Endogenous viruses: insights into viral evolution and impact on host biology

C. Feschotte, C. Gilbert

Endogenous viruses: insights into viral evolution and impact on host biology

Cédric Feschotte[1] and Clément Gilbert[2]

Abstract | Recent studies have uncovered myriad viral sequences that are integrated or 'endogenized' in the genomes of various eukaryotes. Surprisingly, it appears that not just retroviruses but almost all types of viruses can become endogenous. We review how these genomic 'fossils' offer fresh insights into the origin, evolutionary dynamics and structural evolution of viruses, which are giving rise to the burgeoning field of palaeovirology. We also examine the multitude of ways through which endogenous viruses have influenced, for better or worse, the biology of their hosts. We argue that the conflict between hosts and viruses has led to the invention and diversification of molecular arsenals, which, in turn, promote the cellular co-option of endogenous viruses.

Horizontally
In the context of genetic information, horizontal transmission is the transfer of genetic material by means other than sex.

Vertical transmission
Sexual transmission of genetic material from parent to offspring.

Fixation
A mutation reaches fixation when it is present in all individuals of a given species.

Transposable elements
Pieces of DNA (typically genomic elements) that are able to move from one locus to another, often duplicating themselves in the process.

[1]*Department of Biology, University of Texas, Arlington, Texas 76016, USA.*
[2]*Université de Poitiers, UMR CNRS 7267, Ecologie et Biologie des Interactions, Equipe Ecologie Evolution Symbiose, France.*
e-mails: cedric@uta.edu; clement.gilbert@univ-poitiers.fr
doi:10.1038/nrg3199

Viruses are the most numerous and diverse genetic entities on earth. They are environmentally ubiquitous and are capable of infecting organisms from all three domains of life, as well as other viruses[1]. Through various interactions, viruses have profoundly influenced the evolution of cellular life ever since its origin[2,3]. In addition to the infectious properties of viruses, which enable them to spread horizontally between individuals and across species, many viruses can also become part of the genetic material of their host species, a process that is called endogenization. Such endogenous viral elements (EVEs)[4] result from the chromosomal integration of viral DNA (or DNA copies of viral RNA) in the host germ cells, which allows for vertical transmission and potential fixation in the host population. Although endogenization was for a long time thought to be limited to retroviruses, recent studies have shown that all major types of eukaryotic viruses can give rise to EVEs.

The recent availability of large numbers of eukaryotic genome sequences, combined with an enhanced bioinformatics ability to detect and date ancient EVEs, has laid the foundation for the emerging field of palaeovirology[5]. After describing the discovery of EVEs, we discuss how this burgeoning area has already generated an important shift in our perception of the evolutionary origins and dynamics of several notorious groups of eukaryotic viruses.

Another exciting facet of palaeovirology is that it uncovers how EVEs have contributed to the evolution of the host genome by introducing genetic variation and innovation, which we discuss in sections on mutation of the host genome, influences on host gene expression and domestication into new protein-coding genes with cellular functions. Although other types of invasive, parasitic genetic elements, such as transposable elements, have a similarly broad impact on their host[6], EVE sequences may have a greater propensity to be co-opted for certain regulatory and physiological functions. Most transposable elements are engaged in a long-term co-evolutionary relationship with their hosts, pressing them to evolve strategies to minimize their deleterious impact, which, in some cases, may reach a form of commensalism or symbiosis[6,7]. By contrast, most viruses assume a more destructive parasitic lifestyle, often leading to the death of infected cells, which sets them in an intense and perpetual conflict with their host. This arms race continuously selects for the emergence of genomic adaptations in viruses to manipulate the host cell machinery and to counteract antiviral defences. Ironically, as we will see, such viral weaponry often forms the foundation for the subsequent recruitment of EVEs to cellular functions.

Discoveries of endogenous viruses

Retroviruses are the only known eukaryotic viruses that require chromosomal integration for successful completion of their lifecycle. As such, they encode all of the enzymatic machinery that is necessary to perform this step autonomously. These properties probably explain why endogenous retroviruses (ERVs) were the first EVEs to be characterized and why they still make up the vast majority of EVEs that are recognized today. However, it has been known for some time that other viruses can be endogenized. For example, endogenous

Table 1 | **EVEs identified in eukaryotic genomes**

Group/type	Family or genus	Taxa	Number per haploid genome	Refs
Group I/dsDNA	Baculovirus	Insects	Unknown (hybridization data, no sequencing)	124
Group I/dsDNA	Herpesviridae	Humans	1	125,126
Group I/dsDNA	Nudivirus	Parasitic wasps	Several	127
Group I/dsDNA	Phycodnaviridae	Brown algae	1	128,129
Group II/ssDNA	Circoviridae	Mammals	1 to 2	4,20,130
Group II/ssDNA	Geminiviridae	Tomentosae (tobacco and three other species)	5 to 120	130–132
Group II/ssDNA	Parvoviridae	Mammals; shrimp	1 to 3	4,20,89, 133,134
Group III/dsRNA	Partitiviridae	Plants; arthropods; Protozoa	1 to 4	135
Group III/dsRNA	Reovirus	*Aedes* spp. mosquitoes	1	4
Group III/dsRNA	Totiviridae	Fungi; plants; ticks	1 to 6	16,135,136
Group IV/+ssRNA	Dicistroviridae	Honeybees	1	137
Group IV/+ssRNA	Flaviviridae	Medaka fish; mosquitoes	1 to 4	4,21,138, 139
Group IV/+ssRNA	Potyviridae	Grapes	Several	140
Group V/–ssRNA	Bornaviridae	Vertebrates	1 to 17	4,21,17
Group V/–ssRNA	Bunyaviridae	Ticks	14	4
Group V/–ssRNA	Filoviridae	Mammals	1 to 13	4,21,141
Group V/–ssRNA	Nyavirus	Zebrafish	6	21
Group V/–ssRNA	Orthomyxoviridae	Ticks	1	4
Group V/–ssRNA	Rhabdoviridae	Insects (ticks and mosquitoes)	1 to 28	4
Group VI/ssRNA-RT	Retroviridae	Vertebrates	Several hundreds to several hundreds of thousands	36
Group VII/dsDNA-RT	Hepadnavirus	Passerine birds	15	4,19
Group VII/dsDNA-RT	Pararetrovirus	Plants	A dozen to a thousand	8,11

+, positive sense; –, negative sense; RT, reverse transcriptase.

caulimoviruses (also called pararetroviruses) have been identified in plant genomes since the late 1990s[8]. Like retroviruses, caulimoviruses replicate via reverse transcription, and they may have originated from the fusion of a retrotransposon with an envelope (*env*)-like gene derived from a distinct virus[9,10]. Unlike retroviruses, however, plant caulimoviruses possess a double-stranded DNA genome, and they do not normally integrate into the host chromosome[11]. Accordingly, endogenous caulimoviruses are thought to result from fortuitous integration into the genome of germ cells (or meristematic cell progenitors of germ cells), and they generally correspond to partial viral genomes that are incapable of further replication[11]. In rare cases, some of the integrated caulimoviruses appear to have retained the capacity to produce infectious particles but, intriguingly, only under stress conditions[12,13].

Although a few instances of endogenization of various other viruses (shown in brackets) have been known for some time in some plants (Geminiviridae and Potyviridae), arthropods (Baculoviridae and Flaviviridae) and algae (Phycodnaviridae), it is only in the past few years that the pervasiveness of viral endogenization has been appreciated. This progress was largely fuelled by the advent of whole-genome sequencing and bioinformatics, which led to the identification of dozens of non-retroviral EVEs in various eukaryotes (TABLE 1). As for caulimoviruses, endogenization of these non-retroviral EVEs is thought to result from accidental chromosomal integration events[14]. Little is known about the underlying molecular mechanisms, but several studies have identified sequence signatures at the EVE–host genome junction that point to retroposition events, suggesting involvement of the enzymatic machinery encoded by retrotransposons residing in the host genome[15–17]. In addition, it has been shown that viral DNA can be used to patch double-strand breaks in the host chromosome by non-homologous end joining[18], which might provide a mechanism for the endogenization of viruses with DNA genomes or DNA replication intermediates (for example, REF. 19). Finally, it has been proposed that, for some viruses, integration of the viral genome could be facilitated by viral proteins[20]. The discovery of a wide range of non-retroviral EVEs suggests that almost any major type of eukaryotic virus may be endogenized, sometimes in multiple hosts

Reverse transcription
Synthesis of DNA from an RNA template.

Retrotransposon
Mobile intracellular genetic elements that replicate via reverse transcription of an RNA intermediate.

Envelope
(Env). A glycoprotein encoded by many viruses that binds to host receptors located on the cell surface in order to promote viral entry.

Non-homologous end joining
A DNA double-strand break repair pathway that does not make use of a template and is therefore intrinsically error-prone.

independently and over wide evolutionary periods[4] (TABLE 1). These findings also reveal that some viruses have had and may still have a much broader host range than was previously appreciated. For example, endogenous hepadnaviruses[19] and filoviruses[21] have been found in passerine birds and marsupials, respectively, which are thus far not known to be infected by these viruses. The discovery of EVEs may also be helpful to identify candidate reservoir species of zoonotic viruses. In this regard, the apparent over-representation of filovirus-like EVEs in rodent, insectivore and bat genomes compared with those of other mammals is intriguing given that these taxa rank among the most likely candidate reservoirs for their notorious modern relatives, such as Ebola and Marburg viruses[21].

Uncovering the deep evolution of modern viruses

EVEs fill in gaps in viral evolution. In spite of the numerical and ecological importance of viruses in the biosphere and their frequently devastating impact on human health, our understanding of viral evolution remains fragmentary. Regarding the early origins of viruses and the nature of their evolutionary relationships with eukaryotes, bacteria and archaea, the emerging consensus posits that viruses may be older than cellular life and may have actually triggered its emergence or at least may have profoundly influenced the early evolution of cells[2,3,14]. The inferred ancient origin of viruses is based, in part, on the observation that several genes involved in DNA or RNA replication and capsid assembly are structurally more similar among diverse viruses infecting the three domains of life than between viruses and any cellular organisms[2,22].

In addition to these inferences on the earliest steps of viral evolution, several facets of the evolution of modern viruses have been extensively investigated. A number of studies have produced robust estimates of the ages, evolutionary rates and epidemiological dynamics of many viruses that currently infect humans, domestic animals and crops[23,24]. According to these studies, the evolutionary timescale of modern viruses lies within the past million years: most human RNA viruses emerged less than 1,000 years ago[24]. Hence, a huge gap in our understanding of viral evolution lies between these very recent events and the pre-cellular origins of viruses.

The study of EVEs that are closely related to, or even fall within the diversity of, currently circulating viruses offers an opportunity to start filling in this gap. EVEs can provide missing links for deciphering the structural evolution of viral genomes and the origin of new viral genes (BOX 1). In addition, various methods can be used to infer the integration times of EVEs, which, in turn, yield minimum ages for the family of modern viruses to which they belong (BOX 2). In many cases in which EVEs that are closely related to modern viruses could be dated, the minimum age inferred for the most recent ancestor of the viral family has turned out to be far older than was previously estimated using sequence data from circulating viruses. For example, the Circoviridae and Hepadnaviridae families were thought to be <500 and <30,000 years old, respectively[25,26], but EVEs from these two families have been dated at >40 million and >19 million years old, respectively[20,19]. Similarly, EVEs that are clearly related to Lentiviridae and Spumaviridae were traced back to >12 million and >100 million years ago, respectively: far older than had been inferred from sequence comparison of their modern relatives[27–29].

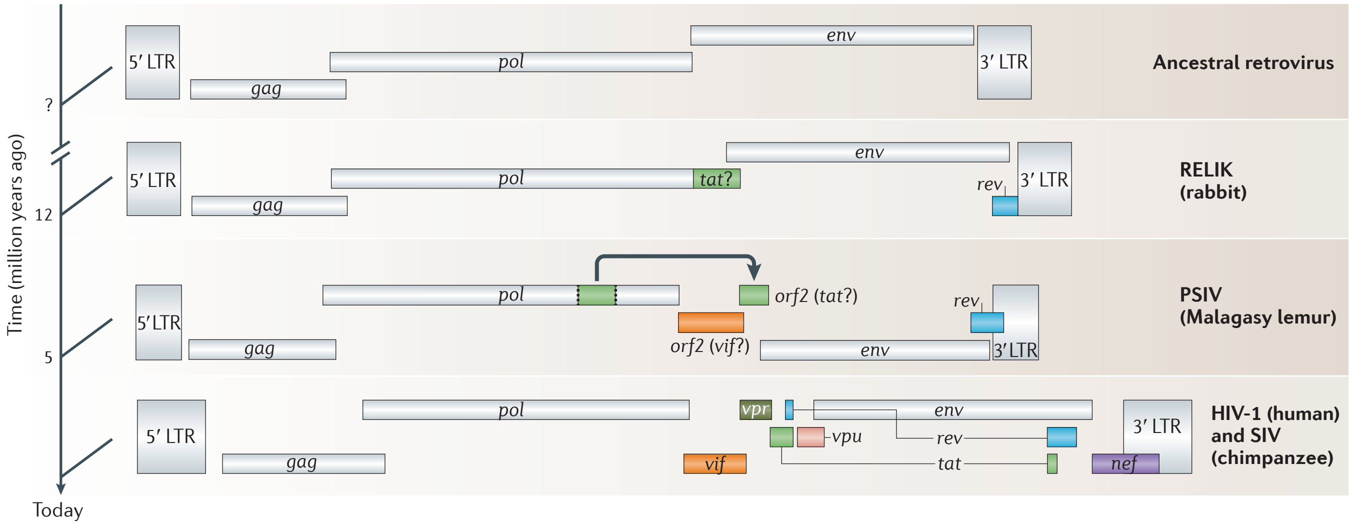
Substitution rates. Most exogenous viruses are characterized by extremely rapid substitution rates that are often three to six orders of magnitude faster than those of their host[30]. Typically, viral substitution rates are calculated using samples of modern viruses that have circulated over short periods of time spanning tens or hundreds of years (see the figure in BOX 2). The discovery of EVE sequences that are fossilized in genomes for millions of years but that are still related to and directly alignable to those of modern viruses offers an opportunity to derive viral substitution rates on a much deeper timescale (BOX 2). Surprisingly, such long-term viral substitution rates are considerably slower than short-term rates estimated using only modern viral sequences. For example, in the case of hepadnaviruses and begomoviruses, long-term substitution rates were found to be two to three orders of magnitude slower than short-term rates[19,31]. At first glance, it is tempting to explain this discrepancy by the fact that the substitution rate of EVEs dramatically plummets following endogenization, as EVE sequences become subject to the much slower mutation rate of the host genome (see the figure in BOX 2). However this 'mutational freezing' of the EVE sequence at the time of endogenization has essentially no bearing on the calculation of long-term viral rate because the number of substitutions accumulated at the host rate (indicated by the red triangles in the figure in BOX 2) represents a small fraction of the substitutions observed when the EVE and its closest modern viral relative are compared. That small number of mutations (accumulated at the host rate) can easily be estimated and subtracted in the calculation of long-term viral rate[19].

How then can we explain the vast discrepancy between short- and long-term rates of viral evolution? The phenomenon is reminiscent of the so-called time dependency of substitution rates observed in cellular organisms, where rates measured on a shallow timescale (for example, in pedigrees or populations) are invariably faster than rates measured over a deeper, geological timescale[32]. Several factors have been proposed to explain this incongruity[5,14,19,32,33]. For instance, it is possible that short-term rates are artificially inflated by a large proportion of slightly deleterious mutations, which are yet to be removed from the population through purifying selection[5,34]. To assess whether this effect may substantially bias short-term estimates of viral rates, it would be necessary to evaluate the lifespan of deleterious mutations in viral populations. Mutational saturation is a factor that could play in the other direction, leading to an underestimate of the genetic distances between EVEs and circulating viruses and thereby of long-term substitution rates.

Zoonotic
Describes a virus that can be transmitted between animals and humans or vice versa.

Mutational saturation
A given site in a DNA sequence is saturated when the number of observed or inferred mutations is lower than the number of mutations that truly occurred at this site.

Box 1 | **EVEs shed light on the structural evolution of viruses**



As representatives of viral families fossilized at various evolutionary time points, endogenous viral elements (EVEs) can provide insights into the origins of components found in modern viral genomes[99,100]. For example, the origin of the accessory genes of complex retroviruses (such as *nef* or *rev* in HIV) has long been a mystery. These genes have diverse functions (for example, regulation of proviral expression, export of retroviral mRNA from the nucleus to the cytoplasm and blocking host restriction factors) and — unlike the structural and enzymatic genes, *gag*, *pol* and *env*, which are common to all retroviruses (shown as grey rectangles in the figure) — they are only found in a subset of retroviral genera (shown as coloured rectangles in the figure).

Until recently, only the *vpx* gene that is present in HIV-2 and in various simian immunodeficiency viruses (SIVs) had been traced: it appears to have arisen by non-homologous recombination between two different SIVs[101–103]. The discovery of an endogenous lentivirus, PSIV, in the genome of prosimian primates (namely, Malagasy lemurs) helped to shed light on the origin of a second lentiviral accessory gene. This gene, called *orf2* (see the figure), shows substantial sequence similarity to the 3′ end of the primate lentiviruses reverse transcriptase domain, suggesting that it might have arisen by partial duplication of this domain, possibly via template jumping during reverse transcription[104]. Interestingly, *orf2* is of the same size and is located at the same position (within the *pol*–*env* intervening region) as the *tat* gene of modern lentiviruses, suggesting that it may be a *tat* orthologue. The *orf2* sequence, however, does not share any substantial similarity with known *tat* genes, which themselves are extremely diverged from each other and are barely alignable. Whether *orf2* and *tat* are truly homologous (that is, whether they have a common ancestor) therefore remains an open question that would need to be addressed by functional studies. Unlike *vpx*, which is only present in HIV-2 and in a subset of SIVs and must be of very recent origin, *tat* is present in all primate lentiviruses and in other lentiviruses (such as bovine immunodeficiency virus (BIV) from cows, equine infectious anaemia virus (EIAV) from horses and rabbit endogenous lentivirus type K (RELIK)), which suggests it emerged much earlier than *vpx*. Thus the link between *tat* and the 3′ end domain of the lentiviral reverse transcriptase in PSIV might be our best opportunity to trace the origin of *tat*. LTR, long terminal repeat.

The discordance observed between EVE-based studies and modern virus-based studies could also reflect genuine biological differences in the viruses under study. In particular, variations in the replication rate (which could possibly be caused by latency) or in the fidelity of viral polymerases over time[19] could have a substantial effect; for example, the fidelity of reverse transcriptases from diverse retroviruses is known to vary by 20-fold[35]. Finally, an interesting question is whether viral lineages are best characterized by a single rate with little variation through time or whether their evolution follows a more saltational mode that is typified by rapid changes in substitution rates that are directed by environmental or ecological variables, such as the colonization of new hosts[33] or the efficiency of the host's immune response. According to the saltational model, elevated short-term viral rates might reflect a transient state of adaptation associated with an intense conflict with a new host, whereas most of the remaining part of a virus evolutionary history would be characterized by a slower rate that is closer to the estimated long-term rate. In line with this scenario, we note that the sample of viruses used to infer short-term rates is strongly biased for zoonotic viruses that may conceivably have been caught in the midst of acute adaptation to their new host. Future efforts should be directed towards developing a simulation approach in order to define the conditions under which a virus could evolve at vastly different rates, depending on the timescale considered, and to assess whether these conditions are realistic or not. It will also be necessary to improve nucleotide substitution models and to continue characterizing new EVEs. Perhaps an ideal prospect to reconcile short- and long-term viral rates would be to identify EVEs with a range of ages spanning a large breadth of the evolutionary history of a viral family and to incorporate these genomic fossils in a phylodynamic analysis of extant viruses.

Latency
A period during which a virus replicates at a low rate without causing any symptoms to the host.

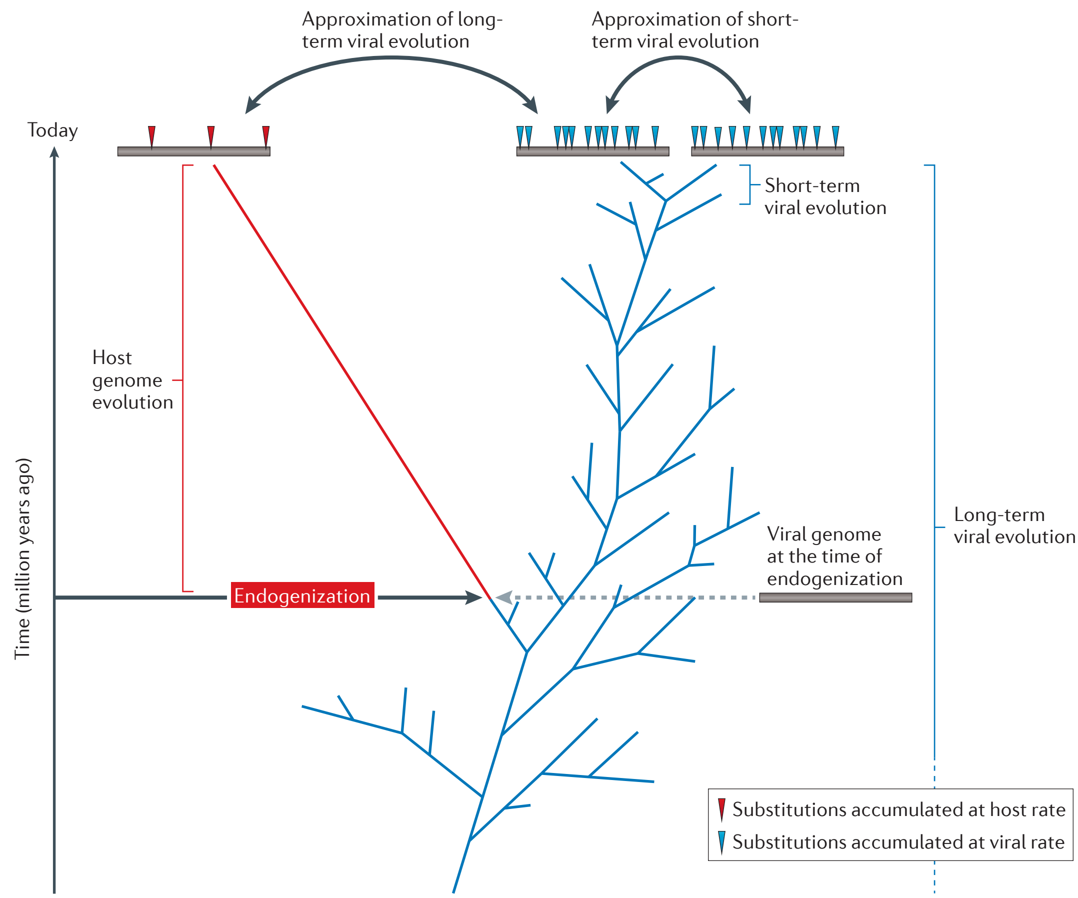
Saltational
A saltational change is a profound and rapid change in the evolutionary dynamics of a viral lineage.

Phylodynamic
The joint study of the epidemiological and evolutionary dynamics of a virus.

Host genome mutagenesis

Insertional mutagens. Population genetics predict that, for every fixed EVE insertion, thousands must have occurred in germ cells but were lost from the host population. Measuring the impact of this largely

Box 2 | **Dating EVEs and inferring long-term viral substitution rates**



Various methods can be applied to infer the age of endogenous viral elements (EVEs) and to calibrate the evolutionary timescale of modern virus families. A traditional way of estimating the integration time of endogenous retroviruses (ERVs) is to calculate the divergence between their 5′ and 3′ long terminal repeats (LTRs) and to divide this distance by the substitution rate of the host genome[105]. This calculation is based on the fact that the reverse transcription of a given retrovirus inherently generates a provirus that is flanked by identical 5′ and 3′LTRs following integration, such that the nucleotide differences that are observed between these LTRs are expected to result from mutations that occurred after integration at the neutral rate of the host genome. This method has been used to infer the age of many ERVs, including most of the ERV families present in the human genome, and various corrections have been proposed that take into account potential biases introduced by gene conversion and heterogeneity in substitution rates between LTRs[106,107]. A second method of inferring the age of EVEs relies on the frequency of stop codons in the various endogenous open reading frames[21]. Finally, EVEs can be dated by screening for the presence or absence of orthologous EVE loci in the genome of species that present a diverse degree of relatedness to the host species in which the EVE was detected[4,19,20,27,28]. The presence of an EVE locus at the same genomic location in two species indicates that integration predates the divergence of these species, and this provides a lower age boundary. Characterization of an 'empty site' (that is, genomic flanking regions without the EVE insertion) in a third more distantly related species can provide an upper age boundary. EVEs that have been dated using this approach can be sequenced in the two more distantly related species in which they were found, and another time estimate can be derived by dividing half of the distance separating these EVEs by the host neutral substitution rate. In instances in which the host neutral rate is based on a different palaeontological calibration point than the one used for inferring species divergence times, congruence between the two approaches solidifies age estimates[19]. After the age of EVEs has been estimated, it offers a calibration point to derive a long-term viral substitution rate. An interesting property of EVEs is that, after endogenization, they no longer mutate or evolve at the fast exogenous viral rate. Instead, they are replicated along with their host genome and evolve much more slowly than their exogenous counterparts. Consequently the sequence of an EVE as we see it today is very close to the sequence of the viral genome that became endogenized. It differs from this ancestral sequence by only a few mutations (shown by the red triangles in the figure). Long-term viral substitution rates can therefore be estimated by simply dividing the substitutions (shown by the blue triangles in the figure) between this approximate ancestral sequence (neglecting the few substitutions that occurred at the host rate) and its closest circulating counterpart by the age of the EVE (time of endogenization). To obtain a more accurate estimate of long-term substitution rate, it is possible in some instances to estimate the amount of substitutions accumulated at the host rate after endogenization (shown by the red triangles in the figure) and to remove this amount from the total distance separating EVEs from their closest exogenous counterparts[19].

invisible assault on the host species is hardly possible, but considering that nearly half a million viral insertions have reached fixation in the human genome alone, whereas many more have concurrently colonized other vertebrate lineages, it is easy to predict that viral integration represents a substantial source of natural genomic variation in vertebrate populations. ERVs possess a greater mutagenic potential than other types of EVE, in that each insertion event typically deposits a full-length provirus that is capable of producing new germline or somatic insertions for a long time after the initial endogenization event. Further genomic propagation of ERVs can occur either by reinfection or by intracellular retrotransposition. An ERV may undergo either autonomous replication using its own enzymatic machinery, or it may be complemented by enzymes produced by exogenous viruses or by other retroelements that coexist in the genome. There is evidence that all of these processes have contributed to the genomic expansion of human ERVs over extensive periods of primate evolution[36,37]. The persistent threat posed by the repeated infiltration of ERVs in vertebrates explains why ancient and highly adaptable host defence mechanisms have co-evolved to suppress ERV expression[38], notably in germ cells and early embryonic cells, where new insertions become inheritable[39] (BOX 3).

For unknown reasons, the rate of ERV insertion seems to have plummeted in recent human evolution. Consistent with a dearth of recent activity, ERV insertional polymorphisms are very rare in the human population, and less than a dozen full-length proviruses are currently known to be dimorphic[40,41]. They all belong to the human ERV-K (HERV-K) subfamily HML2, which is apparently the most recent ERV wave to have entered our genome, perhaps producing insertions as recently as 150,000 years ago[42]. These data stand in contrast to the high frequency of insertion polymorphisms generated by *Alu* and LINE-1, which are two retrotransposon families that are still transpositionally active in humans and that occasionally cause disease through insertional mutagenesis[43]. As of yet, there are no reports of *de novo* insertion of viral sequences directly causing inherited disease in humans. Furthermore, there is no evidence that any ERV currently catalogued in the reference human genome is capable of autonomous retrotransposition or of producing infectious particles, although it remains conceivable (BOX 4). Of course, it is not possible to rule out that some viruses regularly integrate into the genome of human germ cells, but perhaps at too low a frequency or with effects that are too deleterious to spread in the general population.

In contrast to humans, *de novo* ERV insertions are common in the laboratory mouse and, indeed, they account for ~10% of all mutant phenotypes reported since the early 1980s[44]. Interestingly, the most active ERV families in mice have lost their *env* gene and thereby their infectious capacity, but they have morphed into retrotransposons with a high level of germline activity[44,45]. Consistent with recent transposition, ERV insertions are often polymorphic among mice strains. For example, ~3,000 (60%) of ~5,000 intracisternal A particle (IAP) insertions in the B6 reference genome are absent at orthologous positions in at least one of three other common strains examined[46]. In a recent systematic study of structural variation in the mouse genome, ERV insertional polymorphisms accounted for ~150,000 (14%) of the ~700,000 structural variants greater than 100 bp in length that were identified among 17 strains[47]. A high-copy-number, transpositionally active ERV family was also recently identified in rats, where it has engendered extensive genomic variation among inbred strains and at least one heritable *de novo* insertion with phenotypic effects[48]. Thus, murine rodents offer valuable models for studying the short-term consequences of ERV mutagenesis on genome function and phenotype.

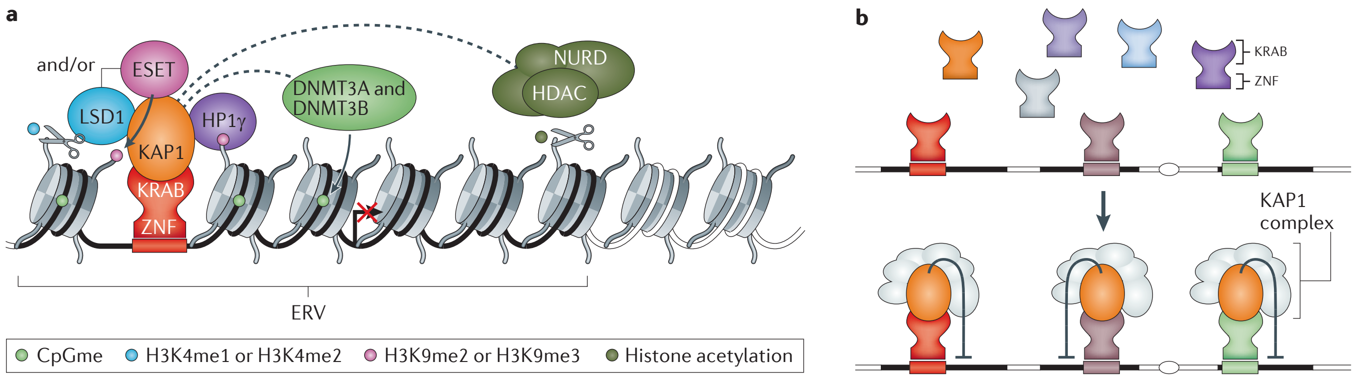
Post-insertional genomic rearrangements. As with other repeated sequences, non-allelic homologous recombination (NAHR) between members of the same EVE family that are located on the same or different chromosomes may lead to genomic rearrangements. Depending on the orientation and location of the elements relative to one another, NAHR may result in deletion, duplication, inversion or translocation events. The frequency and biological impact of NAHR events largely depend on the length, homology, density and distribution properties of the repeats throughout the genome. Generally speaking, families with large numbers of long, highly homogeneous copies are more likely to engage in rearrangements than rare, short and strongly diverged elements. All cases of EVE-mediated rearrangements so far described in humans have implicated ERVs, which is not surprising, given the predominance of these EVEs in our genome. A detailed study of the HERV-K (HML2) subfamily shows that 6 (17%) of the 35 full-length copies examined have undergone some form of NAHR following insertion[49], despite their fairly recent origin (<20 million years ago). Given that this subfamily accounts for a minuscule fraction (<<1%) of human ERVs, these data underscore the potentially profound contribution of ERVs to remodelling genome architecture over a large timescale. ERV-mediated NAHR events can also be pathogenic. For example, a non-reciprocal recombination event between two HERV-I copies on the Y chromosome caused a 792 kb deletion containing the azoospermia factor gene (*AZFA*; also known as *AZF1*), apparently causing male infertility[50]. Similarly, HERV-mediated deletion of the eye absent homologue 1 (*EYA1*) gene was recently linked to branchio-oto-renal syndrome[51]. But somatic rearrangements involving ERVs might be the most frequent and most medically relevant, especially for tumorigenesis, given the long chain of studies linking ERV to cancer (see below). For instance, a recurrent translocation event that creates an oncogenic HERV-K–ETS translocation variant 1 (*ETV1*) fusion gene has been implicated in prostate cancer[52]. These studies beg for more systematic analyses to assess the importance of ERV-mediated rearrangements in disease and evolution.

Provirus
The integrated form of a retrovirus.

Reinfection
Repeated infection of the germ cells of the individual carrying a provirus, with possible horizontal transmission to other individuals.

Dimorphic
Full-length insertion present in some individuals but absent in others.

Box 3 | **The KRAB-ZNF army: a genome defence system against ERVs?**



Several recent studies point to the existence of a previously unrecognized genome surveillance system aimed at repressing the transcription of endogenous retroviruses (ERVs) and possibly other vertebrate retroelements during early embryonic development. A first hint at the system came from a series of elegant studies on the transcriptional silencing of murine leukaemia virus in mouse embryonic stem cells. These studies[108] established that silencing occurs through the tethering of the KAP1 (also known as TRIM28) co-repressor complex to proviral DNA by ZFP809, which is a member of the Krüppel-associated box zinc finger (KRAB-ZNF) family of DNA-binding proteins. KAP1 is known to bind several KRAB-ZNF proteins through its amino-terminus, whereas the carboxy-terminus recruits a suite of effectors that promotes the local formation of silenced chromatin[39,109] (shown in part **a** of the figure). In early embryonic development, KAP1 is known to bind ESET (also known as SETDB1), which di- or trimethylates histone H3 at lysine 9 (H3K9me2 or H3K9me3, respectively) and to heterochromatin protein 1 (HP1) family members, which 'read' H3K9me2 and H3K9me3 marks. KAP1 is also known to interact with another epigenetic silencer, LSD1 (also known as KDM1A), which demethylates lysine 4 on H3 (H3K4me1 or H3K4me2). Finally, KAP1 can also recruit the NURD histone deacetylase (HDAC) complex, which removes acetyl groups on histones. Through each or a combination of these events, it has been shown that KAP1 is a central player for the transcriptional silencing of a wide range of ERV families in mouse embryonic stem cells[39,56,109–111]. At later stages and in differentiated cells, the epigenetic marks that were deposited by KAP1 in early development may also contribute to the recruitment of DNA methyltransferase 3A (DNMT3A) and DNMT3B, which methylates cytosines at CpG sites (CpGme), reinforcing ERV transcriptional silencing. Together, these data suggest a model in which KAP1 and its associated co-repressor complex can be tethered to ERV DNA in a sequence- or family-specific fashion by a battery of KRAB-ZNF adaptors[39] (shown in part **b** of the figure). This model is supported by the exceptional expansion, diversification and turnover of KRAB-ZNF genes in tetrapod genomes. These genes are encoded in the hundreds in all tetrapod species examined, but they tend to be poorly conserved in sequence and genomic location even between closely related species[112]. They have been gained and diversified at an extraordinary rate through lineage-specific tandem duplication followed by bouts of positive selection in their ZNF domains, all of which is predicted to change DNA-binding specificity[38,112] — an evolutionary pattern that is reminiscent of genes that are involved in antiviral immunity (BOX 5). Furthermore, there is a striking positive correlation across species between the number and age of KRAB-ZNF genes and those of their ERV content[38]. These data are consistent with an arms-race model in which the need to silence newly integrated retroviruses, and possibly other retroelements, has driven the duplication and divergence of KRAB-ZNF genes. This model might explain why KRAB-ZNF genes and ERVs tend to cluster together in the genome, and why several KRAB-ZNF genes are themselves under the transcriptional control of ERV-derived LTRs. These fortuitous associations might allow the establishment of a negative-feedback loop to modulate the KRAB-ZNF response. It could also explain the apparent co-option of a subset of ERV-controlled KRAB-ZNF genes for early embryonic development in mice[56,63].

Effects on host gene expression

Disruption of gene regulation. Viruses have evolved countless strategies for hijacking and manipulating the machinery of the host to control the expression of their own genes. In particular, the proviral form of retroviruses is adapted to recruit cellular factors to promote their transcription in the context of chromatin. Each of their long terminal repeats (LTRs) contains a basal promoter for RNA polymerase II and enhancers that are responsive to diverse conditions and signals, and the LTRs are bound by many transcription factors (see below). These features enable spatiotemporal control of proviral expression, as well as control of transcription termination and polyadenylation signals[53]. After they have been integrated into the host chromosome, any of these *cis*-elements has the potential to interfere with the expression of an adjacent host gene (or genes) through myriad mechanisms, including epigenetic effects[37,53,54].

These regulatory activities are likely to augment the deleterious effects that are associated with new EVE insertions and, for those that have reached fixation, to pose a long-lasting burden on the genome. This is reflected in the human genome by a strong statistical depletion of ERVs in the vicinity of promoter regions and near-intron splice sites, presumably because these are functionally sensitive regions in which insertions are rapidly eliminated by purifying selection[55]. The bulk of ERV insertions reaching fixation are transcriptionally silenced in most embryonic and adult tissues through repressive epigenetic marks[54] that are deposited and maintained by what is increasingly recognized to be a dedicated host surveillance system[39] (BOX 3). Failure to maintain or contain these silencing marks would result in the reactivation of dormant ERV insertions, which may trigger new waves of infection or transposition or may perturb local gene expression[54,56]. This was recently demonstrated by a report showing that derepression of a *THE1B* ERV in Hodgkin's lymphoma transcriptionally activates the adjacent colony-stimulating factor 1 receptor (*CSF1R*) proto-oncogene[57].

Box 4 | **'Resurrection' of extinct viruses**

The characterization of endogenous viral elements (EVEs) offers a unique opportunity to carry out functional studies of extinct viruses and to shed light on various aspects of the co-evolution between viral pathogens and their hosts. Two groups independently reconstructed an infectious version of human endogeneous retrovirus (HERV)-K, the most recently active but apparently now defunct ERV in the human genome[113,114]. These studies led to a detailed characterization of the HERV-K replication cycle and of its level of infectivity in various cultured cells. The ability of the researchers to generate infectious HERV-K sequences that were built from segments of different endogenous HERV-K loci suggests that human cells could do the same through recombination events between HERV-K genomic loci or transcripts. These results are consistent with occasional observations of HERV-K particles, notably in teratocarcinoma cell lines[75], and they substantiate the hypotheses according to which HERV-K reinfection or retrotransposition could be involved in several pathologies, including cancer[77] and AIDS[79]. Reconstruction of full-length or partial genomes of extinct EVEs can also be used to learn about the evolution and activity spectrum of host factors that are known to protect the cell against modern viruses. For example, studies of the resurrected HERV-K (a betaretrovirus) and of partially reconstructed chimpanzee retrovirus 1 and 2 (CERV1 (also known as PTERV1) and CERV2, which are both gammaretroviruses) have shown that the replication of these ancient retroviruses was most likely to be resistant to restriction by tripartite motif-containing 5α (TRIM5α) — a host factor that currently protects primates against various retroviruses[114,115] (but see REF. 116) (BOX 5). By contrast, the replication of HERV-K, CERV1 and CERV2 was strongly affected by apolipoprotein B mRNA-editing, enzyme-catalytic, polypeptide-like 3G (APOBEC3G)[115], which may explain, in part, why this host factor has evolved under strong positive selection throughout the primate radiation[117] (BOX 5). In another study, the TRIM5α protein from the ring-tailed lemur was shown to restrict the now-extinct lemur prosimian immunodeficiency virus (PSIV), as well as several extant lentiviruses[118]. Functional studies of ancient viral proteins may also enhance our comprehension of the molecular mechanisms underlying tropism and governing viral host range. For instance, Soll *et al.*[119] reconstructed and expressed a functional CERV2 *env* gene and identified the receptor used by this ancient virus to enter human cells. Interestingly, comparative sequence analysis of the receptor unveiled a series of mutational events that have rendered hamster cells resistant to CERV2. In a recent tour de force in palaeovirology, Goldstone *et al.*[120] expressed and produced crystal structures for the capsid domain of two prehistoric lentiviruses (namely, PSIV and rabbit endogenous lentivirus type K (RELIK)). Despite low levels of sequence identity, the structures display remarkable similarity to each other and to those of modern lentiviral capsids and, likewise, they are bound by cyclophilin A (cypA), which is a host factor that is known to be essential for HIV-1 infectivity.

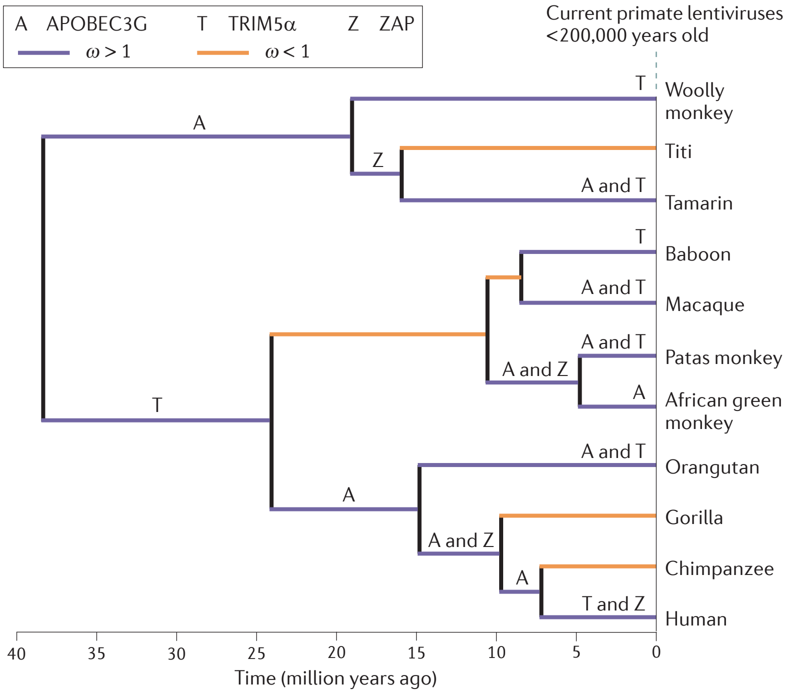
ERVs as an abundant source of promoters. Although the battery of *cis*-regulatory sequences carried by ERVs poses a hazard to host gene expression, it is also likely to predispose them to occasional co-option or 'exaptation' for regulatory function. Examples abound in the literature of ERV-derived sequences that have been incorporated into the 'normal' regulation of mammalian genes, most frequently as promoters, enhancers or polyadenylation signals[53]. Promoters acquired from LTRs typically function as alternative, tissue-specific promoters in addition to the broadly used ancestral promoters[58–60]. Hence, LTR-derived promoters are often co-opted during evolution to spatially increase or restrict gene expression patterns (for a recent example, see REF. 60). Because ERVs have been inserted and co-opted at various time points during evolution, these effects are confined to certain lineages and sometimes to certain individuals, thereby contributing to expression divergence among species and even within species[61,62].

High-throughput capture of transcription start sites (TSSs) and ultra-deep sequencing of poly(A)$^+$ RNA populations in mouse tissues and human cell lines have uncovered a staggering quantity of ERV-driven unique transcripts (~8,000 in mice and ~50,000 in humans), accounting for ~4% and ~7% of all TSSs mapped in these species, respectively[59,62]. Around 40% of these transcripts show highly restricted spatio-temporal expression[59] and most are lineage-specific, as only a small fraction of ERV insertions (~10%) are shared between human and mouse genomes. Given that a large number of ERV insertions are not yet fixed in mice, it would be interesting to examine the impact of ERV insertional polymorphism on intraspecific variation in gene expression and the possible phenotypic consequences.

Although some of the ERV-derived transcripts are produced from LTRs that function as traditional sense promoters in the 5′ region of protein-coding genes[63], most are actually driving the transcription of non-coding RNAs (ncRNAs)[59]. This category encompasses so-called *cis*-natural antisense transcripts that originate from ERVs that have been inserted in introns or in the 3′ region of protein-coding genes[64], as well as a profusion of other ncRNAs that emanate from intergenic or intronic elements[59]. Furthermore, it is common for ERV LTRs to function as bidirectional promoters, thereby serving not one but two transcriptional units, upstream and downstream of their integration site. Even more surprising, promoter activity is not restricted to LTRs but frequently arises from sequences mapping within the internal, coding region of ERVs[59]. Presumably, most of these internal promoters arose *de novo* from cryptic motifs that had been activated through post-insertional point mutations or positional effects. Together, these data reveal an extraordinarily complex, tightly regulated and evolutionarily labile ERV-derived 'subtranscriptome'. It may be an integral part of the 'flexible RNA scaffolds'[65] that are increasingly recognized as having a central role in genome regulation[66,67].

Dispersal of novel transcription factor binding sites. Another common route to EVE co-option is through the donation of enhancers and binding sites for transcription factors. Recently, several landmark studies have provided support to the four-decade-old hypothesis that mobile genetic elements had a pivotal role in the dispersion and turnover of transcription factor binding sites that coordinate gene regulatory networks in mammals[68,69]. Although any type of interspersed repetitive DNA may become involved in this process[69,70], the existing literature points to a disproportionate participation of ERVs in wiring large regulatory circuits. In one study, Wang *et al.*[71] found that >1,500 binding sites for p53 in the human genome lie within the LTRs of class I ERVs, and that most occurred in just two families (namely, the LTR10 and MER61 families) that invaded an ancestral anthropoid genome 40–63 million years ago. Remarkably, the ERV-derived p53 sites account for approximately one-third of those mapped in human colorectal cancer cells, where p53 is

Box 5 | **Antiviral defence and palaeovirology**



The long-lasting and conflicting interactions between viruses and their eukaryotic hosts have triggered the emergence of diverse antiviral defence strategies. In addition to transcriptional and post-transcriptional mechanisms of viral silencing found in various eukaryotes (BOX 3), a number of restriction factors have been identified in mammals that block the activity of retroviruses (and sometimes of other viruses) at various steps of their replication cycle[121]. The more taxonomically widespread and best characterized restriction factors are: the apolipoprotein B mRNA-editing, enzyme-catalytic, polypeptide-like (APOBEC) family of cytidine deaminases, which induce hypermutation of viral genomes by converting cytidines to uridines; tripartite motif-containing 5α (TRIM5α) and TRIMcyp proteins, which promote premature capsid uncoating following viral entry; a zinc finger antiviral protein (ZAP), which degrades viral RNA in the cytoplasm; and tetherin, a protein that prevents release of viral particles from the surface of producer cells. A series of elegant evolutionary analyses of these genes have shown that they have been subject to recurrent episodes of positive selection (shown by the purple branches in the figure), reflecting both the ongoing arms race (terminal purple branches in the figure) and the ancient arms race (internal purple branches in the figure) that take place at the host–virus interface during the evolution of primates and other mammals[117,122,123]. In the figure, ω values (taken from REFS 117,122,123) correspond to ratios of the number of nonsynonymous substitutions per nonsynonymous site to the number of synonymous substitutions per synonymous site. Typically, a value of ω that is lower than 1 indicates that the gene has evolved under negative or purifying selection, whereas a value higher than 1 indicates that the gene has evolved under positive or adaptive selection. These studies also revealed that in the case of ZAP and TRIM5α, positive selection was concentrated in a specific domain of the genes (the poly(ADP-ribose) polymerase (PARP) domain and the SPRY domain, respectively), which, after functional analysis, was confirmed to be the main site of interaction with retroviruses. Thus, together with the study of EVEs, evolutionary and functional analyses of intrinsic immunity genes are a key component of palaeovirology, providing invaluable clues on the nature of the forces at play in the host–virus battles and on the antiquity of these battles. In addition, the delineation of the precise molecular determinants evolving in conflict at the host–virus interface may open avenues for the development of new antiviral drugs.

strongly activated. Furthermore, Wang *et al.*[71] were able to confirm that binding of p53 to some of these ERVs correlated with p53-dependent activation of the nearest adjacent gene. Thus ERV-derived p53 sites seem to represent evolutionarily young functional *cis*-elements with physiological relevance. It is remarkable that such a large subset of the network of human p53 targets has apparently emerged from primate-specific ERVs because p53 is a deeply conserved, pleiotropic regulator of many biological pathways, and it is often coined the 'guardian of the genome' because of its tumour suppressor activity[72].

No less astounding are the results reported by two studies[73,74] showing that ERVs have contributed thousands of binding sites for OCT4 (also known as POU5F1) and Nanog (accounting for 7–15% of all binding sites mapped in human or mouse embryonic stem cells (ESCs)), which are two master regulators of the pluripotency gene network of ESCs. Because these ERVs are specific to either primate or rodent lineages, the data imply a massive emergence of novel binding sites for these two transcription factors in each lineage. Some ERV families have extensively contributed to the wiring of this network. For example, 255 (33.2%) of 767 LTR9B elements are bound by OCT4 in human ESCs — an 82-fold enrichment over the number expected based on the sheer density of LTR9B elements in the genome[73]. Alignments of the elements that are bound by OCT4 to the LTR9B consensus ancestral sequence suggest that the OCT4 binding sites pre-existed at the time of LTR9B integration in the genome. As for the aforementioned p53 binding sites, which also appear to have pre-existed in some ERV1 LTR families[71], it is tempting to speculate that these binding sites were used by the ancestral retroviruses to link their own transcription to certain conditions of the host cell (namely, stress for p53) or to a specific developmental stage (namely, ESCs in the case of OCT4 and Nanog) that is favourable to their propagation. Paradoxically, the molecular manipulations exerted by exogenous viruses on their host may have promoted the subsequent co-option of some of their endogenous descendants for host regulatory functions.

Impact of EVE gene products

Cellular expression of viral genes. Another mechanism by which EVEs can have an impact on host biology is through the cellular expression of viral coding components (but potentially also non-coding sequences). It is well-established that expression of some retroviral Env proteins, including those that are encoded by some ERVs, can modulate the host immune response[37,75,76]. Although thousands of *env* genes have been deposited in the human genome, most of them are corrupted by mutations and/or are not expressed in normal tissues or conditions owing to the defence mounted by the host to repress proviral expression[39] (BOX 3). However, ERVs can escape from silencing, at least transiently, through genetic or epigenetic changes triggered by certain cellular states, including cancer[77], or by environmental factors, such as diet[78] or infection by related or unrelated exogenous viruses[76,79]. For example, the induction of interferon-α in T cells by Epstein–Barr virus (a herpesvirus) infection leads to transcriptional activation of a normally silent HERV-K18 element at the CD48

Table 2 | **Examples of cellular genes of viral origin**

Gene	Virus progenitor (viral gene or domain)	Species distribution (age)	Function and activities	Refs
Syncytin 1 (also known as *ERVW1*)	HERV-W (*env*)	Catarrhine primates: humans, apes, Old World monkeys (25–40 million years)	Placenta-specific expression, fusogenic activities	142,143
Syncytin 2 (also known as *ERVFRD1*)	HERV-FRD (*env*)	Anthropoid primates: catarrhines and New World monkeys (40–65 million years)	Placenta-specific expression, fusogenic and immunosuppressive activities	144
Syncytin A (*Syna*)	HERV-F or HERV-H (*env*)	Murid rodents (20–30 million years)	Placenta formation (layer I of syncytiotrophoblast); placenta-specific expression, fusogenic activities *ex vivo*	94
Syncytin B (*Synb*)	HERV-F or HERV-H (*env*)	Murid rodents (20–30 million years)	Placenta formation (layer II of syncytiotrophoblast); placenta-specific expression, fusogenic and immunosuppressive activities	95
Syncytin-Ory1	Type D retroviruses (*env*)	Leporids: rabbits and hares (12–30 million years)	Placenta-specific expression, fusogenic activities	145
Syncytin-Car1	CarERV3 (class I)	Carnivores (65–80 million years)	Placenta-specific expression, fusogenic activities	146
ERVV1 and *ERVV2*	HERV-V (*env*)	Anthropoid primates (40–65 million years)	Placenta-specific expression, unknown function	147
Fv1	MuERV-L (*gag*) (class III)	*Mus* subgenera: mice (5–10 million years)	Confer resistance to murine leukaemia virus (MLV), binds MLV capsid	84,85,148
CGIN1 (also known as *NYNRIN*)	Retrovirus (*pol* (RNaseH, integrase)*)	Therians: placental and marsupial mammals (125–180 million years)	Unknown	149
EBLN1, *EBLN2*, *EBLN3* and *EBLN4*	Bornavirus (nucleoprotein)	Anthropoid primates (40–65 million years)	Unknown, but EBLN2 appears to interact with several cellular proteins	4,17,21
Iris	*Kanga* errantivirus (F-type *env*)	*Drosophila* melanogaster and obscura subgroups (25–35 million years)	Third instar larva- and adult-specific expression, localized to mitochondria	87,150

*Refers to the section of the gene that encodes RNaseH and integrase. CarERV, carnivore endogenous retrovirus; *EBLN1*, endogenous Bornavirus-like nucleoprotein 1; *ERVV1*, endogenous retrovirus group V, number 1; *Fv1*, Friend virus susceptibility 1; HERV, human endogenous retrovirus; MuERV, murine endogenous retrovirus.

locus[80]. Activation of this element results in overexpression of a truncated Env with superantigen activity, which induces an inflammatory cascade that is reminiscent of the onset of several autoimmune diseases. Intriguingly, overexpression of several HERV loci has been linked, with various degrees of consistency, to several autoimmune and/or neurological diseases, such as multiple sclerosis, type I diabetes, rheumatoid arthritis and schizophrenia[36,37,75,76]. Some ERV-encoded Env proteins also possess immunosuppressive activities that can promote tumorigenesis, as indicated by several studies with mouse cancer models[77]. Interestingly, the same immunosuppressive properties may be beneficial at a specific developmental stage or in certain tissues: for example, in the placenta syncytiotrophoblast, as described below.

Domestication of viral genes by the host. Despite the potentially disastrous consequences of expressing cellular genes of viral origins, there are circumstances in which viral gene products have been usefully recruited by the host. One common functional theme of domestication is 'fighting fire with fire': an EVE produces a protein that offers immunity against an exogenous virus. A classic example is endogenous Jaagsiekte sheep retrovirus (enJSRV), which protects the sheep genital tract from exogenous JSRV infection at two levels[81]. First, enJSRVs encode an ENV that binds a subset of the cellular receptors used by JSRV, decreasing their availability to exogenous particles and leading to a substantial reduction in viral entry[82]. Second, enJSRVs express a misfolded form of Gag protein, which co-assembles with JSRV Gag to form chimeric viral particles that are targeted for degradation by the proteasome[83]. Another well-known example is the Friend virus susceptibility 1 (*Fv1*) gene of mice, which derives from the *gag* gene of an extinct spumavirus-like ERV that infiltrated an ancestral mouse genome about 7 million years ago[84] and now restricts murine leukaemia virus (MLV), a very distantly related retrovirus. Although the molecular mechanism by which FV1 defeats MLV remains poorly understood, recent results indicate that it requires direct interaction between FV1 and the MLV capsid[85]. Following co-option, enJSRVs and *Fv1* have undergone rapid evolution by positive (or diversifying) selection[84,86]. This is a tell-tale signature of many antiviral host factors (BOX 5) that is thought to reflect a molecular arms race arising from conflicting interactions with exogenous retroviruses[5]. In the case of sheep enJSRV and JSRV, the arms race is illustrated by the recent emergence (<200 years ago) of JSRV variants that are capable of escaping enJSRV-mediated restriction[81]. A function in viral defence has also been hypothesized for several other EVE-derived genes that appear to have evolved, at least transiently, under selective constraint[16,21,27,87–91].

Superantigen
A class of antigens that cause nonspecific activation and uncontrolled proliferation of T cells, often resulting in a chronic inflammatory response.

Gag
A retroviral protein that is one of the structural proteins of the viral capsid.

The second functional theme emerging from analyses of mammalian genes of retroviral or retroelement origin is the physiology and development of the placenta[92]. This trend may, in part, be attributed to the remarkable developmental and structural plasticity of the placenta and its atypical pattern of global hypomethylation[92], which allows many ERVs and retroelements to remain transcriptionally active in this tissue[53,58,93]. These properties are likely to set the stage for the co-option of both coding and regulatory sequences of retro-origin for placental function. The best documented examples are the so-called syncytins, which are derived from the *env* gene of multiple ERVs (TABLE 2). Genetic studies in mice have established that the proteins encoded by syncytin A (*Syna*) and *Synb*, which arose independently in the rodent lineage from different ERV copies, are both required for the formation of the bilayered syncytiotrophoblast of the murine placenta[94,95]. These proteins appear to promote trophoblast cell fusion through a mechanism reminiscent of ENV-mediated retroviral entry. Remarkably, two other *env*-derived genes in humans, one gene in rabbits and one gene in carnivores (TABLE 2) — which all independently emerged from primate-, lagomorph- and carnivore-specific ERVs, respectively — display restricted expression in the developing placenta as well as fusogenic activities. Thus, functional properties inherited from a viral lifestyle (such as virus–cell fusion) combined with a transcriptionally permissive environment (such as the placenta) seem to have predisposed *env* genes to convergent domestication on at least six occasions to invent a novel tissue in several mammals, the syncytiotrophoblast, which is now key to host reproduction[92]. Interestingly, genetic experiments have established that the antiretroviral enJSRV elements have also become essential for the development of the sheep placenta[96]. Based on the enJSRV trajectory, it is tempting to envision antiviral defence as the initial selective pressure for some EVE genes to be retained and to diversify their function, thereby providing a stepping stone towards developmental and/or reproductive exaptation.

Concluding remarks and perspectives

The bounty of EVEs that has recently been unearthed from eukaryotic genomes is providing a treasure trove of information on the vast period separating the early origin of primitive viruses and the evolutionary dynamics of contemporary viruses. Notably, these genomic fossils are revealing that many modern viral families have much deeper evolutionary roots than was previously anticipated. In turn, the dating of these ancient viral sequences yields new calibration points enabling estimates of viral substitution rates across a macro-evolutionary timescale. These long-term rates turn out to be dramatically slower than those inferred from the analysis of modern viruses. Although this finding might seem surprising, it is inescapable, as one would predict that viral sequences that have diverged from their modern relatives for dozens of millions of years would be unrecognizable over such extended periods if they had evolved under the substitution rates inferred for circulating viruses. Whether the differences between short-term and long-term rates of viral evolution have biological underpinnings or whether they can merely be accounted for by methodological issues remains an open question. Improving our understanding of viral substitution rates is important for making accurate inferences on virus origin and evolution, but it is also crucial for the better monitoring and prediction of the epidemiological trajectory of clinically relevant viruses.

Although the impact that EVEs have on eukaryotic genome evolution has long been appreciated, until recently, most studies in this area were limited to one or a few candidate loci. The development of high-throughput technologies is beginning to provide a genome-wide perspective on the contribution of EVEs to cellular function, and the magnitude of this contribution is exceeding even the boldest predictions. Currently, most of our knowledge of the impact of EVEs on the host genome is based on the study of ERVs, in part because of their numerical dominance, but we can anticipate that further analysis of non-retroviral EVEs will yield new insights. For example, it will be interesting to evaluate whether EVEs can generate antiviral immunity through other mechanisms than those known for ERVs and, in particular, through an RNA-mediated response[97]. Furthermore, reconstruction and expression of ancient viruses, as well as the study of their interactions with host factors, could be extended to non-retroviral families using EVEs as template sequences. Thus far, such studies have dealt with the interaction of ancient 'resurrected' retroviral sequences with contemporary host factors (BOX 4). However, to delineate the history and processes of host–virus co-evolution more finely, it would be necessary to study the interaction of ancestral sequences that have been reconstructed for both virus and host factors within a robust phylogenetic framework[98].

Undoubtedly, the discovery of new EVEs will continue to widen our knowledge of viral evolution and of the impact that viruses have on their host beyond their immediate pathogenic effects. In addition, the study of EVEs offers an alternative source of information on viruses that is often directly related to those plaguing humans, crops or domestic animals, which has great potential to reveal new targets and avenues for the development of innovative antiviral strategies.

1. Breitbart, M. & Rohwer, F. Here a virus, there a virus, everywhere the same virus? *Trends Microbiol.* **13**, 278–284 (2005).
2. Koonin, E. V., Senkevich, T. G. & Dolja, V. V. The ancient Virus World and evolution of cells. *Biol. Direct* **1**, 29 (2006).
3. Forterre, P. & Prangishvili, D. The origin of viruses. *Res. Microbiol.* **160**, 466–472 (2009).
4. Katzourakis, A. & Gifford, R. J. Endogenous viral elements in animal genomes. *PLoS Genet.* **6**, e1001191 (2010).
This paper presents a systematic *in silico* mining of EVEs in animal genomes (that were available at the time), revealing that all major types of eukaryotic viruses can be endogenized.
5. Patel, M. R., Emerman, M. & Malik, H. S. Paleovirology — ghosts and gifts of viruses past. *Curr. Opin. Virol.* **1**, 304–309 (2011).
6. Kidwell, M. G. & Lisch, D. R. Perspective: transposable elements, parasitic DNA, and genome evolution. *Evolution* **55**, 1–24 (2001).

7. Levin, H. L. & Moran, J. V. Dynamic interactions between transposable elements and their hosts. *Nature Rev. Genet.* **12**, 615–627 (2011).
8. Iskra-Caruana, M. L., Baurens, F. C., Gayral, P. & Chabannes, M. A four-partner plant–virus interaction: enemies can also come from within. *Mol. Plant Microbe Interact.* **23**, 1394–1402 (2010).
9. Koonin, E. V., Mushegian, A. R., Ryabov, E. V. & Dolja, V. V. Diverse groups of plant RNA and DNA viruses share related movement proteins that may possess chaperone-like activity. *J. Gen. Virol.* **72**, 2895–2903 (1991).
10. Malik, H. S., Henikoff, S. & Eickbush, T. H. Poised for contagion: evolutionary origins of the infectious abilities of invertebrate retroviruses. *Genome Res.* **10**, 1307–1318 (2000).
Along with reference 45, this study blurs the boundary between retrotransposons and retroviruses and suggests an evolutionary continuum between the two.
11. Staginnus, C. & Richert-Pöggeler, K. R. Endogenous pararetroviruses: two-faced travelers in the plant genome. *Trends Plant Sci.* **11**, 485–491 (2006).
12. Lockhart, B. E., Menke, J., Dahal, G. & Olszewski, N. E. Characterization and genomic analysis of tobacco vein clearing virus, a plant pararetrovirus that is transmitted vertically and related to sequences integrated in the host genome. *J. Gen. Virol.* **81**, 1579–1585 (2000).
13. Gayral, P. *et al.* A single *Banana streak virus* integration event in the banana genome as the origin of infectious endogenous pararetrovirus. *J. Virol.* **82**, 6697–6710 (2008).
14. Holmes, E. C. The evolution of endogenous viral elements. *Cell Host Microbe* **10**, 368–377 (2011).
15. Geuking, M. B. *et al.* Recombination of retrotransposon and exogenous RNA virus results in nonretroviral cDNA integration. *Science* **323**, 393–396 (2009).
16. Taylor, D. J. & Bruenn, J. The evolution of novel fungal genes from non-retroviral RNA viruses. *BMC Biol.* **7**, 88 (2009).
17. Horie, M. *et al.* Endogenous non-retroviral RNA virus elements in mammalian genomes. *Nature* **463**, 84–87 (2010).
This was one of the first reports of non-retroviral EVEs in mammalian genomes and an experimental demonstration that Borna disease virus DNA can spontaneously integrate in the genome of human infected cells.
18. Bill, C. A. & Summers, J. Genomic DNA double-strand breaks are targets for hepadnaviral DNA integration. *Proc. Natl Acad. Sci. USA* **101**, 11135–11140 (2004).
19. Gilbert, C. & Feschotte, C. Genomic fossils calibrate the long-term evolution of hepadnaviruses. *PLoS Biol.* **8**, e1000495 (2010).
This paper provides a clear illustration of the discrepancy between short-term and long-term evolutionary rates of a virus family.
20. Belyi, V. A., Levine, A. J. & Skalka, A. M. Sequences from ancestral single-stranded DNA viruses in vertebrate genomes: the parvoviridae and circoviridae are more than 40 to 50 million years old. *J. Virol.* **84**, 12458–12462 (2010).
21. Belyi, V. A., Levine, A. J. & Skalka, A. M. Unexpected inheritance: multiple integrations of ancient bornavirus and ebolavirus/marburgvirus sequences in vertebrate genomes. *PLoS Pathog.* **6**, e1001030 (2010).
22. Krupovic, M. & Bamford, D. H. Virus evolution: how far does the double β-barrel viral lineage extend? *Nature Rev. Microbiol.* **6**, 941–948 (2008).
23. Pybus, O. G. & Rambaut, A. Evolutionary analysis of the dynamics of viral infectious disease. *Nature Rev. Genet.* **10**, 540–550 (2009).
24. Holmes, E. C. Evolutionary history and phylogeography of human viruses. *Annu. Rev. Microbiol.* **62**, 307–328 (2008).
The above two references provide an excellent review of the concepts and methods used to delineate the epidemiological dynamics of clinically relevant viruses.
25. Firth, C., Charleston, M. A., Duffy, S., Shapiro, B. & Holmes, E. C. Insights into the evolutionary history of an emerging livestock pathogen: porcine circovirus 2. *J. Virol.* **83**, 12813–12821 (2009).
26. Orito, E. *et al.* Host-independent evolution and a genetic classification of the hepadnavirus family based on nucleotide sequences. *Proc. Natl Acad. Sci. USA.* **86**, 7059–7062 (1989).
27. Katzourakis, A., Gifford, R. J., Tristem, M., Gilbert, M. T. & Pybus, O. G. Macroevolution of complex retroviruses. *Science* **325**, 1512 (2009).
28. Keckesova, Z., Ylinen, L. M., Towers, G. J., Gifford, R. J. & Katzourakis, A. Identification of a RELIK orthologue in the European hare (*Lepus europaeus*) reveals a minimum age of 12 million years for the lagomorph lentiviruses. *Virology* **384**, 7–11 (2009).
29. Cui, J. & Holmes, E. C. Endogenous lentiviruses in the ferret genome. *J. Virol.* **86**, 3383–3385 (2012).
30. Duffy, S., Shackelton, L. A. & Holmes, E. C. Rates of evolutionary change in viruses: patterns and determinants. *Nature Rev. Genet.* **9**, 267–276 (2008).
31. Lefeuvre, P. *et al.* Evolutionary time-scale of the begomoviruses: evidence from integrated sequences in the *Nicotiana* genome. *PLoS ONE* **6**, e19193 (2011).
32. Ho, S. Y. *et al.* Time-dependent rates of molecular evolution. *Mol. Ecol.* **20**, 3087–3101 (2011).
33. Gibbs, A. J., Fargette, D., García-Arenal, F. & Gibbs, M. J. Time—the emerging dimension of plant virus studies. *J. Gen. Virol.* **91**, 13–22 (2010).
34. Wertheim, J. O. & Kosakovsky Pond, S. L. Purifying selection can obscure the ancient age of viral lineages. *Mol. Biol. Evol.* **28**, 3355–3365 (2011).
35. Drake, J. W., Charlesworth, B., Charlesworth, D. & Crow J. F. Rates of spontaneous mutation. *Genetics* **148**, 1667–1686 (1998).
36. Gifford, R. & Tristem, M. The evolution, distribution and diversity of endogenous retroviruses. *Virus Genes.* **26**, 291–315 (2003).
37. Jern, P. & Coffin, J. M. Effects of retroviruses on host genome function. *Annu. Rev. Genet.* **42**, 709–732 (2008).
38. Thomas, J. H. & Schneider, S. Coevolution of retroelements and tandem zinc finger genes. *Genome Res.* **21**, 1800–1812 (2011).
This paper reports a striking correlation in the number and evolutionary emergence of KRAB-ZNF genes and ERVs within a wide range of vertebrate genomes.
39. Rowe, H. M. & Trono, D. Dynamic control of endogenous retroviruses during development. *Virology* **411**, 273–287 (2011).
40. Turner, G. *et al.* Insertional polymorphisms of full-length endogenous retroviruses in humans. *Curr. Biol.* **11**, 1531–1535 (2001).
41. Kidd, J. M. *et al.* A human genome structural variation sequencing resource reveals insights into mutational mechanisms. *Cell* **143**, 837–847 (2010).
42. Jha, A. R. *et al.* Human endogenous retrovirus K106 (HERV-K106) was infectious after the emergence of anatomically modern humans. *PLoS ONE* **6**, e20234 (2011).
43. Beck, C. R., Garcia-Perez, J. L., Badge, R. M. & Moran, J. V. LINE-1 elements in structural variation and disease. *Annu. Rev. Genomics Hum. Genet.* **12**, 187–215 (2011).
44. Maksakova, I. A. *et al.* Retroviral elements and their hosts: insertional mutagenesis in the mouse germ line. *PLoS Genet.* **2**, e2 (2006).
45. Ribet, D. *et al.* An infectious progenitor for the murine IAP retrotransposon: emergence of an intracellular genetic parasite from an ancient retrovirus. *Genome Res.* **18**, 597–609 (2008).
See summary for reference 10.
46. Zhang, Y., Maksakova, I. A., Gagnier, L., van de Lagemaat, L. N. & Mager, D. L. Genome-wide assessments reveal extremely high levels of polymorphism of two active families of mouse endogenous retroviral elements. *PLoS Genet.* **4**, e1000007 (2008).
47. Yalcin, B. *et al.* Sequence-based characterization of structural variation in the mouse genome. *Nature.* **477**, 326–329 (2011).
48. Wang, Y. *et al.* A novel active endogenous retrovirus family contributes to genome variability in rat inbred strains. *Genome Res.* **20**, 19–27 (2010).
49. Hughes, J. F. & Coffin, J. M. Evidence for genomic rearrangements mediated by human endogenous retroviruses during primate evolution. *Nature Genet.* **29**, 487–489 (2001).
50. Sun, C. *et al.* Deletion of azoospermia factor a (AZFa) region of human Y chromosome caused by recombination between HERV15 proviruses. *Hum. Mol. Genet.* **9**, 2291–2296 (2000).
51. Sanchez-Valle, A. *et al.* HERV-mediated genomic rearrangement of EYA1 in an individual with branchio-oto-renal syndrome. *Am. J. Med. Genet.* **152A**, 2854–2860 (2010).
52. Tomlins, S. A. *et al.* Distinct classes of chromosomal rearrangements create oncogenic ETS gene fusions in prostate cancer. *Nature* **448**, 595–599 (2007).
53. 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).
54. Maksakova, I. A., Mager, D. L. & Reiss, D. Keeping active endogenous retroviral-like elements in check: the epigenetic perspective. *Cell. Mol. Life Sci.* **65**, 3329–3347 (2008).
55. Zhang, Y., Romanish, M. T. & Mager, D. L. Distributions of transposable elements reveal hazardous zones in mammalian introns. *PLoS Comput. Biol.* **7**, e1002046 (2011).
56. Macfarlan, T. S. *et al.* Endogenous retroviruses and neighboring genes are coordinately repressed by LSD1/KDM1A. *Genes Dev.* **25**, 594–607 (2011).
This provides an explicit demonstration of how the epigenetic machinery repressing ERV expression may be co-opted for the coordinated control of neighbouring host gene expression in early mammalian development.
57. Lamprecht, B. *et al.* Derepression of an endogenous long terminal repeat activates the CSF1R proto-oncogene in human lymphoma. *Nature Med.* **16**, 571–579 (2010).
58. Seifarth, W. *et al.* Comprehensive analysis of human endogenous retrovirus transcriptional activity in human tissues with a retrovirus-specific microarray. *J. Virol.* **79**, 341–352 (2005).
59. Faulkner, G. J. The regulated retrotransposon transcriptome of mammalian cells. *Nature Genet.* **41**, 563–571 (2009).
60. Beyer, U., Moll-Rocek, J., Moll, U. M. & Dobbelstein, M. Endogenous retrovirus drives hitherto unknown proapoptotic p63 isoforms in the male germ line of humans and great apes. *Proc. Natl Acad. Sci. USA* **108**, 3624–3629 (2011).
61. van de Lagemaat, L. N., Landry, J. R., Mager, D. L. & Medstrand, P. Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. *Trends Genet.* **19**, 530–536 (2003).
62. Conley, A. B., Piriyapongsa, J. & Jordan, I. K. Retroviral promoters in the human genome. *Bioinformatics* **24**, 1563–1567 (2008).
References 59, 61 and 62 provide compelling evidence for an extensive, tightly regulated and lineage-specific ERV-derived transcriptome in mammals.
63. Peaston, A. E. *et al.* Retrotransposons regulate host genes in mouse oocytes and preimplantation embryos. *Dev. Cell* **7**, 597–606 (2004).
64. Conley, A. B., Miller, W. J. & Jordan, I. K. Human *cis* natural antisense transcripts initiated by transposable elements. *Trends Genet.* **24**, 53–56 (2008).
65. Zappulla, D. C. & Cech, T. R. RNA as a flexible scaffold for proteins: yeast telomerase and beyond. *Cold Spring Harb. Symp. Quant. Biol.* **71**, 217–224 (2006).
66. Mattick, J. S., Taft, R. J. & Faulkner, G. J. A global view of genomic information—moving beyond the gene and the master regulator. *Trends Genet.* **26**, 21–28 (2010).
67. Guttman, M. & Rinn, J. L. Modular regulatory principles of large noncoding RNAs. *Nature* **482**, 339–346 (2012).
68. Britten, R. J. & Davidson, E. H. Gene regulation for higher cells: a theory. *Science* **165**, 349–357 (1969).
69. Feschotte, C. Transposable elements and the evolution of regulatory networks. *Nature Rev. Genet.* **9**, 397–405 (2008).
70. Lynch, V. J., Leclerc, R. D., May, G. & Wagner, G. P. Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals. *Nature Genet.* **43**, 1154–1159 (2011).
71. 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).
72. Belyi, V. A. *et al.* The origins and evolution of the p53 family of genes. *Cold Spring Harb. Perspect. Biol.* **2**, a001198 (2010).
73. Kunarso, G. *et al.* Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nature Genet.* **42**, 631–634 (2010).

74. Xie, D. *et al.* Rewirable gene regulatory networks in the preimplantation embryonic development of three mammalian species. *Genome Res.* **20**, 804–815 (2010).
References 71, 73 and 74 show how transcription factor binding sites dispersed by ERV wire extensive gene regulatory networks in a lineage-specific fashion.
75. Bannert, N. & Kurth, R. Retroelements and the human genome: new perspectives on an old relation. *Proc. Natl Acad. Sci. USA* **101**, 14572–14579 (2004).
76. Perron, H. & Lang, A. The human endogenous retrovirus link between genes and environment in multiple sclerosis and in multifactorial diseases. *Clin. Rev. Allergy Immunol.* **39**, 51–61 (2010).
77. Kurth, R. & Bannert, N. Beneficial and detrimental effects of human endogenous retroviruses. *Int. J. Cancer* **126**, 306–314 (2010).
78. Waterland, R. A. & Jirtle, R. L. Transposable elements: targets for early nutritional effects on epigenetic gene regulation. *Mol. Cell. Biol.* **23**, 5293–5300 (2003).
79. Contreras-Galindo, R. A. *et al.* Characterization of human endogenous retroviral elements in the blood of HIV-1-infected individuals. *J. Virol.* **86**, 262–276 (2012).
80. Stauffer, Y. *et al.* Interferon-α-induced endogenous superantigen. A model linking environment and autoimmunity. *Immunity* **15**, 591–601 (2001).
81. Arnaud, F., Varela, M., Spencer, T. E. & Palmarini, M. Coevolution of endogenous betaretroviruses of sheep and their host. *Cell. Mol. Life Sci.* **65**, 3422–3432 (2008).
82. Spencer, T. E., Mura, M., Gray, C. A., Griebel, P. J. & Palmarini, M. Receptor usage and fetal expression of ovine endogenous betaretroviruses: implications for coevolution of endogenous and exogenous retroviruses. *J. Virol.* **77**, 749–753 (2003).
83. Mura, M. *et al.* Late viral interference induced by transdominant Gag of an endogenous retrovirus. *Proc. Natl Acad. Sci. USA* **101**, 11117–11122 (2004).
84. Yan, Y., Buckler-White, A., Wollenberg, K. & Kozak, C. A. Origin, antiviral function and evidence for positive selection of the gammaretrovirus restriction gene *Fv1* in the genus *Mus. Proc. Natl Acad. Sci. USA* **106**, 3259–3263 (2009).
85. Hilditch, L. *et al.* Ordered assembly of murine leukemia virus capsid protein on lipid nanotubes directs specific binding by the restriction factor, Fv1. *Proc. Natl Acad. Sci. USA* **108**, 5771–5776 (2011).
86. Arnaud, F. *et al.* A paradigm for virus-host coevolution: sequential counter-adaptations between endogenous and exogenous retroviruses. *PLoS Pathog.* **3**, e170 (2007).
This paper provides a comprehensive characterization of the various events that took place during the molecular arms race between domesticated endogenous betaretroviruses and their exogenous counterparts in sheep.
87. Malik, H. S. & Henikoff, S. Positive selection of *Iris*, a retroviral envelope-derived host gene in *Drosophila melanogaster*. *PLoS Genet.* **1**, e44 (2005).
88. Taylor, D. J., Dittmar, K., Ballinger, M. J. & Bruenn, J. A. Evolutionary maintenance of filovirus-like genes in bat genomes. *BMC Evol. Biol.* **11**, 336 (2011).
89. Liu, H. *et al.* Widespread endogenization of densoviruses and parvoviruses in animal and human genomes. *J. Virol.* **85**, 9863–9876 (2011).
90. Kobayashi, Y., Horie, M., Tomonaga, K. & Suzuki, Y. No evidence for natural selection on endogenous borna-like nucleoprotein elements after the divergence of Old World and New World monkeys. *PLoS ONE* **6**, e24403 (2011).
91. Fort, P. *et al.* Fossil rhabdoviral sequences integrated into arthropod genomes: ontogeny, evolution, and potential functionality. *Mol. Biol. Evol.* **29**, 381–390 (2012).
92. Rawn, S. M. & Cross, J. C. The evolution, regulation, and function of placenta-specific genes. *Annu. Rev. Cell Dev. Biol.* **24**, 159–181 (2008).
93. Blikstad, V., Benachenhou, F., Sperber, G. O. & Blomberg, J. Evolution of human endogenous retroviral sequences: a conceptual account. *Cell. Mol. Life Sci.* **65**, 3348–3365 (2008).
94. Dupressoir, A. *et al.* Syncytin-A knockout mice demonstrate the critical role in placentation of a fusogenic, endogenous retrovirus-derived, envelope gene. *Proc. Natl Acad. Sci. USA* **106**, 12127–12132 (2009).
95. 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).
The above two references provide genetic evidence for an essential role of two *env*-derived murine syncytins in placenta formation.
96. Dunlap, K. A. *et al.* Endogenous retroviruses regulate periimplantation placental growth and differentiation. *Proc. Natl Acad. Sci. USA* **103**, 14390–14395 (2006).
97. Flegel, T. W. Hypothesis for heritable, anti-viral immunity in crustaceans and insects. *Biol. Direct* **4**, 32 (2009).
98. Harms, M. J. & Thornton, J. W. Analyzing protein structure and function using ancestral gene reconstruction. *Curr. Opin. Struct. Biol.* **20**, 360–366 (2010).
99. 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).
100. 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).
101. Sharp, P. M., Bailes, E., Stevenson, M., Emerman, M. & Hahn, B. H. Gene acquisition in HIV and SIV. *Nature* **383**, 586–587 (1996).
102. Hu, J. *et al.* Characterization and comparison of recombinant simian immunodeficiency virus from drill (*Mandrillus leucophaeus*) and mandrill (*Mandrillus sphinx*) isolates. *J. Virol.* **77**, 4867–4880 (2003).
103. Tristem, M., Purvis, A. & Quicke, D. L. Complex evolutionary history of primate lentiviral *vpr* genes. *Virology* **240**, 232–237 (1998).
104. 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).
105. 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).
106. Kijima, T. E. & Innan, H. On the estimation of the insertion time of LTR retrotransposable elements. *Mol. Biol. Evol.* **27**, 896–904 (2010).
107. Martins, H. & Villesen, P. Improved integration time estimation of endogenous retroviruses with phylogenetic data. *PLoS ONE* **6**, e14745 (2011).
108. Wolf, D. & Goff, S. P. Embryonic stem cells use ZFP809 to silence retroviral DNAs. *Nature* **458**, 1201–1204 (2009).
109. Leung, D. C. & Lorincz, M. C. Silencing of endogenous retroviruses: when and why do histone marks predominate? *Trends Biochem. Sci.* 16 Dec 2011 (doi: 10.1016/j.tibs.2011.11.006).
110. Rowe, H. M. *et al.* KAP1 controls endogenous retroviruses in embryonic stem cells. *Nature* **463**, 237–240 (2010).
111. Matsui, T. *et al.* Proviral silencing in embryonic stem cells requires the histone methyltransferase ESET. *Nature* **464**, 927–931 (2010).
References 108–111 unveil fundamental components and principles of a recently discovered system of proviral silencing in mammal ESCs.
112. Nowick, K., Hamilton, A. T., Zhang, H. & Stubbs, L. Rapid sequence and expression divergence suggest selection for novel function in primate-specific KRAB-ZNF genes. *Mol. Biol. Evol.* **27**, 2606–2617 (2010).
113. Dewannieux, M. *et al.* Identification of an infectious progenitor for the multiple-copy HERV-K human endogenous retroelements. *Genome Res.* **16**, 1548–1556 (2006).
114. Lee, Y. N. & Bieniasz, P. D. Reconstitution of an infectious human endogenous retrovirus. *PLoS Pathog.* **3**, e10 (2007).
The above two references reconstitute an infectious progenitor for a human endogenous retrovirus.
115. Perez-Caballero, D., Soll, S. J. & Bieniasz, P. D. Evidence for restriction of ancient primate gammaretroviruses by APOBEC3 but not TRIM5α proteins. *PLoS Pathog.* **4**, e1000181 (2008).
116. Kaiser, S. M., Malik, H. S. & Emerman, M. Restriction of an extinct retrovirus by the human TRIM5α antiviral protein. *Science*. **316**, 1756–1758 (2007).
117. Sawyer, S. L., Emerman, M. & Malik, H. S. Ancient adaptive evolution of the primate antiviral DNA-editing enzyme APOBEC3G. *PLoS Biol.* **2**, e275 (2004).
118. Rahm, N. *et al.* Unique spectrum of activity of prosimian TRIM5α against exogenous and endogenous retroviruses. *J. Virol.* **85**, 4173–4183 (2011).
119. Soll, S. J., Neil, S. J. & Bieniasz, P. D. Identification of a receptor for an extinct virus. *Proc. Natl Acad. Sci. USA* **107**, 19496–19501 (2010).
120. Goldstone, D. C. *et al.* Structural and functional analysis of prehistoric lentiviruses uncovers an ancient molecular interface. *Cell Host Microbe* **8**, 248–259 (2010).
121. Wolf, D. & Goff, S. P. Host restriction factors blocking retroviral replication. *Annu. Rev. Genet.* **42**, 143–163 (2008).
122. Sawyer, S. L., Wu, L. I., Emerman, M. & Malik, H. S. Positive selection of primate TRIM5α identifies a critical species-specific retroviral restriction domain. *Proc. Natl Acad. Sci. USA* **102**, 2832–2837 (2005).
This study provides a vivid demonstration of the power of evolutionary sequence analysis to shed crucial insight into the interaction of host restriction factors with their viral targets.
123. Kerns, J. A., Emerman, M. & Malik, H. S. Positive selection and increased antiviral activity associated with the PARP-containing isoform of human zinc-finger antiviral protein. *PLoS Genet.* **4**, e21 (2008).
124. Lin, C. L. *et al.* Persistent Hz-1 virus infection in insect cells: evidence for insertion of viral DNA into host chromosomes and viral infection in a latent status. *J. Virol.* **73**, 128–139 (1999).
125. Arbuckle, J. H. *et al.* The latent human herpesvirus-6A genome specifically integrates in telomeres of human chromosomes *in vivo* and *in vitro*. *Proc. Natl Acad. Sci. USA* **107**, 5563–5568 (2010).
126. Morissette, G. & Flamand, L. Herpesviruses and chromosomal integration. *J. Virol.* **84**, 12100–12109 (2010).
127. Bézier, A., Herbinière, J., Lanzrein, B. & Drezen, J. M. Polydnavirus hidden face: the genes producing virus particles of parasitic wasps. *J. Invertebr. Pathol.* **101**, 194–203 (2009).
128. Delaroque, N., Maier, I., Knippers, R. & Mueller, D. G. Persistent virus integration into the genome of its algal host, *Ectocarpus siliculosus* (Phaeophyceae). *J. Gen. Virol.* **80**, 1367–1370 (1999).
129. Cock, J. M. *et al.* The *Ectocarpus* genome and the independent evolution of multicellularity in brown algae. *Nature* **465**, 617–621 (2010).
130. Liu, H. *et al.* Widespread horizontal gene transfer from circular single-stranded DNA viruses to eukaryotic genomes. *BMC Evol. Biol.* **11**, 276 (2011).
131. Bejarano, E. R., Khashoggi, A., Witty, M. & Lichtenstein, C. Integration of multiple repeats of geminiviral DNA into the nuclear genome of tobacco during evolution. *Proc. Natl Acad. Sci. USA* **93**, 759–764 (1996).
132. Ashby, M. K. *et al.* Analysis of multiple copies of geminiviral DNA in the genome of four closely related *Nicotiana* species suggest a unique integration event. *Plant. Mol. Biol.* **35**, 313–321 (1997).
133. Kapoor, A., Simmonds, P. & Lipkin, W. I. Discovery and characterization of mammalian endogenous parvoviruses. *J. Virol.* **84**, 12628–12635 (2010).
134. Tang, K. F. & Lightner, D. V. Infectious hypodermal and hematopoietic necrosis virus (IHHNV)-related sequences in the genome of the black tiger prawn *Penaeus monodon* from Africa and Australia. *Virus Res.* **118**, 185–191 (2006).
135. Liu, H. *et al.* Widespread horizontal gene transfer from double-stranded RNA viruses to eukaryotic nuclear genomes. *J. Virol.* **84**, 11876–11887 (2010).
136. Frank, A. C. & Wolfe, K. H. Evolutionary capture of viral and plasmid DNA by yeast nuclear chromosomes. *Eukaryot. Cell* **8**, 1521–1531 (2009).
137. Maori, E., Tanne, E. & Sela, I. Reciprocal sequence exchange between non-retro viruses and hosts leading to the appearance of new host phenotypes. *Virology* **362**, 342–349 (2007).
138. Crochu, S. *et al.* Sequences of flavivirus-related RNA viruses persist in DNA form integrated in the genome of *Aedes* spp. mosquitoes. *J. Gen. Virol.* **85**, 1971–1980 (2004).
139. Roiz, D., Vázquez, A., Seco, M. P., Tenorio, A. & Rizzoli, A. Detection of novel insect flavivirus sequences integrated in *Aedes albopictus* (Diptera: Culicidae) in Northern Italy. *Virol. J.* **6**, 93 (2009).

140. Tanne, E. & Sela, I. Occurrence of a DNA sequence of a non-retro RNA virus in a host plant genome and its expression: evidence for recombination between viral and host RNAs. *Virology* **332**, 614–622 (2005).
141. Taylor, D. J., Leach, R. W. & Bruenn, J. Filoviruses are ancient and integrated into mammalian genomes. *BMC Evol. Biol.* **10**, 193 (2010).
142. Blond, J. L. *et al.* An envelope glycoprotein of the human endogenous retrovirus HERV-W is expressed in the human placenta and fuses cells expressing the type D mammalian retrovirus receptor. *J. Virol.* **74**, 3321–3329 (2000).
143. Mi, S. *et al.* Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. *Nature* **403**, 785–789 (2000).
144. Blaise, S., de Parseval, N., Bénit, L. & Heidmann, T. Genome wide screening for fusogenic human endogenous retrovirus envelopes identifies syncytin 2, a gene conserved on primate evolution. *Proc. Natl Acad. Sci. USA* **100**, 13013–13018 (2003).
145. Heidmann, O., Vernochet, C., Dupressoir, A. & Heidmann, T. Identification of an endogenous retroviral envelope gene with fusogenic activity and placenta-specific expression in the rabbit: a new "syncytin" in a third order of mammals. *Retrovirology* **6**, 107 (2009).
146. Cornelis, G. *et al.* Ancestral capture of syncytin-Car1, a fusogenic endogenous retroviral envelope gene ivolved in placentation and conserved in Carnivora. *Proc. Natl Acad. Sci. USA* **109,** e432–e441 (2012).
147. Blaise, S., de Parseval, N. & Heidmann, T. Functional characterization of two newly identified human endogenous retrovirus coding envelope genes. *Retrovirology* **2**, 19 (2005).
148. Best, S., Le Tissier, P., Towers, G. & Stoye, J. P. Positional cloning of the mouse retrovirus restriction gene *Fv1*. *Nature* **382**, 826–829 (1996).
149. Marco, A. & Marín, I. CGIN1: a retroviral contribution to mammalian genomes. *Mol. Biol. Evol.* **26**, 2167–2170 (2009).
150. Lung, O. & Blissard, G. W. A cellular *Drosophila melanogaster* protein with similarity to baculovirus F envelope fusion proteins. *J. Virol.* **79**, 7979–7989 (2005).

Acknowledgements

We apologize to many colleagues who have produced primary research on the topic that could not be cited or discussed owing to space limitations. We thank the three anonymous reviewers for their constructive comments and useful suggestions. This work was supported by grant GM77582 from the US National Institutes of Health to C.F.

Competing interests statement

The authors declare no competing financial interests.

FURTHER INFORMATION

Cédric Feschotte's homepage:
http://www.uta.edu/faculty/cedric
CNRS 'Ecology, Evolution, Symbiosis' research unit:
http://ecoevol.labo.univ-poitiers.fr/?lang=en
University of Utah Department of Human Genetics:
http://www.genetics.utah.edu

ALL LINKS ARE ACTIVE IN THE ONLINE PDF