion of the virion and cellular membranes (FIG. 4). For many retroviruses, expression of Env also interferes with cell surface expression of the receptor, rendering the cell resistant to reinfection — a phenomenon known as superinfection interference[84,85].

Well-documented examples of *env* exaptation fall into two distinct categories: the first comprises syncytins, which are ERV-encoded Env proteins that function in mammalian placental morphogenesis[7], and the second comprises ERV Envs that confer resistance to exogenous viral infection through mechanisms analogous to superinfection interference[86].

Syncytins. The placental syncytins are the focus of several recent reviews[7,87,88]. Briefly, these ERV-encoded glycoproteins drive fusion of cytotrophoblasts to form the multinucleate syncytiotrophoblast layer[7]. The underlying mechanism involves cell–cell fusion and is analogous to viral entry (which depends on receptor binding to trigger fusion of virion and cellular membranes) (FIG. 4). The syncytins are a striking example of convergent evolution, having originated independently across multiple mammalian lineages, including marsupials[89], as well as in at least one species of live-bearing reptile[90].

Syncytin function has been confirmed in mice[91]. However, because most reported syncytins arose independently in different mammalian clades, they are not homologues, and it is therefore risky to extrapolate results of mouse experiments to nonrodent species. Identifying ERV-encoded syncytins in nonmodel organisms is instead based on rigorous but indirect criteria[7]. These include conservation within a clade of related taxa, placenta-specific expression and fusogenicity in cell culture (FIG. 3c). In the case of human syncytins, additional histological and tissue-culture-based evidence is also consistent with the proposed function (reviewed elsewhere[88]). Confirming other syncytins may require additional experiments in representative nonmodel organisms or genetic association studies correlating variant *syncytin* alleles with relevant phenotypes. Finally, it remains possible that some of the syncytins have additional, as yet unrecognized, functions.

Whether the receptors used by syncytins are the same as those used by the originating retroviruses is difficult to establish — most syncytins are tens of millions of years old, and the retroviruses that produced them are probably extinct. However, there is a precedent for reconstituting Env proteins from ERV sequences and using these to identify the receptors used by ancient retroviruses[57,75] (FIG. 3); similar approaches may be useful for establishing whether a syncytin and related ERVs shared the same receptor.

Env-mediated entry restriction. Viral interactions with host macromolecules fall into two broad categories: those exploited by viruses to ensure optimal fitness and those that have evolved to block infection. Host cell factors in the latter category are often referred to as restriction factors. Examples of restriction factors that inhibit replication of retroviruses include the APOBEC3 family DNA editing enzymes, tetherin (also known as BST2), SAMHD1 and TRIM5α[92,93]. Viral genes acquired by endogenization also have the potential to become restriction factors[86]. Among these restriction factors,

Exaptation
A trait that evolved on the basis of one function that has subsequently evolved to provide a different function.

Superinfection interference
A phenomenon by which prior infection of a cell renders it resistant to reinfection by retroviruses using the same entry receptor; often mediated by the viral Env glycoprotein.

Syncytins
Glycoproteins of retroviral origin that fulfil cellular functions involving receptor-mediated membrane fusion; thus far, all reported syncytins function as placental syncytins.

Syncytiotrophoblast
A multinuclear layer that forms through fusion of mononuclear cytotrophoblasts.

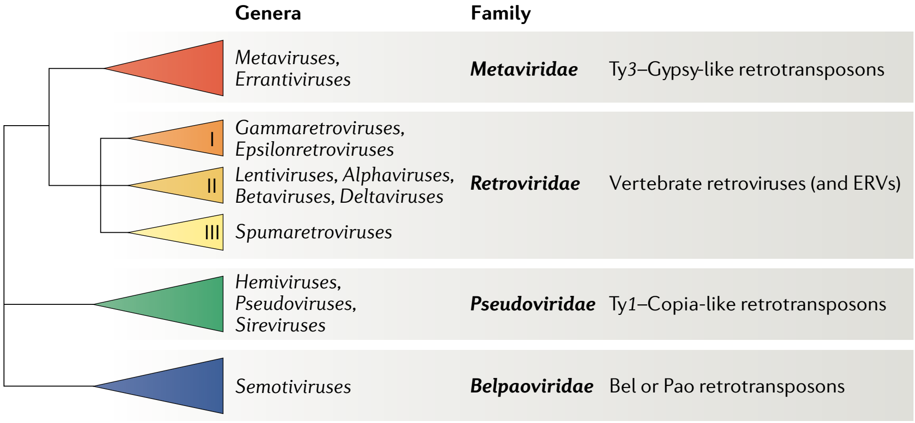
Restriction factors
Host-encoded factors that have evolved by natural selection to suppress or prevent viral replication at the cellular level.

Box 2 | Endogenous retroviruses or LTR retrotransposons?

The distinction between endogenous retroviruses (ERVs) and long-terminal repeat (LTR) retrotransposons is not always clear, particularly in the case of defective ERVs that have undergone post-endogenization expansion by 'piggybacking' on the replicative machinery of other elements. In the broadest sense, LTR retrotransposons are elements that have evolved to propagate intracellularly by reverse transcription and reinsertion into host cell DNA and are often adapted to the regulatory milieu of germline cells. LTR retrotransposons have several features indicating common ancestry with retroviruses. These include *gag*-like, *pro*-like and *pol*-like genes; encapsidation of RNA genomes in a nucleocapsid complex; reverse transcription primed by a cellular tRNA (first strand synthesis) and an RNase resistant RNA primer (second strand synthesis); and production of an integrated DNA genome flanked by identical LTRs. Generally speaking, LTR retrotransposons lack an extracellular phase (virion).

ERV insertions arise, at least initially, as a random consequence of horizontal transmission and replication of exogenous retroviruses, which are adapted for intercellular and interhost transmission in the form of extracellular virions. However, whereas some ERV insertions may remain limited to one or a few loci, others may undergo expansions in copy number by adapting to germline transmission — effectively becoming LTR retrotransposons[24,25].

The ambiguous terminology can be resolved by distinguishing between evolutionary origins and current status (see the figure). Within the spectrum of elements in vertebrate genomes frequently grouped together as ERVs, it is possible to identify two broad categories of sequences. The first is related to ancient lineages of LTR retrotransposons, such as those found in the *Pseudoviridae*, *Belpaoviridae* and *Metaviridae* families[180]. Collectively, these are widely distributed among vertebrates, fungi, plants and protists and have deep evolutionary origins likely predating the appearance of vertebrates. Despite their distant relationships to retroviruses and the occasional presence of an *env*-like gene, these viruses are distinct from the *Retroviridae* in reverse-transcriptase-based phylogenies (see the figure). The second category comprises elements found exclusively in vertebrate genomes and cluster within the family *Retroviridae*, for the most part, interleaving within and between genera typified by extant retroviruses[5,27,30]. This pattern indicates a shared common ancestry and includes ERVs with obvious relationships to exogenous retroviruses as well as ERVs that either lack or have lost features associated with extracellular spread (for example, *env* genes). Thus, 'ERV' and 'LTR retrotransposon' need not be mutually exclusive terms, depending on context — the former refers to a particular phylogenetic origin, whereas the latter is consistent with adaptation to an intracellular niche. When clarity is essential, 'ERV' can be used to specify elements with a retroviral origin (recent or in the distant past), including those with evolutionarily derived features reflecting adaptation for intracellular replication.


Genera
Family
Metaviruses, Errantiviruses
Metaviridae
Ty3–Gypsy-like retrotransposons
I
Gammaretroviruses, Epsilonretroviruses
II
Lentiviruses, Alphaviruses, Betaviruses, Deltaviruses
III
Spumaretroviruses
Retroviridae
Vertebrate retroviruses (and ERVs)
Hemiviruses, Pseudoviruses, Sireviruses
Pseudoviridae
Ty1–Copia-like retrotransposons
Semotiviruses
Belpaoviridae
Bel or Pao retrotransposons

Roman numerals (I, II and III) indicate the three classes of retroviral reverse transcriptase.

the most common are ERV-encoded proteins that block viral entry through receptor interference.

In 1981, it was reported that three endogenous loci of chickens (*EV3*, *EV6* and *EV9*) confer entry-level blocks to infection by avian leukosis virus, most likely by receptor interference[94] (FIG. 4). The authors correctly predicted that similar functions would be found in other species known to harbour ERVs. The prototypical example of ERV-mediated entry restriction is the murine *Fv4* gene (also known as *Akvr-1*). *Fv4* was first defined as a locus conferring resistance to experimental infection of laboratory mice by ecotropic murine leukaemia virus (MLV) and subsequently was correlated with expression of a novel MLV-related Env protein[95]. A similar resistance phenotype was observed in a population of feral mice in California, United States[96]. Cloning of *Fv4* revealed that the same gene was responsible for the observed resistance in both cases, and sequencing revealed that *Fv4* comprises a defective MLV provirus that retains an intact *env* ORF but lacks most of the 5′ half of the provirus including the 5′ LTR[97]. *Fv4* expression is instead regulated by cellular sequences adjacent to the insertion[98]. *Fv4* was likely selected by virtue of its ability to block infection by ecotropic strains of MLV. Two additional examples of genes encoding Env-mediated restriction in mice, *Rcmf* and *Rcmf2*, confer resistance to polytropic MLV strains; as with *Fv4*, both genes are incapable of expressing infectious, replication-competent virus[99,100] (FIG. 5).

Env glycoproteins are normally anchored in the viral and cellular membranes by a membrane-spanning domain in the TM subunit (FIG. 2). However, ERV-mediated entry restriction can also involve secreted Env. In such cases, the secreted proteins have mutations resulting in a premature truncation, thereby eliminating

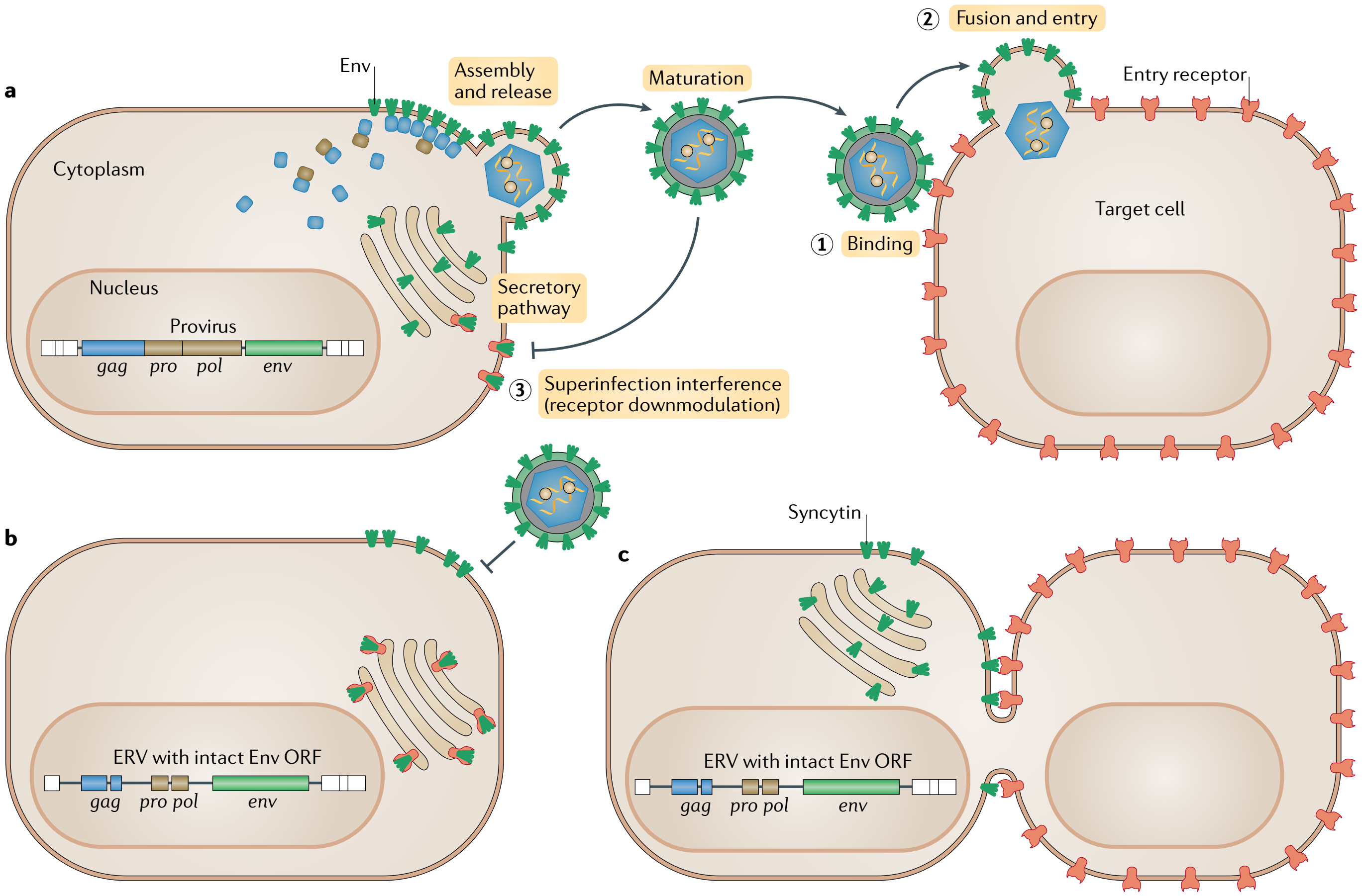


Fig. 4 | **Env exaptation and the relationship between ancient viral functions and current genome functions.** **a** | Normal functions of retroviral Env glycoproteins are presented. The Env proteins of retroviruses assemble as heterotrimeric complexes of the surface unit–transmembrane (SU–TM) domain that traffic through the secretory pathway to sites of viral assembly and are incorporated into nascent virions composed of Gag and Gag–Pro–Pol proteins (blue and blue–brown) budding through the cellular membrane (left). Env complexes on mature virions function to recognize the cell surface entry receptors (red rectangles) on target cells (step 1) and drive fusion of the virion and cellular membranes to release the nucleocapsid core of the virus into the cytoplasm of the host target cell (right; step 2). In addition, newly synthesized Env proteins of some retroviruses, including gammaretroviruses, interact with the cognate receptor proteins in the producer cell (left), thereby rendering the infected cell resistant to reinfection, a phenomenon known as superinfection interference (step 3). **b** | Endogenous Env proteins that restrict viral entry retain the receptor-binding properties of the original exogenous viral Env glycoprotein in order to inhibit entry by a mechanism analogous to superinfection interference. Fusogenicity is not required for this form of resistance, and such proteins appear to have lost the ability to drive membrane fusion, either by drift or selection (see the main text). The other viral genes (*gag*, *pro* and *pol*) are often disrupted by mutations. **c** | Syncytins are also endogenous Env proteins, whose functions require both the receptor-binding and membrane fusion functions of the ancient retroviral Env glycoproteins from which they are derived. Cell surface syncytin molecules bind to a cognate receptor on target cells and drive cell–cell membrane fusion in a manner analogous to virion–cell membrane fusion during entry.

the membrane-spanning domain. For example, feline REFREX proteins are truncated Env proteins derived from endogenous feline leukaemia virus (FeLV) that block entry of exogenous FeLV[101]. A truncated Env in the human genome, encoded by the *suppressyn* gene (also known as *ERVH48-1*), binds the receptor ASCT2 (also known as ATB^0) used by syncytin 1 and several retroviruses; thus, *suppressyn* may have evolved to block entry of a virus that uses ASCT2, as a negative regulator of syncytin 1 (REF.[102]), or both.

Could ERV genes have evolved to restrict viruses that are now extinct? Proof of principle can be accomplished through reconstruction and functional analysis of ERV *env* genes (FIG. 3). This was recently done to demonstrate the antiviral function of *HsaHTenv*, which encodes a fusion-defective HERV-T Env in the human genome[57]. To test the hypothesis that *HsaHTenv* expression results in entry restriction, a functional HERV-T Env (representing the ancestral retrovirus) was first reconstructed and then used to identify the corresponding receptor. Expression of native *HsaHTenv* was found to block infection by virions bearing functionally reconstituted HERV-T Env through receptor interference[57].

More than two dozen *env* ORFs have been identified in the human genome[103,104]; for most of these, there is, as yet, no direct evidence that they confer resistance to retroviruses in vivo. Intriguingly, HIV-1 infection of primary human CD4$^+$ T cells induces expression of HERV-K (HML2) loci[105]. Some HERV-K(HML2) loci encode intact *env* ORFs, and transfection and expression of at

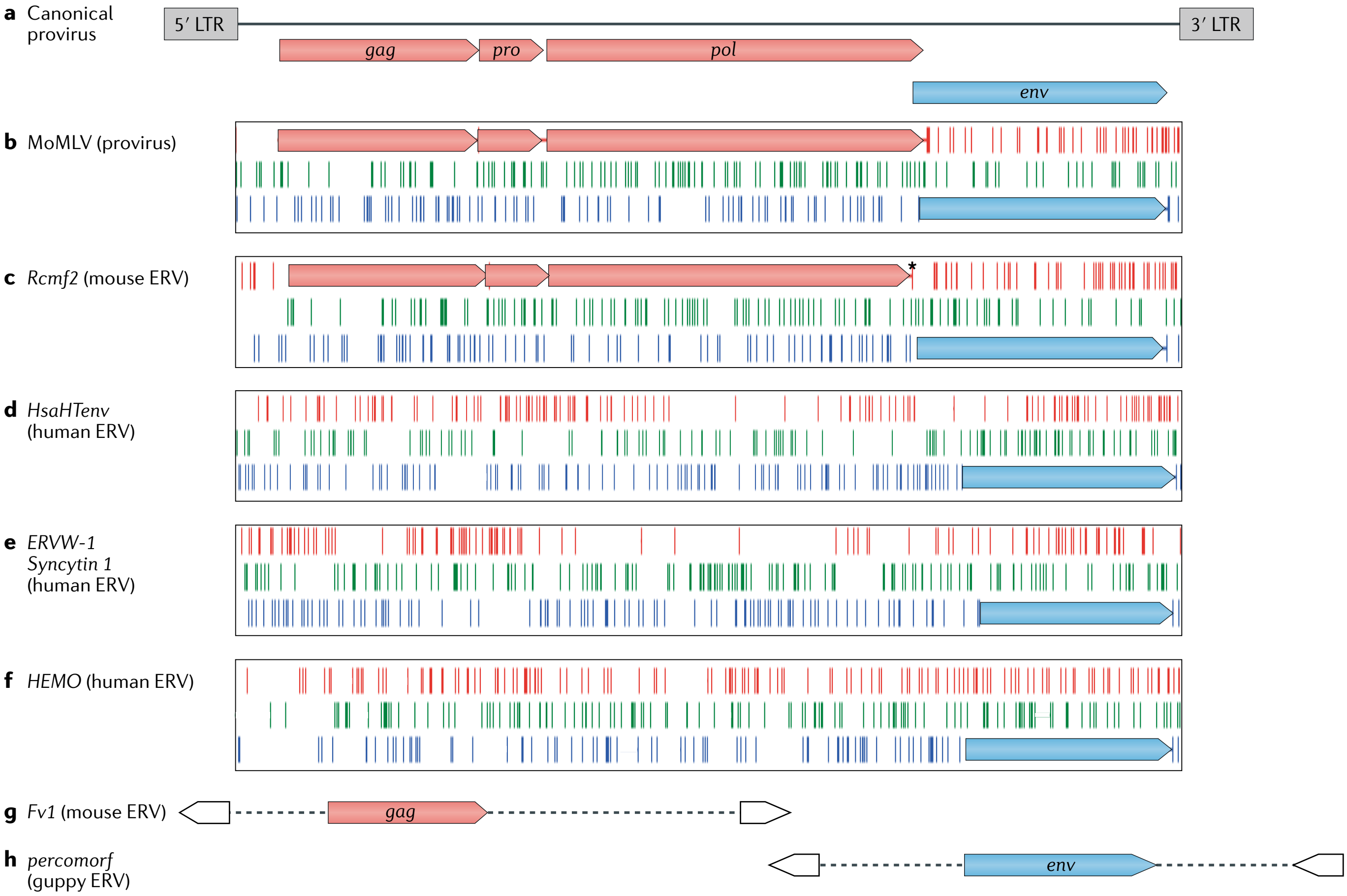


Fig. 5 | **The effects of drift and selection on endogenous retrovirus genes.** Endogenous retroviruses (ERVs) arise by the same mechanisms that produce integrated proviruses in somatic cells and have an identical structure at the time of insertion. However, as a consequence of random genetic drift, the ERV locus will acquire random substitutions over time. In the absence of selection, the ORFs encoded by the viral genes will eventually become disrupted through the accumulation of missense mutations, owing to either nucleotide substitutions or insertions or deletions that shift the reading frame. The rate at which the ERV diverges from the original sequence will mirror the background neutral substitution rate of the organisms' genomes in which it resides. The top panel (part **a**) represents a hypothetical provirus produced by integration of a simple retrovirus without any accessory genes. A provirus produced by the Moloney isolate of murine leukaemia virus (MoMLV) (part **b**); beneath the MoMLV genome is a display showing the positions of every stop codon in all three forward reading frames (stop codons are indicated by vertical lines in red (frame 1), green (frame 2) and blue (frame 3)). The intact *gag–pro–pol* ORFs in frame 1 and the *env* ORF in frame 3 are shown as horizontal red and blue box arrows, respectively. The murine *Rcmf2* locus is an MLV-related ERV that confers resistance to exogenous MLV infection (part **c**)[100]. Resistance is due to receptor interference mediated by expression of an MLV Env protein. The *Rcmf2* ERV is a recent unfixed insertion into the mouse germ line, with nearly intact ORFs except for a premature stop codon in *pol* that truncates the integrase protein (asterisk). The human *HsaHTenv* locus (part **d**) encodes an entry-restricting Env protein that may have provided resistance against an extinct retrovirus[57]. Orthologues of *HsaHTenv* are found in the same location in the genomes of chimpanzees, gorillas and orangutans, consistent with an estimated integration time of more than 13 million years ago. Since that time, the *gag–pro–pol* ORFs have been heavily disrupted by stop codons. Assuming this reflects the background accumulation of neutral substitutions, the *env* reading frame should also have acquired up to a dozen stop codons in the same time frame[57]. Instead, the pattern reflects purifying selection focused on the *env* gene, consistent with the proposed function as an antiviral defence gene. *ERVW-1* is the human gene encoding syncytin 1 (part **e**), an ERV-derived Env protein that plays an essential role in placental development in humans, apes and old-world monkeys. The ERV is estimated to be more than 25 million years old and displays the dichotomous pattern signifying exclusive preservation of the *env* gene by purifying selection[115]. The more than 100 million-year-old *HEMO* locus (part **f**) expresses a truncated Env glycoprotein of unknown function[108]. The pattern is consistent with purifying selection, strong evidence that the HEMO protein serves (or had previously served) one or more important functions. The murine *Fv1* gene (part **g**) encodes a partial Gag protein and functions as an early post-entry restriction to a variety of retroviruses. Recent estimates place the original insertion at ~45–50 million years ago[122,123]. Although selection has maintained the *Fv1* ORF, there is little to no remaining trace of the original provirus. The horizontal dashed line represents a non-ERV sequence, and the open box arrows indicate unrelated genes flanking the *Fv1* locus. The *percomorf* locus (part **h**) of ray-finned fish encodes a gamma-type Env protein of unknown function. The ORF lies in an intron of the *dnajc6* gene in the opposite orientation. The box arrows depict the ultimate and penultimate exons of *dnajc6*. On the basis of the distribution of *percomorf* among the genomes of extant fish, the ORF is estimated to be 109–140 million years old[107]. As with *Fv1*, all traces of the original provirus have been lost.

least one of these in laboratory cell lines inhibit production of infectious HIV-1 (REF.[106]). Inhibition is not due to receptor interference, raising the possibility that one or more of these loci may exert antiviral effects through a novel mechanism; whether inhibition manifests in vivo remains to be determined.

Highly conserved ERV* env *genes of unknown function. The oldest intact ERV *env* genes reported are the *percomorf* gene of ray-finned fish[107] and the primate *HEMO* gene[108]. The preservation of these as intact ORFs for more than 100 million years reflects long-term purifying selection and strongly suggests that both genes are likely to encode novel cellular functions. The age and conservation of *percomorf* argue against an antiviral function and instead suggest that *percomorf* may represent a new category of exapted function involving receptor-mediated membrane fusion. HEMO lacks a furin cleavage site and a hydrophobic fusion-peptide, indicating that it cannot be a fusogen, although it may retain its receptor-binding activity. Characterization of the human homologue reveals that HEMO is expressed as a full-length Env protein, which is cleaved by an unknown cellular protease to release a truncated extracellular form[108]. The secreted form is detectable in the blood of pregnant women and in placental blood and tissues, but its functions remain unknown.

Exaptation and features of Env. A majority of retrovirus Env proteins belong to one of two types, the gamma-type and the beta-type[33] (FIG. 2c). Intriguingly, ERV loci with gamma-type *env* sequences are widely distributed among vertebrate genomes, whereas beta-type *env* sequences are largely found in mammalian genomes[28,33]. The reason for these markedly different distributions is unknown.

Intriguingly, almost all known examples of Env exaptation involve gamma-type Envs, including all the mammalian syncytins. The reasons for this bias are also unknown, but certain features may predispose gamma-type Env to exaptation. Gamma-type Envs have a modular arrangement, with discrete receptor-binding domains (RBDs) within the amino-terminal half of the Env surface (SU) subunit[109]. One speculative possibility is that modularity uncouples evolution of receptor specificity from functions located outside of the RBD (that is, by recombination), allowing these to evolve independently. Additionally, the carboxyl termini of gammaretrovirus Envs suppress fusogenicity[110,111]. These short R peptides are removed by the viral protease after virion assembly such that only Env complexes present on mature virions are fusion competent[110,111]. However, immature Env complexes within the virus-infected cell are still able to bind their cognate receptors and mediate superinfection interference. Thus, by preventing spontaneous cell–cell fusion, R domains may enhance the probability of fixation of gamma-type ERV *env* genes. Additional mutations might be selected to prevent activation in *trans* (for example, by other gammaretroviruses). Indeed, several reported entry-blocking ERVs are fusion defective[57,112,113]. Similarly, ERV–Fc *env* ORFs found in the genomes of multiple mammals, including humans[49,114], have defects that prevent fusion and preclude a syncytin-like function (K. Halm, personal communication). Discovery of additional entry-blocking Envs may establish whether loss of fusogenicity is a common feature of such loci and whether the loss is the result of drift or selection. By contrast, syncytins require both receptor-binding and membrane fusion activities to function, whereas features that prevent fusion should be eliminated or modified by selection. Indeed, this is the case for human syncytin 1, which has lost R-peptide-mediated regulation and can direct viral protease-independent fusion[115]. Whether similar adaptations are found in other syncytins remains to be determined.

There are relatively few reports of beta-type ERV Envs with exapted functions[116]. Perhaps, beta-type Envs contribute novel functions, distinct from those associated with gamma-type Envs. Retroviruses with beta-type *env* genes often encode additional ORFs overlapping *env*[117], which could influence the selection of endogenized forms.

Exaptation of other ERV proteins

There are a few reports of exaptation involving *gag* and *pol*[118]. The prototypical example is the *Fv1* gene of mice[119], which confers resistance to MLV[120]. *Fv1* is an endogenous *gag* gene related to ERV-L elements[119,121]; expression of *Fv1* blocks incoming viral capsid cores shortly after entry. *Fv1* orthologues have been identified in a broad range of rodent species, and the estimated insertion time is 45–50 million years ago[122,123]. Indeed, some *Fv1* homologues restrict retroviruses unrelated to MLV[124], suggesting that Fv1 does not recognize conserved amino acid motifs but may instead detect structurally conserved spatial patterns in the hexameric lattice typical of retroviral capsid cores[125].

The *EnJS56A1* locus of domestic sheep (*Ovis aries*) also encodes a Gag protein, which can act as a trans-dominant inhibitor of a related exogenous virus known as Jaagsiekte sheep retrovirus (JSRV)[126,127]. Unlike Fv1, which blocks MLV replication shortly after entry, *EnJS56A1* acts at a late stage in the JSRV replication cycle, interfering with proper trafficking and assembly of progeny virions[126,127].

Gag-mediated antiviral functions have not been reported for human ERVs, although HERV-K(HML2) Gag has been shown to inhibit HIV-1 in cell culture[128], raising the possibility that one or more HERV-K(HML2) loci may encode a protein that confers a late-stage block to the lentiviral replication cycle. It is not yet known whether this effect manifests in vivo, and HERV-K(HML2) loci have not been identified in reported genetic surveys of HIV-positive cohorts or in cellular screens for HIV-1-interacting factors. It was predicted that several human proteins are structurally related to the retrovirus Gag and Gag–Pro–Pol polyproteins (although many are likely derived from LTR retrotransposons)[129]; one of these, ARC, assembles into capsid-like structures that are strikingly similar to retroviral capsid cores[130,131]. Another, SASPase, is structurally and functionally analogous to retroviral proteases[132].

Evidence for accessory genes is sometimes present in ERVs, particularly those related to exogenous retroviruses with complex genomes[34,36–38,133]. In some cases,

Purifying selection
A component of natural selection; refers to selection that eliminates deleterious or suboptimal variants of a gene or sequence that arise by mutation.

R peptides
The last 17–20 residues of the cytoplasmic carboxyl termini of gammaretroviral Env proteins, which are cleaved off by the viral protease during virion maturation to activate fusogenic potential.

ERV-L elements
An ancient family of related endogenous retrovirus (ERV) elements found in the genomes of all mammals; distantly related to spumaretroviruses.

Exogenous virus
A horizontally transmitted virus, as distinguished from endogenous viruses.

these bear little resemblance to the accessory genes of their exogenous relatives, and any viral or exaptive functions remain speculative. A possible example of accessory gene exaptation involves the *Mls* (also known as *Mtv*) genes of mice, which originate from mouse mammary tumour virus (MMTV) *sag* genes[134]. These encode superantigens that activate T cells[135]. Expression of different endogenous *sag* loci (*Mls* genes) result in clonal deletion of different cognate T cell subsets; by eliminating target cells that support viral infection and dissemination, *Mls* expression may provide resistance to exogenous MMTV strains of the same Sag specificity[136].

Betaretroviruses encode proteins required for optimal expression of unspliced viral RNA[137,138]. This raises the possibility that ERV-encoded versions of these proteins could also affect cellular transcripts[139]. Several HERV-K (HML2) loci in the human genome have the potential to encode such a protein, an RNA transport factor known as REC[140,141]. REC binds the 3′ end of unspliced viral RNA through a REC-responsive element (RcRE) encoded in HERV-K(HML2) LTRs[140,141], of which there are close to 1,000 in the human genome[64]. Interestingly, the Rev protein of HIV-1 can also bind the HERV-K(HML2) RcRE[140,141]. Direct evidence that Rec or Rev influences transport of cellular transcripts in vivo has not been reported but may be worthy of investigation.

Genomic signatures of exaptation

Initially, exapted ERV ORFs were identified by traditional means, for example, in seeking to explain a specific phenotype or by functional assays of candidate genes. Exapted ERV genes can be identified without a priori knowledge of a phenotype. For example, most *syncytins* and resistance-conferring *env* genes are in proviruses with disrupted *gag*, *pro* and *pol* genes. This reflects the degree to which the locus has accumulated random substitutions. The juxtaposition of an intact *env* ORF is therefore consistent with purifying selection focused on *env* (FIG. 5). Statistical tests of selection can also be applied, such as the dN:dS ratio (ω)[142]. The accumulation of silent changes (dS) sets a baseline expectation for drift, against which the accumulation of nonsynonymous changes can be evaluated. Ratios <1, =1 or >1 indicate purifying selection, drift and positive selection, respectively. Importantly, purifying selection and positive selection are not mutually exclusive; even for genes that have experienced positive selection, a majority of codons still evolve under purifying selection to maintain overall structure and function. Average ω values for *percomorf* and *HEMO* are <1, consistent with long-term purifying selection and strong indications that these genes encode functional proteins[107,108]. In contrast to *percomorf* and *HEMO*, analysis of *Fv1* reveals a combination of long-term positive selection with periodic bouts of lineage-specific selection focused on residues involved in target specificity[122,123], a combination typical of many antiretroviral proteins[92,93]. If there are insufficient taxa to calculate ω, one can also simulate neutral evolution of the ORF to derive a probability distribution for inactivating mutations[57,143].

Envs that have essential roles in organismal development should evolve under continuous purifying selection. By contrast, those that inhibit replication of exogenous viruses may experience shorter-lived bouts of selection — when the exogenous virus becomes extinct or is replaced with a resistant variant, selection should be relaxed and the exapted gene subject to loss by drift[57,144] (FIG. 1). Consistent with these predictions, syncytins have estimated ages ranging from approximately 12 million years to more than 80 million years[7,89,90], whereas receptor-blocking Envs are younger, as reflected by narrower taxonomic distributions, insertional polymorphism and estimated integration times that are less than 20 million years ago[57,96,145,146]. Conceivably, many of the defective *env* sequences in the genomes of humans and other vertebrates may have once functioned to block viral entry but have since decayed owing to extinction of the selective agent[57].

Exaptation of ERV non-coding elements

Integrated proviruses, and by extension ERVs, can alter the regulation of nearby genes[12,147–149] and potentially influence the control of genes thousands of base pairs away[150]. Indeed, there are numerous examples of ERV LTRs functioning as novel promoters or transcription-factor-binding sites for genes, and there are now also examples of ERVs giving rise to novel regulatory long non-coding RNAs[151–153]. Several recent comprehensive reviews discuss the potential involvement of ERVs in both normal and aberrant gene regulation[12,149,154]. Importantly, thanks to ongoing acquisition and loss of ERV loci over evolutionary timescales, even closely related species vary in the composition and genomic distribution of ERV LTRs. Thus, through their effects on regulation of key genes, these elements may contribute to phenotypic diversification and, as a consequence, will be subject to exaptation by natural selection.

Recently, several lines of evidence suggest that ERVs may facilitate the concerted evolution of sets of genes that are regulated in coordination within so-called gene regulatory networks (GRNs)[154–160]. The coordinated regulation of genes can involve shared *cis*-acting regulatory elements (CREs), and the evolutionary rewiring of GRNs may be a source of phenotypic variation and species diversification[161]. At issue is whether shared CREs evolve de novo, which depends on random substitutions generating similar or identical motifs for multiple genes in a GRN, or whether there are mechanisms that facilitate concerted evolution of loci linked within GRNs[162]. Endogenization and parallel fixation of related ERV LTRs, containing similar or identical viral promoters and associated CREs, provide a compelling solution to the difficulties of the de novo hypothesis[154,155].

Hypothetically, several unique properties of ERV could facilitate a role in GRN evolution. First, LTRs are densely packed with regulatory elements, including promoters and transcription-factor-binding sites (FIG. 2). These reflect the host range and tissue tropism of the virus at the time of integration, which may dictate the exaptive potential of any resulting ERVs. Second, although retroviral integration does not target specific motifs, it is also not perfectly random, with some retroviruses displaying preferences for transcriptional units or for promoter regions[163]. Thus, although endogenization

Positive selection
The selection that favours fixation of changes in a gene, such as when a virus escapes from virus-specific antibodies through changes in a target epitope.

may produce insertions distributed widely across the genome (and the host population), for some types of retrovirus, these insertions may be enriched in or near transcription units. Most of these are probably lost by drift or negative selection, but those that alter GRNs in beneficial ways will be favoured by natural selection. Third, sequence similarity within LTR families could facilitate the spread of a new motif in one locus to related loci, for example, by ectopic recombination or gene conversion[59,164–166]. Although speculative, a fourth feature of ERVs that may influence their role in GRN evolution in multiple ways is the propensity to form solo-LTRs. For example, solo-LTR formation would eliminate proviral sequences that are known targets for epigenetic silencing[167] and may also activate the regulatory influence of the LTR on adjacent genes[168]. If solo-LTR formation is required to activate regulatory potential, then recombination and deletion would have to precede function, resulting in a temporal separation between the original integration event and the eventual manifestation of novel phenotypes subject to selection — possibly spanning hundreds or thousands of host generations. Moreover, solo-LTR formation may occur repeatedly at the same locus[169], effectively increasing the probability of fixing a solo-LTR allele. Conversely, the probability of solo-LTR formation by homologous recombination decreases once the 5′ and 3′ LTR sequences begin to diverge[170] such that the potential for exaptation may diminish with time.

Conclusions

At present, the most thoroughly documented ERVs are those of mammals, particularly those of mice and humans, although analyses of nonmammalian genomes are beginning to yield novel insights[1,3,4,27,171,172]. Molecular understanding of ERV biology, including viral functions, exapted cellular functions and contributions to disease, is even narrower, being mostly based on specific ERV or ERV families found in model organisms (for example, mice, chickens, livestock and pets). These are often inbred, domesticated species, which may not accurately reflect the process of endogenization as it occurs across generations in natural outbred populations. Broad comparative approaches may be the key to determining which biological properties, if any, predispose some retroviruses to germline invasion and for examining the impact of host biology and population dynamics on endogenization. Insights could come from studying natural populations currently in the early stages of endogenization[55].

As a case in point, and despite the rapidly growing list of published examples, it is unclear whether LTR exaptation represents a major or minor mechanism of vertebrate GRN evolution[173]. To provide a major source of selectable variants, endogenization must produce many more insertions than are ultimately preserved by selection, yet little is known about the origins and initial population genetics of newly formed ERVs in natural populations. Consequently, incorporating LTR exaptation into general models of GRN evolution invokes several important questions: what triggers increases in ERV copy number (amplification bursts) in some lineages but not others? Have bursts of endogenization occurred with sufficient frequency during vertebrate evolution to explain the observed levels of diversity? Are these bursts temporally correlated with major speciation events or the appearance of novel phenotypes?

Similarly, the literature on exapted ERV proteins mostly relates to the identification and confirmation of cellular functions, with little attention given to understanding whether and how these genes undergo further modification for optimal function. Do they acquire additional regulatory refinements, and if they do, how? Do they experience additional adaptions in structure and function of the encoded proteins? Indeed, exapted ERV ORFs may prove generally useful for understanding how newly formed protein-coding genes gain interactions with other host factors and become integrated into existing regulatory circuits.

As a complement to molecular evolutionary analysis of ERVs, several new technologies now make it possible to test functional hypothesis directly. For example, deep-sequencing methodologies have been used for transcriptional profiling of ERV loci, for population-level analysis of germline integration and for detecting rare integration events[45,46,54,55,63,174]. Such approaches can be coupled with new techniques enabling analysis of individual ERV loci in primary cells and tissues and assessment of their regulatory potential. These include methods for identifying epigenetic modifications and DNA–nucleic acid interactions and protocols for analysing events at the single-cell level. As ERVs often belong to closely related, multilocus families, unambiguous assignment of sequencing reads to specific loci can be problematic, particularly when analysing younger, less divergent families. Thus, correlations between transcription and expression of ERV families and external triggers or various disease phenotypes have been observed, but such studies may lack the resolution to attribute observed biological effects to specific loci within a larger ERV family[11]. Useful insights come when care is taken to map reads precisely[65,105,175] or to assess candidate ERV genes individually[106]. Finally, advances in genetic manipulation, including small interfering RNA and CRISPR–Cas, provide tools for perturbing and analysing native ERVs, including protocols for altering multiple loci in parallel at the cellular[150] and organismal[176] levels.

More than 100 years have passed since the discovery of the first retroviruses[177,178], and a similar time span marks the origins of evolutionary genetics as a distinct discipline[179]. The study of endogenous retroviruses combines concepts from both fields, while the potential for ERVs to facilitate evolution of developmental and morphological diversity touches on fundamental questions in evolutionary developmental biology. The potential connections to cancer and autoimmune diseases have also drawn considerable interest from scientists in a variety of fields[11]. Going forward, ERV research encompassing any combination of these areas should be embedded in a framework of population genetics theory while incorporating knowledge and methods gained from over a century's worth of research on all aspects of retrovirus biology.

Published online 8 April 2019

1. Herniou, E. et al. Retroviral diversity and distribution in vertebrates. *J. Virol.* **72**, 5955–5966 (1998).
2. Aiewsakun, P. & Katzourakis, A. Endogenous viruses: connecting recent and ancient viral evolution. *Virology* **479–480**, 26–37 (2015).
3. Xu, X., Zhao, H., Gong, Z. & Han, G.-Z. Endogenous retroviruses of non-avian/mammalian vertebrates illuminate diversity and deep history of retroviruses. *PLOS Pathog.* **14**, e1007072 (2018).
4. Naville, M. & Volff, J.-N. Endogenous retroviruses in fish genomes: from relics of past infections to evolutionary innovations? *Front. Microbiol.* **7**, 1197 (2016).
5. Gifford, R. & Tristem, M. The evolution, distribution and diversity of endogenous retroviruses. *Virus Genes* **26**, 291–315 (2003).
6. Gifford, R. J. Viral evolution in deep time: lentiviruses and mammals. *Trends Genet.* **28**, 89–100 (2012).
7. Lavialle, C. et al. Paleovirology of 'syncytins', retroviral env genes exapted for a role in placentation. *Phil. Trans. R. Soc. B Biol. Sci.* **368**, 20120507 (2013).
8. Delviks-Frankenberry, K., Cingöz, O., Coffin, J. M. & Pathak, V. K. Recombinant origin, contamination, and de-discovery of XMRV. *Curr. Opin. Virol.* **2**, 499–507 (2012).
9. Groom, H. C. T. & Bishop, K. N. The tale of xenotropic murine leukemia virus-related virus. *J. Gen. Virol.* **93**, 915–924 (2012).
10. Suling, K., Quinn, G., Wood, J. & Patience, C. Packaging of human endogenous retrovirus sequences is undetectable in porcine endogenous retrovirus particles produced from human cells. *Virology* **312**, 330–336 (2003).
11. Young, G. R., Stoye, J. P. & Kassiotis, G. Are human endogenous retroviruses pathogenic? An approach to testing the hypothesis. *Bioessays* **35**, 794–803 (2013).
12. Babaian, A. & Mager, D. L. Endogenous retroviral promoter exaptation in human cancer. *Mob. DNA* **7**, 24 (2016).
13. Mager, D. L. & Lorincz, M. C. Epigenetic modifier drugs trigger widespread transcription of endogenous retroviruses. *Nat. Genet.* **49**, 974–975 (2017).
14. Young, G. R. et al. Resurrection of endogenous retroviruses in antibody-deficient mice. *Nature* **491**, 774–778 (2012).
15. Stoye, J. P. & Coffin, J. M. The four classes of endogenous murine leukemia virus: structural relationships and potential for recombination. *J. Virol.* **61**, 2659–2669 (1987).
16. Martinelli, S. C. & Goff, S. P. Rapid reversion of a deletion mutation in Moloney murine leukemia virus by recombination with a closely related endogenous provirus. *Virology* **174**, 135–144 (1990).
17. Stoye, J. P., Moroni, C. & Coffin, J. M. Virological events leading to spontaneous AKR thymomas. *J. Virol.* **65**, 1273–1285 (1991).
18. Benachenhou, F. et al. Evolutionary conservation of orthoretroviral long terminal repeats (LTRs) and ab initio detection of single LTRs in genomic data. *PLOS ONE* **4**, e5179 (2009).
19. Benachenhou, F. et al. Conserved structure and inferred evolutionary history of long terminal repeats (LTRs). *Mob. DNA* **4**, 5 (2013).
20. Copeland, N. G., Hutchison, K. W. & Jenkins, N. A. Excision of the DBA ecotropic provirus in dilute coat-color revertants of mice occurs by homologous recombination involving the viral LTRs. *Cell* **33**, 379–387 (1983).
21. Weiss, R. A. The discovery of endogenous retroviruses. *Retrovirology* **3**, 67 (2006).
22. Bannert, N. & Kurth, R. The evolutionary dynamics of human endogenous retroviral families. *Annu. Rev. Genomics Hum. Genet.* **7**, 149–173 (2006).
23. Gifford, R., Kabat, P., Martin, J., Lynch, C. & Tristem, M. Evolution and distribution of class II-related endogenous retroviruses. *J. Virol.* **79**, 6478–6486 (2005).
24. Belshaw, R., Katzourakis, A., Pac˘es, J., Burt, A. & Tristem, M. High copy number in human endogenous retrovirus families is associated with copying mechanisms in addition to reinfection. *Mol. Biol. Evol.* **22**, 814–817 (2005).
25. Magiorkinis, G., Gifford, R. J., Katzourakis, A., De Ranter, J. & Belshaw, R. Env-less endogenous retroviruses are genomic superspreaders. *Proc. Natl Acad. Sci. USA* **109**, 7385–7390 (2012).
26. Jern, P., Sperber, G. O. & Blomberg, J. Use of endogenous retroviral sequences (ERVs) and structural markers for retroviral phylogenetic inference and taxonomy. *Retrovirology* **2**, 50 (2005).
27. Hayward, A., Cornwallis, C. K. & Jern, P. Pan-vertebrate comparative genomics unmasks retrovirus macroevolution. *Proc. Natl Acad. Sci. USA* **112**, 464–469 (2015).
28. Bénit, L., Dessen, P. & Heidmann, T. Identification, phylogeny, and evolution of retroviral elements based on their envelope genes. *J. Virol.* **75**, 11709–11719 (2001).
29. King, A. M. Q., Adams, M. J., Carstens, E. B. & Lefkowitz, E. J. (eds) *Virus Taxonomy: Classification and Nomenclature of Viruses: The Ninth Report of the International Committee on Taxonomy of Viruses* (Elsevier, 2011).
30. Gifford, R. J. et al. Nomenclature for endogenous retrovirus (ERV) loci. *Retrovirology* **15**, 59 (2018).
31. Martin, J., Herniou, E., Cook, J., O'Neill, R. W. & Tristem, M. Interclass transmission and phyletic host tracking in murine leukemia virus-related retroviruses. *J. Virol.* **73**, 2442–2449 (1999).
32. Hayward, A., Grabherr, M. & Jern, P. Broad-scale phylogenomics provides insights into retrovirus-host evolution. *Proc. Natl Acad. Sci. USA* **110**, 20146–20151 (2013).
33. Henzy, J. E. & Johnson, W. E. Pushing the endogenous envelope. *Phil. Trans. R. Soc. B Biol. Sci.* **368**, 20120506 (2013).
34. Farkašová, H. et al. Discovery of an endogenous deltaretrovirus in the genome of long-fingered bats (Chiroptera: Miniopteridae). *Proc. Natl Acad. Sci. USA* **114**, 3145–3150 (2017).
This paper is the first to identify an ERV related to modern deltaretroviruses, the genus that includes human T-lymphotropic viruses and the bovine leukaemia virus.
35. Hron, T. et al. Remnants of an ancient deltaretrovirus in the genomes of horseshoe bats (Rhinolophidae). *Viruses* **10**, 185 (2018).
36. Katzourakis, A., Tristem, M., Pybus, O. G. & Gifford, R. J. Discovery and analysis of the first endogenous lentivirus. *Proc. Natl Acad. Sci. USA* **104**, 6261–6265 (2007).
37. Gifford, R. J. et al. A transitional endogenous lentivirus from the genome of a basal primate and implications for lentivirus evolution. *Proc. Natl Acad. Sci. USA* **105**, 20362–20367 (2008).
38. Gilbert, C., Maxfield, D. G., Goodman, S. M. & Feschotte, C. Parallel germline infiltration of a lentivirus in two Malagasy lemurs. *PLOS Genet.* **5**, e1000425 (2009).
39. Cui, J. & Holmes, E. C. Endogenous lentiviruses in the ferret genome. *J. Virol.* **86**, 3383–3385 (2012).
40. Han, G.-Z. & Worobey, M. A primitive endogenous lentivirus in a colugo: insights into the early evolution of lentiviruses. *Mol. Biol. Evol.* **32**, 211–215 (2015).
41. Hron, T., Fábryová, H., Pačes, J. & Elleder, D. Endogenous lentivirus in Malayan colugo (Galeopterus variegatus), a close relative of primates. *Retrovirology* **11**, 84 (2014).
42. Hron, T., Farkašová, H., Padhi, A., Pačes, J. & Elleder, D. Life history of the oldest lentivirus: characterization of ELVgv integrations in the dermopteran genome. *Mol. Biol. Evol.* **33**, 2659–2669 (2016).
43. Marchi, E., Kanapin, A., Byott, M., Magiorkinis, G. & Belshaw, R. Neanderthal and Denisovan retroviruses in modern humans. *Curr. Biol.* **23**, R994–R995 (2013).
44. Lee, A. et al. Novel Denisovan and Neanderthal retroviruses. *J. Virol.* **88**, 12907–12909 (2014).
45. Lenz, J. HERV-K HML-2 diversity among humans. *Proc. Natl Acad. Sci. USA* **113**, 4240–4242 (2016).
46. Holloway, J. R., Williams, Z. H., Freeman, M. M., Bulow, U. & Coffin, J. M. Gorillas have been infected with the HERV-K (HML-2) endogenous retrovirus much more recently than humans and chimpanzees. *Proc. Natl Acad. Sci. USA* **116**, 1337–1346 (2019).
This study uncovers multiple young ERVs in gorilla genomes related to human HERV-K(HML2), indicating recent activity in the gorilla lineage and raising the possibility that modern gorillas host an active HML-2 virus.
47. Goldstone, D. C. et al. Structural and functional analysis of prehistoric lentiviruses uncovers an ancient molecular interface. *Cell Host Microbe* **8**, 248–259 (2010).
This paper describes X-ray crystallography of the capsid proteins of two ancient lentiviruses in complex with host factor cyclophilin A. It also uses structures to infer phylogenetic relationships between extinct and extant lentiviruses.
48. Aiewsakun, P. & Katzourakis, A. Marine origin of retroviruses in the early Palaeozoic Era. *Nat. Commun.* **8**, 13954 (2017).
This paper describes the discovery and analysis of foamy-virus-like ERVs in marine vertebrates and suggests retroviruses may have originated early during vertebrate evolution.
49. Diehl, W. E., Patel, N., Halm, K. & Johnson, W. E. Tracking interspecies transmission and long-term evolution of an ancient retrovirus using the genomes of modern mammals. *eLife* **5**, e12704 (2016).
This paper describes the use of ERV loci to retrace the origins and global spread of an ancient gammaretrovirus among mammals between 15 million and 33 million years ago, spanning the late Oligocene and early Miocene epochs.
50. Katzourakis, A. et al. Discovery of prosimian and afrotherian foamy viruses and potential cross species transmissions amidst stable and ancient mammalian co-evolution. *Retrovirology* **11**, 61 (2014).
51. Escalera-Zamudio, M. et al. A novel endogenous betaretrovirus in the common vampire bat (Desmodus rotundus) suggests multiple independent infection and cross-species transmission events. *J. Virol.* **89**, 5180–5184 (2015).
52. Zhuo, X. & Feschotte, C. Cross-species transmission and differential fate of an endogenous retrovirus in three mammal lineages. *PLOS Pathog.* **11**, e1005279 (2015).
53. Holmes, E. C. The evolution of endogenous viral elements. *Cell Host Microbe* **10**, 368–377 (2011).
54. Kamath, P. L. et al. The population history of endogenous retroviruses in mule deer (Odocoileus hemionus). *J. Hered.* **105**, 173–187 (2014).
55. Greenwood, A. D., Ishida, Y., O'Brien, S. P., Roca, A. L. & Eiden, M. V. Transmission, evolution, and endogenization: lessons learned from recent retroviral invasions. *Microbiol. Mol. Biol. Rev.* **82**, e00044–17 (2018).
56. Lee, A., Nolan, A., Watson, J. & Tristem, M. Identification of an ancient endogenous retrovirus, predating the divergence of the placental mammals. *Phil. Trans. R. Soc. B Biol. Sci.* **368**, 20120503 (2013).
57. Blanco-Melo, D., Gifford, R. J. & Bieniasz, P. D. Co-option of an endogenous retrovirus envelope for host defense in hominid ancestors. *eLife* **6**, 11 (2017).
This study uses ancestral node reconstruction to establish that an intact *env* gene in the human genome can mediate superinfection interference and may have functioned to restrict entry of an ancient exogenous virus.
58. Blanco-Melo, D., Gifford, R. J. & Bieniasz, P. D. Reconstruction of a replication-competent ancestral murine endogenous retrovirus-L. *Retrovirology* **15**, 34 (2018).
This paper reports on the resurrection and experimental investigation of an ancient, extinct retrovirus using ancestral node reconstruction. This retrovirus is the oldest ERV (ERV-L) successfully reconstructed so far.
59. Johnson, W. E. & Coffin, J. M. Constructing primate phylogenies from ancient retrovirus sequences. *Proc. Natl Acad. Sci. USA* **96**, 10254–10260 (1999).
60. Martins, H. & Villesen, P. Improved integration time estimation of endogenous retroviruses with phylogenetic data. *PLOS ONE* **6**, e14745 (2011).
61. Dangel, A. W., Baker, B. J., Mendoza, A. R. & Yu, C. Y. Complement component C4 gene intron 9 as a phylogenetic marker for primates: long terminal repeats of the endogenous retrovirus ERV-K(C4) are a molecular clock of evolution. *Immunogenetics* **42**, 41–52 (1995).
62. Magiorkinis, G., Blanco-Melo, D. & Belshaw, R. The decline of human endogenous retroviruses: extinction and survival. *Retrovirology* **12**, 8 (2015).
63. Wildschutte, J. H. et al. Discovery of unfixed endogenous retrovirus insertions in diverse human populations. *Proc. Natl Acad. Sci. USA* **113**, E2326–E2334 (2016).
This study capitalizes on human genomic variation captured in databases, such as the 1000 Genomes Project, to detect and describe rare, unfixed HERV-K(HML-2) loci in the human population.
64. Subramanian, R. P., Wildschutte, J. H., Russo, C. & Coffin, J. M. Identification, characterization, and comparative genomic distribution of the HERV-K (HML-2) group of human endogenous retroviruses. *Retrovirology* **8**, 90 (2011).
65. Bhardwaj, N., Montesion, M., Roy, F. & Coffin, J. M. Differential expression of HERV-K (HML-2) proviruses in cells and virions of the teratocarcinoma cell line Tera-1. *Viruses* **7**, 939–968 (2015).
66. Domansky, A. N. et al. Solitary HERV-K LTRs possess bi-directional promoter activity and contain a negative regulatory element in the U5 region. *FEBS Lett.* **472**, 191–195 (2000).
67. Boeke, J. D., Garfinkel, D. J., Styles, C. A. & Fink, G. R. Ty elements transpose through an RNA intermediate. *Cell* **40**, 491–500 (1985).
68. Heidmann, T., Heidmann, O. & Nicolas, J. F. An indicator gene to demonstrate intracellular

transposition of defective retroviruses. *Proc. Natl Acad. Sci. USA* **85**, 2219–2223 (1988).

69. Esnault, C. et al. APOBEC3G cytidine deaminase inhibits retrotransposition of endogenous retroviruses. *Nature* **433**, 430–433 (2005).
70. Heslin, D. J. et al. A single amino acid substitution in a segment of the CA protein within Gag that has similarity to human immunodeficiency virus type 1 blocks infectivity of a human endogenous retrovirus K provirus in the human genome. *J. Virol.* **83**, 1105–1114 (2009).
71. Chudak, C. et al. Identification of late assembly domains of the human endogenous retrovirus-K(HML-2). *Retrovirology* **10**, 140 (2013).
72. Hanke, K. et al. Reconstitution of the ancestral glycoprotein of human endogenous retrovirus k and modulation of its functional activity by truncation of the cytoplasmic domain. *J. Virol.* **83**, 12790–12800 (2009).
73. Robinson, L. R. & Whelan, S. P. J. Infectious entry pathway mediated by the human endogenous retrovirus K envelope protein. *J. Virol.* **90**, 3640–3649 (2016).
74. Robinson-McCarthy, L. R. et al. Reconstruction of the cell entry pathway of an extinct virus. *PLOS Pathog.* **14**, e1007123 (2018).
This paper and that of Robinson and Whelan (2016) use an infectious rhabdovirus vesicular stomatitis virus (VSV) engineered to express an ancient Env protein in place of the VSVG protein to dissect the entry pathway of an ancient human endogenous retrovirus.
75. Soll, S. J., Neil, S. J. D. & Bieniasz, P. D. Identification of a receptor for an extinct virus. *Proc. Natl Acad. Sci. USA* **107**, 19496–19501 (2010).
76. Kaiser, S. M., Malik, H. S. & Emerman, M. Restriction of an extinct retrovirus by the human TRIM5alpha antiviral protein. *Science* **316**, 1756–1758 (2007).
77. Dewannieux, M. et al. Identification of an infectious progenitor for the multiple-copy HERV-K human endogenous retroelements. *Genome Res.* **16**, 1548–1556 (2006).
78. Lee, Y. N. & Bieniasz, P. D. Reconstitution of an infectious human endogenous retrovirus. *PLOS Pathog.* **3**, e10 (2007).
This paper and that of Dewannieux et al. (2006) describe the first successful reconstructions of functional infectious human endogenous retrovirus particles, in both cases on the basis of the HERV-K (HML2) family of ERV loci.
79. Lee, Y. N., Malim, M. H. & Bieniasz, P. D. Hypermutation of an ancient human retrovirus by APOBEC3G. *J. Virol.* **82**, 8762–8770 (2008).
80. Brady, T. et al. Integration target site selection by a resurrected human endogenous retrovirus. *Genes Dev.* **23**, 633–642 (2009).
This study describes the first global analysis of integration site preferences for an ancient, reconstituted endogenous retrovirus (HERV–Kcon), enabling direct comparison of integration site preferences to the locations of fixed HERV-K(HML2) loci in the human genome.
81. Gould, S. J. & Vrba, E. S. Exaptation—a missing term in the science of form. *Paleobiology* **8**, 4–15 (2016).
82. McClintock, B. Controlling elements and the gene. *Cold Spring Harb. Symp. Quant. Biol.* **21**, 197–216 (1956).
83. Britten, R. J. & Davidson, E. H. Gene regulation for higher cells: a theory. *Science* **165**, 349–357 (1969).
84. Nethe, M., Berkhout, B. & van der Kuyl, A. C. Retroviral superinfection resistance. *Retrovirology* **2**, 52 (2005).
85. Sommerfelt, M. A. & Weiss, R. A. Receptor interference groups of 20 retroviruses plating on human cells. *Virology* **176**, 58–69 (1990).
86. Malfavon-Borja, R. & Feschotte, C. Fighting fire with fire: endogenous retrovirus envelopes as restriction factors. *J. Virol.* **89**, 4047–4050 (2015).
87. Bolze, P.-A., Mommert, M. & Mallet, F. Contribution of syncytins and other endogenous retroviral envelopes to human placenta pathologies. *Prog. Mol. Biol. Transl Sci.* **145**, 111–162 (2017).
88. Dupressoir, A., Lavialle, C. & Heidmann, T. From ancestral infectious retroviruses to bona fide cellular genes: role of the captured syncytins in placentation. *Placenta* **33**, 663–671 (2012).
89. Cornelis, G. et al. Retroviral envelope gene captures and syncytin exaptation for placentation in marsupials. *Proc. Natl Acad. Sci. USA* **112**, E487–E496 (2015).
90. Cornelis, G. et al. An endogenous retroviral envelope syncytin and its cognate receptor identified in the viviparous placental Mabuya lizard. *Proc. Natl Acad. Sci. USA* **114**, E10991–E11000 (2017).
This paper gives the first description of a syncytin in a nonmammalian species.
91. Dupressoir, A. et al. 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 (2011).
92. Johnson, W. E. Rapid adversarial co-evolution of viruses and cellular restriction factors. *Curr. Top. Microbiol. Immunol.* **371**, 123–151 (2013).
93. Meyerson, N. R. & Sawyer, S. L. Two-stepping through time: mammals and viruses. *Trends Microbiol.* **19**, 286–294 (2011).
94. Robinson, H. L., Astrin, S. M., Senior, A. M. & Salazar, F. H. Host susceptibility to endogenous viruses: defective, glycoprotein-expressing proviruses interfere with infections. *J. Virol.* **40**, 745–751 (1981).
95. Ikeda, H. & Odaka, T. A cell membrane 'gp70' associated with Fv-4 gene: immunological characterization, and tissue and strain distribution. *Virology* **133**, 65–76 (1984).
96. Gardner, M. B., Kozak, C. A. & O'Brien, S. J. The Lake Casitas wild mouse: evolving genetic resistance to retroviral disease. *Trends Genet.* **7**, 22–27 (1991).
97. Kozak, C. A., Gromet, N. J., Ikeda, H. & Buckler, C. E. A unique sequence related to the ecotropic murine leukemia virus is associated with the Fv-4 resistance gene. *Proc. Natl Acad. Sci. USA* **81**, 834–837 (1984).
98. Inaguma, Y., Yoshida, T. & Ikeda, H. Scheme for the generation of a truncated endogenous murine leukaemia virus, the Fv-4 resistance gene. *J. Gen. Virol.* **73**, 1925–1930 (1992).
99. Jung, Y. T., Lyu, M. S., Buckler-White, A. & Kozak, C. A. Characterization of a polytropic murine leukemia virus proviral sequence associated with the virus resistance gene Rmcf of DBA/2 mice. *J. Virol.* **76**, 8218–8224 (2002).
100. Wu, T., Yan, Y. & Kozak, C. A. Rmcf2, a xenotropic provirus in the Asian mouse species Mus castaneus, blocks infection by polytropic mouse gammaretroviruses. *J. Virol.* **79**, 9677–9684 (2005).
101. Ito, J. et al. Refrex-1, a soluble restriction factor against feline endogenous and exogenous retroviruses. *J. Virol.* **87**, 12029–12040 (2013).
102. Sugimoto, J., Sugimoto, M., Bernstein, H., Jinno, Y. & Schust, D. A novel human endogenous retroviral protein inhibits cell-cell fusion. *Sci. Rep.* **3**, 1462 (2013).
103. Villesen, P., Aagaard, L., Wiuf, C. & Pedersen, F. S. Identification of endogenous retroviral reading frames in the human genome. *Retrovirology* **1**, 32 (2004).
104. de Parseval, N., Lazar, V., Casella, J.-F., Bénit, L. & Heidmann, T. Survey of human genes of retroviral origin: identification and transcriptome of the genes with coding capacity for complete envelope proteins. *J. Virol.* **77**, 10414–10422 (2003).
105. Young, G. R. et al. HIV-1 infection of primary CD4^{+} T cells regulates the expression of specific human endogenous retrovirus HERV-K (HML-2) elements. *J. Virol.* **92**, e01507–17 (2018).
106. Terry, S. N. et al. Expression of HERV-K108 envelope interferes with HIV-1 production. *Virology* **509**, 52–59 (2017).
107. Henzy, J. E., Gifford, R. J., Kenaley, C. P. & Johnson, W. E. An intact retroviral gene conserved in Spiny-rayed fishes for over 100 My. *Mol. Biol. Evol.* **34**, 634–639 (2017).
This paper describes what may be the oldest reported intact retroviral *env* gene, which inserted between 109 million and 140 million years ago and is shared by thousands of species of modern fish.
108. Heidmann, O. et al. HEMO, an ancestral endogenous retroviral envelope protein shed in the blood of pregnant women and expressed in pluripotent stem cells and tumors. *Proc. Natl Acad. Sci. USA* **114**, E6642–E6651 (2017).
This paper describes the discovery and functional characterization of an unusual ERV-encoded Env expressed as a secreted protein in placental tissues and in the blood of pregnant women.
109. Barnett, A. L., Davey, R. A. & Cunningham, J. M. Modular organization of the Friend murine leukemia virus envelope protein underlies the mechanism of infection. *Proc. Natl Acad. Sci. USA* **98**, 4113–4118 (2001).
110. Brody, B. A., Rhee, S. S. & Hunter, E. Postassembly cleavage of a retroviral glycoprotein cytoplasmic domain removes a necessary incorporation signal and activates fusion activity. *J. Virol.* **68**, 4620–4627 (1994).
111. Rein, A., Mirro, J., Haynes, J. G., Ernst, S. M. & Nagashima, K. Function of the cytoplasmic domain of a retroviral transmembrane protein: p15E-p2E cleavage activates the membrane fusion capability of the murine leukemia virus Env protein. *J. Virol.* **68**, 1773–1781 (1994).
112. Taylor, G. M., Gao, Y. & Sanders, D. A. Fv-4: identification of the defect in Env and the mechanism of resistance to ecotropic murine leukemia virus. *J. Virol.* **75**, 11244–11248 (2001).
113. Ito, J., Baba, T., Kawasaki, J. & Nishigaki, K. Ancestral mutations acquired in refrex-1, a restriction factor against feline retroviruses, during its cooption and domestication. *J. Virol.* **90**, 1470–1485 (2015).
114. Bénit, L., Calteau, A. & Heidmann, T. Characterization of the low-copy HERV-Fc family: evidence for recent integrations in primates of elements with coding envelope genes. *Virology* **312**, 159–168 (2003).
115. Bonnaud, B. et al. Evidence of selection on the domesticated ERVWE1 env retroviral element involved in placentation. *Mol. Biol. Evol.* **21**, 1895–1901 (2004).
116. Nakaya, Y. & Miyazawa, T. The roles of syncytin-like proteins in ruminant placentation. *Viruses* **7**, 2928–2942 (2015).
117. Goff, S. P. in *Fields Virology* (eds Knipe, D. M. & Howley, P. M.) 6th edn 1424–1473 (Lippincott Williams and Wilkins, 2013).
118. Marco, A. & Marín, I. CGIN1: a retroviral contribution to mammalian genomes. *Mol. Biol. Evol.* **26**, 2167–2170 (2009).
119. Best, S., Le Tissier, P., Towers, G. & Stoye, J. P. Positional cloning of the mouse retrovirus restriction gene Fv1. *Nature* **382**, 826–829 (1996).
120. Pincus, T., Hartley, J. W. & Rowe, W. P. A major genetic locus affecting resistance to infection with murine leukemia viruses. I. Tissue culture studies of naturally occurring viruses. *J. Exp. Med.* **133**, 1219–1233 (1971).
121. Bénit, L. et al. Cloning of a new murine endogenous retrovirus, MuERV-L, with strong similarity to the human HERV-L element and with a gag coding sequence closely related to the Fv1 restriction gene. *J. Virol.* **71**, 5652–5657 (1997).
122. Boso, G., Buckler-White, A. & Kozak, C. A. Ancient evolutionary origin and positive selection of the retroviral restriction factor Fv1 in muroid rodents. *J. Virol.* https://doi.org/10.1128/JVI.00850-18 (2018).
123. Young, G. R., Yap, M. W., Michaux, J. R., Steppan, S. J. & Stoye, J. P. Evolutionary journey of the retroviral restriction gene Fv1. *Proc. Natl Acad. Sci. USA* **115**, 10130–10135 (2018).
124. Yap, M. W., Colbeck, E., Ellis, S. A. & Stoye, J. P. Evolution of the retroviral restriction gene Fv1: inhibition of non-MLV retroviruses. *PLOS Pathog.* **10**, e1003968 (2014).
125. Mortuza, G. B. et al. High-resolution structure of a retroviral capsid hexameric amino-terminal domain. *Nature* **431**, 481–485 (2004).
126. Mura, M. et al. Late viral interference induced by transdominant Gag of an endogenous retrovirus. *Proc. Natl Acad. Sci. USA* **101**, 11117–11122 (2004).
127. Arnaud, F., Murcia, P. R. & Palmarini, M. Mechanisms of late restriction induced by an endogenous retrovirus. *J. Virol.* **81**, 11441–11451 (2007).
128. Monde, K., Contreras-Galindo, R., Kaplan, M. H., Markovitz, D. M. & Ono, A. Human endogenous retrovirus K Gag coassembles with HIV-1 Gag and reduces the release efficiency and infectivity of HIV-1. *J. Virol.* **86**, 11194–11208 (2012).
129. Campillos, M., Doerks, T., Shah, P. K. & Bork, P. Computational characterization of multiple Gag-like human proteins. *Trends Genet.* **22**, 585–589 (2006).
130. Pastuzyn, E. D. et al. The neuronal gene Arc encodes a repurposed retrotransposon Gag protein that mediates intercellular RNA transfer. *Cell* **172**, 275–288 (2018).
131. Ashley, J. et al. Retrovirus-like Gag protein Arc1 binds RNA and traffics across synaptic boutons. *Cell* **172**, 262–274 (2018).
This paper and that of Pastuzyn et al. (2018) describe neuronal proteins that are related to retroviral Gag proteins and that form capsid-like structures that package RNA and are released extracellularly.
132. Bernard, D. et al. Identification and characterization of a novel retroviral-like aspartic protease specifically expressed in human epidermis. *J. Invest. Dermatol.* **125**, 278–287 (2005).
133. Katzourakis, A., Gifford, R. J., Tristem, M., Gilbert, M. T. P. & Pybus, O. G. Macroevolution of complex retroviruses. *Science* **325**, 1512–1512 (2009).
134. Frankel, W. N., Rudy, C., Coffin, J. M. & Huber, B. T. Linkage of Mls genes to endogenous mammary tumour viruses of inbred mice. *Nature* **349**, 526–528 (1991).
135. Ross, S. R. Mouse mammary tumor virus molecular biology and oncogenesis. *Viruses* **2**, 2000–2012 (2010).
136. Golovkina, T. V., Chervonsky, A., Dudley, J. P. & Ross, S. R. Transgenic mouse mammary tumor virus

superantigen expression prevents viral infection. *Cell* **69**, 637–645 (1992).

137. Mertz, J. A., Simper, M. S., Lozano, M. M., Payne, S. M. & Dudley, J. P. Mouse mammary tumor virus encodes a self-regulatory RNA export protein and is a complex retrovirus. *J. Virol.* **79**, 14737–14747 (2005).

138. Hofacre, A. & Fan, H. Jaagsiekte sheep retrovirus biology and oncogenesis. *Viruses* **2**, 2618–2648 (2010).

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

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

141. Yang, J. et al. An ancient family of human endogenous retroviruses encodes a functional homolog of the HIV-1 Rev protein. *Proc. Natl Acad. Sci. USA* **96**, 13404–13408 (1999).

142. Yang, Z. & Bielawski, J. Statistical methods for detecting molecular adaptation. *Trends Ecol. Evol. (Amst.)* **15**, 496–503 (2000).

143. Katzourakis, A. & Gifford, R. J. Endogenous viral elements in animal genomes. *PLOS Genet.* **6**, e1001191 (2010).

144. Aswad, A. & Katzourakis, A. Paleovirology and virally derived immunity. *Trends Ecol. Evol. (Amst.)* **27**, 627–636 (2012).

145. Kozak, C. A. Origins of the endogenous and infectious laboratory mouse gammaretroviruses. *Viruses* **7**, 1–26 (2014).

146. Anai, Y. et al. Infectious endogenous retroviruses in cats and emergence of recombinant viruses. *J. Virol.* **86**, 8634–8644 (2012).

147. Jern, P. & Coffin, J. M. Effects of retroviruses on host genome function. *Annu. Rev. Genet.* **42**, 709–732 (2008).

148. Cohen, C. J., Lock, W. M. & Mager, D. L. Endogenous retroviral LTRs as promoters for human genes: a critical assessment. *Gene* **448**, 105–114 (2009).

149. Thompson, P. J., Macfarlan, T. S. & Lorincz, M. C. Long terminal repeats: from parasitic elements to building blocks of the transcriptional regulatory repertoire. *Mol. Cell* **62**, 766–776 (2016).

150. Fuentes, D. R., Swigut, T. & Wysocka, J. Systematic perturbation of retroviral LTRs reveals widespread long-range effects on human gene regulation. *eLife* **7**, e35989 (2018).
This study uses a modified CRISPR system to induce or silence multiple HERV-K(HML2) LTRs in parallel, revealing long-range effects on expression of hundreds of genes.

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

152. Kapusta, A. et al. Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. *PLOS Genet.* **9**, e1003470 (2013).

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

154. Chuong, E. B., Elde, N. C. & Feschotte, C. Regulatory activities of transposable elements: from conflicts to benefits. *Nat. Rev. Genet.* **18**, 71–86 (2017).

155. Lynch, V. J. A copy-and-paste gene regulatory network. *Science* **351**, 1029–1030 (2016).

156. Khodosevich, K., Lebedev, Y. & Sverdlov, E. Endogenous retroviruses and human evolution. *Comp. Funct. Genomics* **3**, 494–498 (2002).

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

158. Wang, T. et al. Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53. *Proc. Natl Acad. Sci. USA* **104**, 18613–18618 (2007).
This paper and that of Chuong et al. (2016) reveal that co-option of ERV LTRs contributed to concerted evolution of interferon-regulated gene networks and many p53 regulated genes, respectively.

159. Ito, J. et al. Systematic identification and characterization of regulatory elements derived from human endogenous retroviruses. *PLOS Genet.* **13**, e1006883 (2017).

160. Simonti, C. N., Pavlicev, M. & Capra, J. A. Transposable element exaptation into regulatory regions is rare, influenced by evolutionary age, and subject to pleiotropic constraints. *Mol. Biol. Evol.* **34**, 2856–2869 (2017).

161. Carroll, S. B. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell* **134**, 25–36 (2008).

162. Monteiro, A. & Podlaha, O. Wings, horns, and butterfly eyespots: how do complex traits evolve? *PLOS Biol.* **7**, e37 (2009).

163. Lesbats, P., Engelman, A. N. & Cherepanov, P. Retroviral DNA integration. *Chem. Rev.* **116**, 12730–12757 (2016).

164. Hughes, J. F. & Coffin, J. M. Human endogenous retroviral elements as indicators of ectopic recombination events in the primate genome. *Genetics* **171**, 1183–1194 (2005).

165. Kijima, T. E. & Innan, H. On the estimation of the insertion time of LTR retrotransposable elements. *Mol. Biol. Evol.* **27**, 896–904 (2010).

166. Trombetta, B., Fantini, G., D'Atanasio, E., Sellitto, D. & Cruciani, F. Evidence of extensive non-allelic gene conversion among LTR elements in the human genome. *Sci. Rep.* **6**, 28710 (2016).

167. Schlesinger, S. & Goff, S. P. Retroviral transcriptional regulation and embryonic stem cells: war and peace. *Mol. Cell. Biol.* **35**, 770–777 (2015).

168. Cullen, B. R., Lomedico, P. T. & Ju, G. Transcriptional interference in avian retroviruses — implications for the promoter insertion model of leukaemogenesis. *Nature* **307**, 241–245 (1984).

169. Hughes, J. F. & Coffin, J. M. Human endogenous retrovirus K solo-LTR formation and insertional polymorphisms: implications for human and viral evolution. *Proc. Natl Acad. Sci. USA* **101**, 1668–1672 (2004).

170. Belshaw, R. et al. Rate of recombinational deletion among human endogenous retroviruses. *J. Virol.* **81**, 9437–9442 (2007).

171. Martin, J., Kabat, P., Herniou, E. & Tristem, M. Characterization and complete nucleotide sequence of an unusual reptilian retrovirus recovered from the order Crocodylia. *J. Virol.* **76**, 4651–4654 (2002).

172. Henzy, J. E., Gifford, R. J., Johnson, W. E. & Coffin, J. M. A novel recombinant retrovirus in the genomes of modern birds combines features of avian and mammalian retroviruses. *J. Virol.* **88**, 2398–2405 (2014).

173. de Souza, F. S. J., Franchini, L. F. & Rubinstein, M. Exaptation of transposable elements into novel *cis*-regulatory elements: is the evidence always strong? *Mol. Biol. Evol.* **30**, 1239–1251 (2013).

174. Hobbs, M. et al. Long-read genome sequence assembly provides insight into ongoing retroviral invasion of the koala germline. *Sci. Rep.* **7**, 15838 (2017).

175. Montesion, M., Bhardwaj, N., Williams, Z. H., Kuperwasser, C. & Coffin, J. M. Mechanisms of HERV-K (HML-2) transcription during human mammary epithelial cell transformation. *J. Virol.* **92**, e01258–17 (2018).

176. Niu, D. et al. Inactivation of porcine endogenous retrovirus in pigs using CRISPR-Cas9. *Science* **357**, 1303–1307 (2017).
This paper describes the parallel inactivation of two dozen related porcine ERV (PERV) loci in a single fetal fibroblast cell using a customized CRISPR–Cas9 protocol followed by nuclear transfer to create a line of pigs free of functional PERV loci.

177. Ellermann, V. & Bang, O. Experimentelle leukämie bei hühnern [German]. *Zentralbl. Bakteriol. Parasitenkd. Infectionskr. Hyg. Abt. Orig.* **46**, 595–609 (1908).

178. Rous, P. A sarcoma of the fowl transmissible by an agent separable from the tumor cells. *J. Exp. Med.* **13**, 397–411 (1911).

179. Dietrich, M. R. in *Evolutionary Genetics: Concepts and Case Studies* (eds Wolf, J. B. & Fox, C. W.) (Oxford Univ. Press, 2006).

180. Krupovic, M. et al. Ortervirales: new virus order unifying five families of reverse-transcribing viruses. *J. Virol.* **92**, e00515–18 (2018).

Acknowledgements

The author thanks J. Butler, B. Howell and the organizers of the 2018 Boston College Intersections Villa Faculty Writing Retreat for the opportunity to complete major portions of this manuscript; R. Gifford, L. Mulder and J. Henzy for helpful discussions; S. Whelan and V. Simon for providing offices for writing while on sabbatical leave at Harvard Medical School and the Icahn School of Medicine at Mount Sinai, respectively. Work in the author's laboratory is supported by grants from the US National Institutes of Health (AI083118) and the US Department of Defense/Congressionally Directed Medical Research Programs (PR172274).

Competing interests

The author declares no competing interests.

Publisher's note

Reviewer information

Nature Reviews Microbiology thanks A. Dupressoir, C. Feschotte, J. Frank and other anonymous reviewer(s) for their contribution to the peer review of this work.