Distribution, diversity and evolution of endogenous retroviruses in perissodactyl genomes.

Running Head: **Endogenous retroviruses in perissodactyl genomes.**

Henan Zhu[1], Robert James Gifford[1*], Pablo Ramiro Murcia[1*]

[1] MRC-University of Glasgow Centre for Virus Research, Glasgow, UK, G61 1QH

* To whom correspondence should be addressed:
Robert James Gifford: robert.gifford@glasgow.ac.uk
Pablo Ramiro Murcia: pablo.murcia@glasgow.ac.uk

Abstract

The evolution of mammalian genomes has been shaped by interactions with endogenous retroviruses (ERVs). In this study, we investigated the distribution and diversity of ERVs in the mammalian order Perissodactyla, with a view to understanding their impact on the evolution of modern equids (family Equidae). We characterize the major ERV lineages in the horse genome in terms of their genomic distribution, ancestral genome organization and time of activity. Our results show that subsequent to their ancestral divergence from rhinos and tapirs, equids acquired four novel ERV lineages. We show that two of these proliferated extensively in the lineage leading to modern horses, and one contains loci that are actively transcribed in specific tissues. In addition, we show that the white rhinoceros has resisted germline colonisation by retroviruses for over 54 million years - longer than any other extant mammalian species. The map of equine ERVs that we provide here will be of great utility to future studies aiming to investigate the potential functional roles of equine ERVs, and their impact on equine evolution.

IMPORTANCE

ERVs in the host genome are highly informative about the long-term interactions of retroviruses and hosts. They are also interesting because they have influenced the evolution of mammalian genomes in various ways. In this study, we derive a calibrated timeline describing the process through which ERV diversity has been generated in the equine germline. We determined the distribution and diversity of perissodactyl ERV lineages and inferred their retrotranspositional activity during evolution, thereby gaining insight into the long-term co-evolutionary history of retroviruses and mammals. Our study provides a platform for future investigations to identify equine ERV loci involved in physiological processes and/or pathological conditions.

Introduction

The genomes of mammalian species contain thousands of sequences derived from retroviruses (1, 2). Retroviruses are characterized by a replication strategy in which the viral genome is stably integrated into the genome of the host cell (a form referred to as 'provirus') (3). Thus, when retroviral infection occurs in cells of the host germline (i.e. sperm, eggs or early embryo), integrated proviruses can be vertically inherited as host alleles. These 'endogenous retrovirus' (ERV) loci may subsequently increase their copy number within host species genome - either through reinfection of germline cells or retrotransposition within them – leading to the generation of multi-copy ERV lineages (4-6). A subset of ERV copies have been fixed in host genomes, and these sequences constitute a genomic 'fossil record' from which the long-term evolutionary history of retroviruses can be inferred (6).

ERV insertions that are only slightly deleterious or selectively neutral may be fixed through chance or genetic hitchhiking (6). However, some appear to have been fixed because they have been domesticated and/or neofunctionalized by host genomes to perform important physiological functions (7-9). Furthermore, even ERV sequences that do not encode proteins can play important physiological roles. For example, ERV loci can exert important impacts on the regulation of gene expression through their impact on epigenetic machinery (10, 11), or by expression of long non-coding RNAs (lncRNAs) (12).

Comparative studies indicate that the myriad of ERV lineages found in the genomes of modern mammals arose from multiple independent genome invasion events. As many of these events occurred after the divergence of mammalian orders, each mammalian order typically has its own distinct ERV composition and history. In fact, some ERVs are unique to individual genera or species. For example, ERVs derived from retroviruses in the genus *Gammaretrovirus* are present in chimpanzees (*Pan trogolodytes*) and Gorillas (*Gorilla gorilla*), but closely related ERVs are absent from the human genome (13). Each distinct mammalian lineage has its own characteristic history of ERV activity (e.g. infection, fixation and expansion). Consequently, characterisation of ERVs - and investigation of their potential physiological roles - has to be performed separately in distinct mammalian groups.

The domestic horse (*Equus caballus*) is an economically and scientifically important mammal that contributed significantly to the development of modern societies. Horses belong to the family Equidae, which comprises extant species of strict herbivores adapted for running and dietary specialization. The family Equidae, in turn, belongs to the

order Perissodactyla (odd-toed ungulates) (14, 15). Living perissodactyls represent a small remnant of a diverse group of mammals that apparently arose in North America ~54 million years ago (Mya), and subsequently became widespread on all continents apart from Australia and Antarctica (16-18). They are divided into two suborders: Hippomorpha containing the Equidae (horses, donkeys and zebras), and Ceratomorpha comprising the Tapiridae (tapirs) and Rhinocerotidae (rhinoceroses).

Several previous studies have examined ERV diversity in the horse genome (19-21). In this study, we use a range of bioinformatic approaches to characterise ERVs across a broad range of perissodactyl genomes, including several equid species and the white rhinoceros (*Ceratotherium simum*). We identify the major ERV lineages in the perissodactyl germline, recover representative genomes for each, and examine the dynamics of their expansion in the branch leading to modern horses. We also investigate the transcriptional profiles among ERV loci in equine-derived cells and tissues.

Results

Identification and phylogenetic classification of equine ERV lineages

We used a 'phylogenetic screening' (3) approach to characterise perissodactyl ERV lineages *in silico*. Because the reverse transcriptase (RT) is relatively refractory to mutation, similarity searches using the RT protein sequence will typically recover all ERV loci that contain an RT gene (22). Moreover, because the RT protein can be used to reconstruct evolutionary relationships across the entire *Retroviridae (23)*, phylogenetic approaches can be used to classify RT loci identified by screening (3).

We used this approach to identify and phylogenetically classify ERV RT sequences in 17 published perissodactyl genome sequences, representing ten distinct species, and seven distinct breeds of domestic horse (24, 25) (**Table S1**). We constructed phylogenies of the RT sequences identified in these screens, and identified all clades comprised exclusively of perissodactyl ERVs. Where these lineages were robustly separated from one another by RT sequences derived from ERVs or exogenous retroviruses found in non-mammalian hosts, we assumed they had arisen in independent germline invasion events (3). On this basis, we estimate that there are at least nine distinct ERV lineages present in the perissodactyl germline. All nine lineages are present in equids, whereas only five are found in the rhinoceros. We did not identify any ERV lineages that were unique to the

rhinoceros, or any that were specific to particular equid species or breeds. The RT phylogeny in **Figure 1** provides an overview of our findings.

Notably, we observed a complete absence in perissodactyl genomes of ERVs that group robustly within the *Gammaretrovirus* clade (as defined *sensu stricto* by exogenous gammaretroviruses). The perissodactyl germline also appears to lack any RT-encoding ERVs that group with HERV-I, despite such ERVs being present in most other mammal groups, and broadly distributed throughout vertebrates as a whole (26). While we did not identify any true members of the *Gammaretrovirus* genus in perissodactyls, we did identify several distinct lineages of clade I ERVs (*Gammaretrovirus*-related). These ERV lineages appear to be more closely related to human endogenous retroviruses (HERVs) than to any known exogenous retroviruses. Here, we refer to these three lineages as Rho (HERV.R(b)-related), Zeta (HERV.H/HERV.W-related) and Theta (HERV.L(b)-related) – see **Table 1** for further details of these ERV lineages and the HERV references they are based upon.

Strikingly, clade II ERVs were completely absent from the rhinoceros genome. In equids, by contrast, four clade II (*Betaretrovirus*-related) lineages are present, one of which (EqERV.b1) represents a *bona fide Betaretrovirus*, and has previously been described in detail (21). We identified two additional clade II lineages that grouped together with representatives of the HERV-K 'supergroup', which we refer to here as 'Kappa' (27). Accordingly, we named these two lineages EqERV.Kappa.1 and EqERV.Kappa.2. The fourth and final lineage of clade II ERVs we identified was found to be distinct from all previously characterized retroviruses and ERVs and was named *unclassified equine ERV 1* (EqERV.U1).

The ERV.L lineage (referred to here as Lambda) is an ancient group of clade III ERVs that is widespread throughout mammalian genomes, and entered the mammalian germline >105 My ago (28, 29). We identified numerous RT sequences belonging to this lineage in perissodactyls (**Table 1**). In addition, we identified a second lineage of clade III RT sequences that were related to the primate HERV.S lineage (referred to here as Sigma) (**Figure 1**) (3). A potential third lineage of clade III ERVs was also identified, grouping immediately basal to the Lambda lineage. However, all sequences within this low copy number group were highly degraded, and we could not determine with confidence whether they should be regarded as genuinely distinct from Lambda, and thus were not analysed further.

The ERV lineages identified here are represented by approximately similar numbers of RT sequences in almost all distinct equid species (**Table S2**). A few equid genomes had lower overall numbers of insertions, but the relative proportion of loci in each lineage was broadly equivalent to other equid species, suggesting that differences were related to low coverage. The Rho and Theta were found to have slightly higher copy numbers in the white rhino genome than in equids. Notably, a total of 908 Lambda RT sequences were identified in the rhino genome versus between 400 and 713 identified in equids.

Distribution and diversity of ERVs in perissodactyl genomes

Having identified distinct, monophyletic lineages of perissodactyl ERVs using the RT gene, we next sought to characterise the genome structure and evolutionary history of those lineages in greater depth. Retroviral proviruses typically encode three principal coding domains (*gag*, *pol* and *env*), flanked at either side by *long terminal repeat* (LTR) sequences, which are identical at the time of integration. However, many ERV loci are comprised of 'solo LTRs', generated when recombination between the 5' and 3' LTRs deletes internal coding sequences (30). To associate ERV RT sequences with full-length proviruses (and thereby map associations between RT lineages and LTRs), we performed a second round of screening using the ERV annotation pipeline (ERVAP) (**Method**). Using this approach, we estimated the total number of proviruses (internal regions bounded by paired LTRs) and solo LTRs associated with each lineage of perissodactyl ERVs, as delineated by RT phylogeny (see **Figure 1**). **Table 1** summarises our findings.

All clade I lineages (Rho, Zeta and Theta), and the single clade III lineage for which we could identify LTRs (Sigma) were associated with multiple, distinct LTR types. Furthermore, these lineages were clearly present in the ancestral perissodactyl germline (i.e. prior to the Hippomorpha-Ceratomorpha divergence), since we identified loci that were orthologous between rhinos and equids (**Figure 2a**). Notably we found far fewer proviruses than RT sequences in the Rho and Theta lineages, and no proviruses at all for the Lambda lineage. This likely reflects that the expansion of these ERV sequences in perissodactyls has been driven primarily by non-LTR mechanisms. These commonly entail reverse transcription and integration of ERV transcripts by non-LTR retrotransposons such as LINE-1 (4, 5), in which case ERV insertions are generated with truncated LTR sequences (e.g. see (31)). Such truncated sequences would in many cases not meet the criteria for classification as LTRs in our analysis pipeline (see **Methods**).

We obtained further evidence that non-LTR mechanisms have been involved in amplification of ancestral ERV RTs when attempting to infer representative/consensus internal sequences for each of the five ancestral lineages: we found that a high proportion of RT sequences belonging to the Lambda lineage are located within 1000bp of L1 domains encoding ORFs >400aa in length (**Table S5 and Table S6**), indicating they have been amplified as components of LINE1 (L1) transcripts. Multiple L1 lineages are believed to have been simultaneously active during the evolution of perissodactyls (32). Interestingly, we identified some L1 sequences encoding a chimeric protein containing ERV Lambda RT sequence fused to an L1 gene product (data not shown).

Since we could not confidently link RT sequences in the Lambda lineage with any LTRs, we could not count LTRs in this lineage. Furthermore, we were only able to generate a poor quality, truncated consensus sequence (data not shown). For the remaining eight remaining lineages, however, we recovered full-length consensus sequences encoding putative *env* genes, and established the links between RT lineages and LTR groups defined in RepBase (**Table 1, Figure 3 and Dataset 1**).

We found no evidence for the presence of any modern ERV lineages in the rhinoceros genome. Indeed, we could not identify any ERVs in the rhinoceros that were not derived from one of the five ancient lineages present in both rhinos and horses, suggesting that the rhinoceros has resisted ERV germline invasions for over 54 million years (16-18). To the best of our knowledge, this is the longest time that any mammalian lineage has existed without newly acquired ERVs becoming fixed in the germline. Only humans, in which have not acquired fixed insertions from any novel ERV lineages since diverging from other great ape speces.

We used data recovered via ERVAP to search equid genomes for ERV loci that were specific to particular breeds or species. We performed this analysis in the awareness that, in general, such loci cannot be comprehensively or rigorously mapped solely comparing whole genome sequences generated using short read sequencing. Firstly, assemblies constructed using a reference genome can include false positive “pseudologs” – ERV insertions that are present in the reference but actually missing in assembled genome (due to multiply mapped reads). Similarly, ERV insertions that are only present in individual horse breeds may not be detected, as reads from these loci may be incorrectly mapped to other loci in the reference genome. However, it is possible to identify a proportion of the loci that are absent from genomes assembled *de novo* (see **Table S1**), but present in the reference genome. We identified a total of ten such ERV loci (see **Table**

S4), all of which were derived from modern ERV lineages. They included an EqERV.b1 insertion that is absent from the genome of the Mongolian horse (an ancient breed of domestic horse), and an EqERV.U1 insertion that is absent from the genome of Przewalski's horse (**Figure 2b**).

Evolution of ERV lineages in the horse germline

We used a molecular clock-based approach, first described by Subramanian *et al* (33), to investigate the historical activity of ERV lineages. For each LTR group listed in **Table 1**, we created alignments including all LTR sequences identified by our screen (subject to sequence quality). We used these alignments to construct consensus LTR sequences for each LTR group. We then calculated pairwise distances between individual LTR loci and their corresponding LTR group consensus sequence. We converted pairwise distances into age estimates by assuming a neutral molecular clock and generated plots of estimated lineage activity over time (**Figure 4**).

We categorised perissodactyl ERV lineages that entered the germline prior to the rhino-equid divergence as 'ancient' and those that entered after as 'modern'. LTR dating (**Figure 4**) indicated that germline expansions of ancestral perissodactyl ERV lineages largely occurred in the Paleogene period (66-23 Mya) and continued for many millions of years after the divergence of the Hippomorpha and Ceratomorpha. In fact, some LTR groups associated with ancestral ERV lineages have undergone more recent expansions. In particular, the Rho and Zeta lineages include LTR groups (LTR1.3 and LTR1 respectively) that appear to have expanded much more recently (from ~25-5 Mya) (**Figure 4**).

Studies of mammalian ERVs indicate that intragenomic proliferation can occur through LTR-driven, intracellular retrotransposition (22). This is characterised by proviral loci with paired LTRs and intact *gag* and *pol* genes, but truncated or missing *env* genes. Several ancestral lineages (Rho, Theta, Zeta), and at least one modern lineage (Kappa.1) contained loci with such genome structures (**Table S3**). However, we also identified proviruses encoding envelope genes in all four ancestral ERV lineages (see **Table 1**), and furthermore each of these lineages contains at least one locus that encodes a near intact envelope protein (**Table S3**). Notably, expression of *env* RNA derived from the Zeta lineage has been reported previously in reproductive tissue (34). We found 252 ancestral ERV loci, and 14 modern ERV lineage loci that that overlapped with lncRNA loci (same

strand with > 1bp overlapping) annotated by Scott (35), including representatives of the Kappa2, Beta1 and U1 lineages.

The four modern ERV lineages identified in equid genomes grouped robustly within clade II. By narrowing the focus of our evolutionary investigations to this groups, we could reconstruct phylogenetic relationships using longer alignments (**Figure 5a**). Among the four lineages, one is a *bona fide* Betaretrovirus called EqERV.b1, and has been described previously (21). The EqERV.b1 lineage is relatively closely related to mouse mammary tumour virus (MMTV) and shares some of its characteristic features (e.g. LTRs >1000bp in length). We established that orthologous EqERV.b1 insertions are shared in the horse and donkey genomes, demonstrating that the lineage was present in the equid germline prior to the divergence of horses and donkeys ~6-10 Mya (36, 37). This establishes a minimum age for the EqERV.b1 lineage that is considerably more ancient than the 0.5 Myr suggested previously (21). Furthermore, by extension, the identification of this ortholog demonstrates a minimum age of 9 Myr for the entire lineage of MMTV-related retroviruses. The EqERV.b1 family contains a relatively large number of solo LTRs (Table 1), and when these sequences are used to estimate lineage activity, they indicate that EqERV.b1 expansion occurred in the late Neogene period, from ~12-5 Mya (**Figure 4**).

The remaining three 'Betaretrovirus-like' lineages group outside the clade defined by exogenous Betaretroviruses (Figure 3a), and together with members of the HERV.K supergroup, which comprises ten distinct groups of ERVs identified in primate genomes and labelled HML1-10. These groups, which were originally defined using DNA hybridisation, have since been shown to comprise at least two, phylogenetically distinct lineages: one containing the HML5 and HML6 lineages, and one containing all the others (Figure 3a). Here we refer to the clade that contains both these lineages, and the related equine ERV lineages, as 'Kappa'. Phylogenies based on pol show that both equine Kappa lineages (k1 and k2) are clearly distinct from related lineages in the human genome. Notably, we found that the EqERV.k1 genome contains a potential homolog of the HERV-K(HML2) *rec* gene with predicted splice sites in the expected locations (data not shown).

The EqERV.U1 lineage is not closely related to any previously characterized retrovirus or ERV, and in phylogenetic trees based on *pol* (**Figure 5a**), it groups as a robustly supported sister clade to ERVs and exogenous betaretroviruses found in birds and reptiles (38, 39). The EqERV.U1 lineage contains the largest number of proviruses (n=45) and solo LTRs (n=705) of any modern perissodactyl ERV lineage in the horse

genome, and intriguingly, also shows indications of relatively recent activity. We therefore investigated the evolution of the EqERV.U1 lineage in greater depth.

Genomic and phylogenetic characterization of the EqERV.U1 lineage

The alignment of full-length proviruses was used to infer a consensus genome structure for the EqERV.U1 lineage. This revealed that there were, in fact, two, distinct types of genomic organizations among EqERV.U1 insertions (**Figure 5c**). In the first of these (type I), the *pro* ORF encodes a dUTPase domain at its 5' end, as is found in Betaretroviruses (40). However, the majority of EqERV.U1 insertions had a more unusual genome structure (type II) in which the dUTPase was encoded by an ORF inserted into the 5' end of *gag*. This second type of genome structure has not previously been reported in any retrovirus.

We used a combination of approaches to calibrate the timescale of EqERV.U1 activity. Where paired LTRs were present, we estimated the age of loci by calculating the divergence between these sequences (which are derived from identical copies) and applying a neutral rate for the host genome. In addition, we examined published genome assemblies of other Perissodactyl species and subspecies for the presence of orthologous EqERV loci. We annotated information about loci ages and genome structure onto a phylogeny constructed from an alignment of EqERV.U1 proviruses (with dUTPase-encoding regions removed). We then annotated information about genome structure (type I versus type II) and insertion age onto this phylogeny (**Figure 5b**). Notably, the midpoint-rooted phylogeny showed the oldest insertions clustering toward the root of the tree. Furthermore, insertions with the more typical 'type I' genome organization were found almost exclusively toward the root, whereas all proviruses that exhibited a 'type II' genome structure clustered together in a single derived clade with robust bootstrap support. We identified two proviral loci that were unique to the horse, both of which exhibited a type II genome structure **(Figure 5c**). All other EqERV-U1 loci in the horse genome had orthologs in the donkey genome.

Together, these data indicate that the germline invasion event that originally generated the EqERV.U1 lineage occurred somewhere between 25-30 MYA (**Figure 4**). The initial expansion of this lineage involved ERVs with type I genome structures. Approximately 15 MYA (**Figure 4**), one EqERV copy underwent the genome rearrangements that generated the type II genome structure, and this element gave rise to

a lineage that has been expanding up until relatively recently (~1 MYA based on integration dates estimated by LTR comparisons).

Analysis of publicly available transcriptome data revealed that 21 EqERV.U1 loci showed evidence of expression, and for nine of these, the entire provirus appeared (based on read coverage) to be transcribed. However, we did not have sufficient resolution in this dataset to determine whether all expressed genes were from the same locus. The transcriptome datasets analysed here encompassed 17 derived from specific equine tissues, and one derived from an equine-derived cell line (E-derm). We found that brainstem, spinal cord, and oviduct only have Type I provirus expressions, whereas E-derms and skin only expressed type II proviruses. Trophectoderm has both type I and type II provirus transcripts. In E-derms, only one complete EqERV.U1 locus on chromosome 29 is transcribed.

Discussion

In this study, we examined ERV diversity in the order Perissodactyla, with the aim of understanding how interactions with retroviruses have shaped equid evolution. We used a "phylogenetic screening" approach to characterise ERV lineages, within which evolutionary relationships between RT-encoding proviral sequences were used as the primary basis for classifying loci. This established that there have been at least nine distinct genome invasion events in the perissodactyl lineage (**Figure 1**). We provide a minimum estimate because it is difficult to be certain that the nine lineages described here are comprised entirely of ERV insertions that arose from the same ancestral founder. This is particularly challenging when ERV lineages have undergone numerous separate expansions – for example many of the ancestral lineages identified here contain multiple LTR subgroups (see **Table 1**): these might reflect multiple distinct genome invasions by related viruses utilising distinct LTRs, or recombination events wherein pre-existing ERV lineages acquire novel LTRs, enabling further waves of intragenomic expansion.

Our efforts to recover representative proviral loci were instructive with regard to determining which equine ERV lineages were more ancient. Proviruses in the Lambda, Rho, Zeta, Theta and Sigma lineages all exhibited multiple frameshifts, in-frame stop codons, and indels. Moreover, for four of these lineages, we identified examples of loci that were orthologous between the Hippomorpha and Ceratomorpha (**Figure 2a**), establishing that they entered the mammalian germline >54 My ago. Given that no intact or near-intact proviruses were identified for any ancestral ERV lineage, it is likely that

amplification *in trans* (probably non-LTR via mechanisms) accounts for the differences in RT copy number observed for these lineages, and the relatively low number of proviruses versus RT sequences.

Overall, the ERV landscape of perissodactyl genomes broadly resembles that found in other large-bodied placental mammal groups (e.g. hominids, cetaceans and artiodactyls). These species generally have lower numbers of ERV sequences in their genomes when compared with many smaller-bodied mammal species (e.g. rodents, bats) (41). Furthermore, all the lineages we have defined as ancestral within perissodactyls (i.e. Lambda, Sigma, Rho, Theta and Zeta) have relatively closely related counterparts in humans, carnivores and artiodactyls. Importantly, when examined in the context of the entire retrovirus family, retroviral lineages that are in fact only distantly related can appear superficially similar, even though they in fact diverged a long time ago. For example, due to the time-dependent phenomenon observed for rates of evolutionary change in virus sequences (42), it is entirely possible that the retroviruses that gave rise to the avian and mammalian Rho lineages (see **Figure 1**) are as distantly related to one another as are the host species they infect.

Although the ERV composition of the horse genome shares broad similarities with other large-bodied mammals, it also exhibits some intriguing differences. Perhaps the most conspicuous of these is the total absence of ERVs grouping within the *Gammaretrovirus* genus (as defined by exogenous isolates) in any of the genomes we screened. In addition, the rhinoceros genome exhibits a total absence of clade II (*Betaretrovirus*-related) ERVs, despite these being present in the genome of most other mammalian species, including equids. The absence of these groups is surprising when considered in the light of previous studies, which have shown that they are extremely widespread in mammalian genomes (43-46). Given the diversity of species that appear to have harboured gamma- and betaretroviruses in the past, it seems likely that perissodactyl ancestors would have been exposed to these viruses. Potentially, the absence of these viruses from all or some perissodactyl lineages might reflect the existence of perissodactyl-specific antiviral factors that potently restrict these particular retrovirus groups, and experimental studies challenging equine cells with gammaretroviruses might allow these factors to be identified. However, it is also important to interpret the distribution of ERVs cautiously. Because it is highly statistically unlikely that any ERV locus will reach fixation, it is entirely possible that perissodactyl genomes have been invaded by ERV lineages that are not represented in the genomes of extant perissodactyl species. This

may also have occurred in the case of the rhino, which has acquired no fixed ERV loci from retroviruses that entered the germline after the Hippomorpha-Ceratomorpha divergence (~54 Mya).

The horse and human genomes are similar in that the only ERV lineages that appear likely to have been active recently are betaretrovirus-related. In humans and apes, the HERV.K(HML2) lineage contains some intact proviral loci that are capable of producing infectious particles, and are only present at a low frequency in the human population (47). In horses, two clade II (Betaretrovirus-related) lineages (EqERV.U1 and EqERV.b1) have generated high numbers of fixed loci in the past 20 million years. We identify insertions belonging to these lineages that are polymorphic among horse subspecies and breeds (**Table S3**) – indicating that the EqERV.b1 and EqERV.U1 lineages have remained active up until relatively recently. The annotations generated in our study (**Table S4**) can inform future efforts to map the distribution of polymorphic EqERV loci more precisely (e.g. by using PCR to amplify insertion sites from a range of breeds and subspecies).

Over recent years, it has become increasingly clear that ERVs have played an important role in shaping mammalian genome evolution. One way that ERVs can impact their hosts is by providing genes that are co-opted by host genomes to perform physiological functions in their host species (7-9). For example, syncytins are proteins derived from retroviral envelope (*env*) genes that have been domesticated by mammals to carry out an essential function in placental development (48, 49). We identified intact or nearly intact *env* genes in several ancient ERVs, and some of these might represent genes or pseudogenes that have (or had) syncytin-like properties. Indeed one of the *env* genes identified in our study (belonging to the Zeta lineage) is highly expressed in the placenta, and on this basis has previously been identified as a candidate syncytin-like gene (34). Alternatively, some (or all) of these *env* genes might encode proteins that restrict related retroviruses from infecting the cell via a receptor interference mechanism, as has been described for exogenous retroviruses (50), as well as endogenous *env* genes in other species (51-53). Intriguingly, one modern lineage (EqERV.U1) contains actively transcribed loci, consistent with a potential physiological role. In this lineage, expansion has been associated with the transposition of the dUTPase gene into the 5' end of the *gag* gene (**Figure 4)**, and we found evidence that some of these rearranged forms might express a gag-dUTPase fusion protein via ribosomal frameshifting (**Figure 5c**). The significance of the patterns of genomic rearrangement and transcription in the EqERV.U1

lineage remains unclear. However, to the extent these patterns have been shaped by selection pressures related to the dUTPase gene, they might provide an insight into the functions of this poorly understood retroviral enzyme (54).

Genomic changes mediated by ERV activity are also thought to have facilitated mammalian evolution by providing a platform for the emergence of new layers of epigenetic gene regulation during development (10). Notably, we found that many of the ERVs identified in our study overlapped lncRNAs (**Table S7**), indicating a potential role for equine ERVs in lncRNA-mediated gene regulation (55). We do not yet know to what extent ERV activity has mediated adaptive changes during equid evolution. Nonetheless, insofar as it has, our study offers some insight into which groups of ERVs are likely to have been involved. Equid evolution during the Miocene (15-20 Mya) was associated with physiological adaptations that arose as equine ancestors shifted from being small forest-dwelling animals feeding on leafy vegetation into larger-bodied herbivores adapted for life in open grassland (56). Our investigation indicates that during this period, loci belonging to specific ERV lineages and sublineages were being fixed in the equid germline at an elevated rate. As shown in **Figure 4**, these include several modern ERV lineages (EqERV.U1, EqERV.b1, EqERV.K1) as well as certain LTR subgroups of the ancestral Rho, Zeta and Theta lineages (in particular, the ERV1-2, ERV1, and MER34A1 subgroups of these lineages respectively). Whereas in the case of the modern ERV lineages, expansion appears to have been driven by a mixture of reinfection and intracellular retrotransposition, the expansion of ancestral ERV lineages is more clearly associated non-LTR mechanisms, particularly within the most ancient ERV groups found in the perissodactyl germline - Lambda, Rho and Theta.

Materials and Methods

Genome assembly

The reference genome of the domestic horse (equCab2, GCF_000002305.2) and the white rhinoceros (cerSim1, GCF_000283155.1) were downloaded from the NCBI Genome database (1). The donkey genome sequences (assembly 'willy') were downloaded from the Centre for GeoGenetics website (25). Whole genome sequencing short reads of the Somali wild ass (*Equus asinus somalicus*), the Onager (*Equus hemionus*), the Kiang (*Equus kiang*), the plains zebra (*Equus burchellii boehmi*), the Burchell's zebra (*Equus burchellii quagga*), the Grevy's zebra (*Equus grevyi*), the Hartmann's mountain zebra (*Equus zebra hartmannae*) were obtained from the NCBI

Sequence Read Archive (https://www.ncbi.nlm.nih.gov/sra, accession: PRJEB7446) (24, 25). Read trimming was performed by Trim Galore (57), and reads were mapped to the horse or donkey reference genomes using Bowtie2 with a very-sensitive–local option (equal to –D 20 –R 3 –N 0 –L 20 –i S,1,0.50) (58). Consensus genomes were generated using a combination of SAMtools and BCFtools (59).

Genome screening *in silico*

As a first step towards more definitively characterizing the evolutionary history of equine ERVs, we implemented a 'phylogenetic screening' strategy based on analysis of the reverse transcriptase (RT) peptide sequence. We collated a representative set of RT sequences derived from ERVs and exogenous retroviruses. These sequences were conceptually translated to peptide sequences. RT peptide sequences representing established retrovirus groups and ERV lineages were used as probes for *in silico* screening of perissodactly genomes.

The screening was performed using the database-integrated genome screening (DIGS) tool (60). Genomic sequences that disclosed statistically significant similarity to RT probes were extracted and classified by BLAST comparison to the RT reference library. A subset of these RT sequences was extracted and entered into a multiple sequence alignment (MSA) with RT sequences from our reference set. This MSA was then used as input for a maximum likelihood (ML) phylogenetic analysis. We used the phylogeny to identify well-supported clades that were comprised entirely of perissodactly ERVs. We then created RT reference sequences based on recovered equine RT sequences to represent these clades and repeated the DIGS process.

This enabled to identify a complete set of RT encoding ERVs in each of the species examined. For these loci, we then attempted to recover a more complete provirus, using the ERVAP pipeline (see **Figure 5**). In this pipeline, RT sequences were extracted along with 10 kilobases (Kb) of flanking sequence on each side. The LTRharvest (61) program is used to search for potential LTR sequences flanking RT matches. To be counted as LTRs, sequences were required to be >100bp in length and <20% divergent from one another. Where putative LTRs were identified, these were classified by BLAST comparison to a library of repetitive sequences obtained from RepBase (62). For proviral sequences with paired LTRs, the LTRdigest program (63) is used to annotate internal regions (i.e. by demarcating putative codings domains). For sequences that disclosed similarity to retroviral RTs, but were not flanked by identifiable LTRs, the HMMR program

was used to search for these domains. Annotations generated by LTRdigest and HMMR were based on retrovirus protein libraries obtained from PFAM (64) and a tRNA library obtained from GtRNAdb (65).

We used BLAST to search for ERV loci that were unique to individual species or breeds. We generated probe sequences that comprised 100bp of insertion site sequences, and 30bp of ERV sequence. Potential empty insertion sites were identified as genomic sequences that matched the probe in the flanking sequence region, but not in the ERV region.

Phylogenetic analysis

Maximum likelihood phylogenies were generated using RAxML (66) and model parameters selected using IQ-TREE model selection function (67). Support for phylogenies was assessed via 1000 non-parametric bootstrap replicates. A phylogeny based on RT was used to infer the relationships of equine ERVs to one another and to previously characterized retroviral RT sequences. This phylogeny was based on an alignment spanning 135 amino acid residues in RT and was reconstructed using the rtREV amino acid substitution matrix as selected by IQ-TREE (68). To investigate the evolutionary relationship of EqERV.U1 to other, closely related retroviruses, we constructed a second dataset by aligning complete Pol polyprotein sequences. Phylogenies were reconstructed using a codon-based alignment spanning RT, RNaseH and Integrase domains.

Dating

For LTR comparisons we excluded pairs that did not group together in LTR phylogenies since these pairs could reflect proviruses that have undergone non-homologous recombination in the internal region or artefacts generated during genome assembly. To date solo LTRs, we applied an approach described by Subramanian *et al.*, in which each LTR is dated by measuring divergence from a subgroup consensus and applying a neutral rate calibration.

Transcriptomics

Equine transcriptome data were obtained from the European Nucleotide Archive (ENA) (**Table S10**). Adapter sequences were removed using the Trim Galore! script. Trimmed reads were aligned to the *E.caballus* reference using TopHat and an annotation

file generated in-house from ENSEMBL 84 gene annotations combined with ERV annotations obtained via genome screening. Expression levels were inferred using Cuffquant, and values obtained from distinct experiments were normalised using Cuffnorm. Approximately 4551 million reads were obtained, which were then mapped to the equine reference genome (EquCab2). Mapping to Ensembl and ERV annotation resulted in 80.91% of reads (~3683 million) being assigned to host genes or ERV loci.

Journal of Virology JVI Journal of Virology

512 **Table 1. Profile of nine perissodactyl ERV lineages in the domestic horse genome**
513

Genus/ Group	Clade	Prototype	Prototype citation	Name[a]	PBS*	RepBase LTR subgroups*	Copy number			
							RT[b]	provirus[c]	*env*$^{(+)}$ provirus	Solo LTR
Rho	I	HERV.R(b)	(3)	Rho.1*	Arg(CCG)	1-2, 1-3, 15, 45, 72A, 72B, 8B, 8E, 8F	151	20	6	4057
Zeta	I	HERV.W	(69)	Zeta.1*	Leu(TAA)	1, 14, 1420	37	13	5	3862
Theta	I	HERV.L(b)	(70-72)	Theta.1*	*ND*	1-4, 27_FC	251	11	2	351
	I			Theta.2	*ND*	1-4B, 1-6, 13A, 19, 23B, 6, 6B, MER34A_CF, MER34A1	67	9	6	8675
Betaretrovirus	II	MMTV		Beta.1	Lys(TTT)	[4]	10	3*	3*	350
Kappa	II	HERV.K(HML2)	(33)	Kappa.1	Lys(CTT)	2-2	5	4	4	79
	II			Kappa.2	Lys(CTT)	This study	3	1	1	35
U1	II	N/A		U1	Trp(CCA)	2-1	45	32	32	705
U2	III	N/A		U2	*ND*	*ND*	54	NA	NA	NA
Lambda	III	HERV.L	(28)	Lambda*	*ND*	None identified	691	NA	0	NA
Sigma	III	HERV.S	(3)	Sigma	Ser(AGA) Ser(CGA)	3-1C, 74	67	1	0	296
Totals							1381	92	57	18410

514
515 [a] Refers to lineages demarcated in Figure 1. [b] Number of RT loci [c] Only loci that contained RT plus at least two retroviral coding
516 domains represented in PFAM (64), and were flanked by paired LTRs >100bp in length were counted as proviruses. [d] Number of
517 proviruses for which we detected the presence of (intact or fragmentary) *env* genes. [e] Number of solo LTRs. NA: not detected.

518 **FIGURE LEGENDS**
519
520 **Figure 1. Evolutionary relationships between perissodactyl endogenous, previously**
521 **characterised ERVs, and exogenous retroviruses.** The figure shows a maximum
522 likelihood phylogeny reconstructed from an alignment of retroviral reverse transcriptase
523 (RT) peptide sequences. Sequences extracted from the horse, donkey and rhino genomes
524 are indicated by gray circles, following the key shown top left. For previously characterised
525 ERVs and exogenous retroviruses, taxa labels show the abbreviated name (see **Table S8**)
526 for full details. Sequences derived from exogenous virus references are marked by open
527 circles aligned with taxa labels. Sequences identified in non-mammalian hosts are
528 indicated by red font. Retrovirus subfamilies and orthoretroviral clades (I, II and III) are
529 indicated on basal branches. Established retroviral genera and ERV lineages defined in
530 this study are indicated by coloured brackets. For each of these groups, the presence of
531 sequences in the rhinoceros, donkey and horse in each genus is indicated by grey bars,
532 following the the key (top left). Asterisks indicate nodes with bootstrap support >=70%.
533 The scale bar shows evolutionary distance in substitutions per site.
534
535 **Figure 2. Examples of orthologous and polymorphic ERV loci in perissodactyls.** The
536 DNA sequences of the extreme 5' and 3' ends of orthologous ERV internal are shown
537 enclosed by red boxes (with the majority of intervening ERV sequence being omitted).
538 Target site duplication (TSD) sequences flanking ERV insertions are shown enclosed by
539 blue boxes. 20-30 base pair (bp) regions of upstream and downstream flanking genomic
540 DNA sequence are shown for each locus. Panel **(a)** shows examples of insertions
541 belonging to the Theta, Rho and Sigma lineages (top to bottom) that occur at orthologous
542 loci in the horse, donkey and rhinoceros genomes. Panel **(b)** shows examples of the ERVs
543 in EqERV.b1 and EqERV.U1 lineages (top to bottom) that are polymorphic within horses.
544
545 **Figure 3. Schematic representation of proviruses.** The putative locations of *gag*, *pro*,
546 *pol*, and *env* coding domains within consensus proviral genomes is indicated by grey
547 boxes. Long terminal repeat (LTR) sequences are shown as white boxes. The estimated
548 positions of PBS and PPT sequences are indicated by black bars. A scale bar indicating
549 length in kilobases is shown above each genome diagram. Abbreviations: PBS: primer
550 binding site; MA: matrix; CA; capsid: NC; nucleocapsid; PR protease; DU: dUTPase; RT:
551 reverse transcriptase, IN: integrase; SU: Surface glycoprotein; TM: transmembrane
552 domain; PPT: polypurine tract.
553
554 **Figure 4. Inferred timeline of ERV lineage expansions show. Empirical cumulative**
555 **distribution function (ECDF) plots, representing** the accumulation of observed LTRs
556 over time. The ages of LTRs were inferred by estimating divergence from an LTR
557 consensus, and applying a molecular clock-based calibration. The x-axis shows time in
558 millions of years before present, and y-axis shows the proportion of LTR sequences
559 accumulated. Distinct LTR groups found to occur within the same ERV lineage are shown
560 within the same plot, using distinct colours as indicated by the plot-associated key. The
561 panel bottom right shows a time-scaled perissodactyl phylogeny obtained from the
562 TimeTree website (73). All x axes were adjusted to the same scale.
563
564 **Figure 5. Characteristics of modern equine ERVs.**
565 **Panel (a):** a maximum likelihood phylogeny representing the estimated evolutionary
566 relationships between Pol sequences derived from clade II ERVs in perissodactyl
567 genomes, and those of previously characterised ERVs and exogenous retroviruses. Taxa
568 labels for RT sequences detected in this study indicate the species in which they were

identified. Other taxa labels show the abbreviated name of the virus or ERV. Sequences identified in non-mammalian hosts are indicated in red. Brackets on the right indicate ERV lineages and retroviral genera. Asterisks indicate nodes with bootstrap support above 70%. The scale bar shows evolutionary distance in substitutions per site. Details of taxa are provided in **Table S8**.

Panel (b): consensus genome structures of EqERV.U1 proviruses. Viral coding domains are shown as dark grey bars. Long terminal repeats (LTRs) are shown as boxes. Crooked arrows indicate where we have inferred translational frameshifting. For type II proviruses, we show a putative frameshift site (indicated with a question mark) that would allow expression of a matrix-dUTPase fusion protein. Abbreviations: LTR (long terminal repeat); MA (matrix); CA (capsid); NC (nucleocapsid), DU (dUTPase); PR (protease); RT (reverse transcriptase); IN (integrase); SU (surface); TM (transmembrane).

Panel (c): a maximum likelihood phylogeny of EqERV.U1 loci based on the aligned nucleotide sequences of 25 full-length proviruses. The sidebar boxes to the right of taxa indicate the type of genome found in the element (see panel b) as indicated in the key below the tree. An asterisk on the sidebar shows the youngest provirus based on the paired LTR dating. Open circles indicate loci that show evidence of transcription based on analysis of transcriptomic datasets. Asterisks indicate nodes with bootstrap support above 70%. The scale bar shows evolutionary distance in substitutions per site.

Acknowledgements
Dr Robert J. Gifford and Dr Pablo R. Murcia were supported by grants from the UK Medical Research Council (No. MC_UU_12014/10 and MC_UU_12014/9, respectively).

595 **References**
596

597 1. Wade CM, Giulotto E, Sigurdsson S, Zoli M, Gnerre S, Imsland F, Lear TL,
598 Adelson DL, Bailey E, Bellone RR, Blocker H, Distl O, Edgar RC, Garber M, Leeb
599 T, Mauceli E, MacLeod JN, Penedo MC, Raison JM, Sharpe T, Vogel J, Andersson
600 L, Antczak DF, Biagi T, Binns MM, Chowdhary BP, Coleman SJ, Della Valle G,
601 Fryc S, Guerin G, Hasegawa T, Hill EW, Jurka J, Kiialainen A, Lindgren G, Liu J,
602 Magnani E, Mickelson JR, Murray J, Nergadze SG, Onofrio R, Pedroni S, Piras
603 MF, Raudsepp T, Rocchi M, Roed KH, Ryder OA, Searle S, Skow L, Swinburne
604 JE, Syvanen AC, Tozaki T, Valberg SJ, Vaudin M, White JR, Zody MC, Lander ES,
605 Lindblad-Toh K. 2009. Genome sequence, comparative analysis, and population
606 genetics of the domestic horse. Science 326:865-7.
607 2. Mouse Genome Sequencing C. 2002. Initial sequencing and comparative analysis
608 of the mouse genome. Nature 420:520.
609 3. Tristem M. 2000. Identification and characterization of novel human endogenous
610 retrovirus families by phylogenetic screening of the human genome mapping
611 project database. J Virol 74:3715-30.
612 4. Belshaw R, Katzourakis A, Paces J, Burt A, Tristem M. 2005. High copy number in
613 human endogenous retrovirus families is associated with copying mechanisms in
614 addition to reinfection. Mol Biol Evol 22:814-7.
615 5. de Parseval N, Heidmann T. 2005. Human endogenous retroviruses: from
616 infectious elements to human genes. Cytogenet Genome Res 110:318-32.
617 6. Gifford R, Tristem M. 2003. The evolution, distribution and diversity of endogenous
618 retroviruses. Virus Genes 26:291-315.
619 7. Varela M, Spencer TE, Palmarini M, Arnaud F. 2009. Friendly viruses: the special
620 relationship between endogenous retroviruses and their host. Ann N Y Acad Sci
621 1178:157-72.
622 8. Lavialle C, Cornelis G, Dupressoir A, Esnault C, Heidmann O, Vernochet C,
623 Heidmann T. 2013. Paleovirology of 'syncytins', retroviral env genes exapted for a
624 role in placentation. Philos Trans R Soc Lond B Biol Sci 368:20120507.
625 9. Redelsperger F, Raddi N, Bacquin A, Vernochet C, Mariot V, Gache V, Blanchard-
626 Gutton N, Charrin S, Tiret L, Dumonceaux J, Dupressoir A, Heidmann T. 2016.
627 Genetic Evidence That Captured Retroviral Envelope syncytins Contribute to
628 Myoblast Fusion and Muscle Sexual Dimorphism in Mice. PLoS Genet
629 12:e1006289.
630 10. Imbeault M, Helleboid PY, Trono D. 2017. KRAB zinc-finger proteins contribute to
631 the evolution of gene regulatory networks. Nature 543:550-554.
632 11. Rowe HM, Jakobsson J, Mesnard D, Rougemont J, Reynard S, Aktas T, Maillard
633 PV, Layard-Liesching H, Verp S, Marquis J, Spitz F, Constam DB, Trono D. 2010.
634 KAP1 controls endogenous retroviruses in embryonic stem cells. Nature 463:237-
635 40.
636 12. Kapusta A, Kronenberg Z, Lynch VJ, Zhuo X, Ramsay L, Bourque G, Yandell M,
637 Feschotte C. 2013. Transposable elements are major contributors to the origin,
638 diversification, and regulation of vertebrate long noncoding RNAs. PLoS Genet
639 9:e1003470.
640 13. Yohn CT, Jiang Z, McGrath SD, Hayden KE, Khaitovich P, Johnson ME, Eichler
641 MY, McPherson JD, Zhao S, Paabo S, Eichler EE. 2005. Lineage-specific
642 expansions of retroviral insertions within the genomes of African great apes but not
643 humans and orangutans. PLoS Biol 3:e110.
644 14. Radinsky LB. 1966. The Adaptive Radiation of the Phenacodontid Condylarths and
645 the Origin of the Perissodactyla. Evolution 20:408-417.

15. Wilson DE, Reeder DM. 2005. Mammal species of the world : a taxonomic and geographic reference, 3rd ed. Johns Hopkins University Press, Baltimore.
16. Bowen GJ, Clyde WC, Koch PL, Ting S, Alroy J, Tsubamoto T, Wang Y, Wang Y. 2002. Mammalian dispersal at the Paleocene/Eocene boundary. Science 295:2062-5.
17. Rose KD, Holbrook LT, Rana RS, Kumar K, Jones KE, Ahrens HE, Missiaen P, Sahni A, Smith T. 2014. Early Eocene fossils suggest that the mammalian order Perissodactyla originated in India. Nat Commun 5:5570.
18. Steiner CC, Ryder OA. 2011. Molecular phylogeny and evolution of the Perissodactyla. Zoological Journal of the Linnean Society 163:1289-1303.
19. Brown K, Moreton J, Malla S, Aboobaker AA, Emes RD, Tarlinton RE. 2012. Characterisation of retroviruses in the horse genome and their transcriptional activity via transcriptome sequencing. Virology 433:55-63.
20. Garcia-Etxebarria K, Jugo BM. 2012. Detection and characterization of endogenous retroviruses in the horse genome by in silico analysis. Virology 434:59-67.
21. van der Kuyl AC. 2011. Characterization of a full-length endogenous beta-retrovirus, EqERV-beta1, in the genome of the horse (Equus caballus). Viruses 3:620-8.
22. Magiorkinis G, Gifford RJ, Katzourakis A, De Ranter J, Belshaw R. 2012. Env-less endogenous retroviruses are genomic superspreaders. Proc Natl Acad Sci U S A 109:7385-90.
23. Xiong Y, Eickbush TH. 1990. Origin and evolution of retroelements based upon their reverse transcriptase sequences. EMBO J 9:3353-3362.
24. Jonsson H, Schubert M, Seguin-Orlando A, Ginolhac A, Petersen L, Fumagalli M, Albrechtsen A, Petersen B, Korneliussen TS, Vilstrup JT, Lear T, Myka JL, Lundquist J, Miller DC, Alfarhan AH, Alquraishi SA, Al-Rasheid KA, Stagegaard J, Strauss G, Bertelsen MF, Sicheritz-Ponten T, Antczak DF, Bailey E, Nielsen R, Willerslev E, Orlando L. 2014. Speciation with gene flow in equids despite extensive chromosomal plasticity. Proc Natl Acad Sci U S A 111:18655-60.
25. Orlando L, Ginolhac A, Zhang G, Froese D, Albrechtsen A, Stiller M, Schubert M, Cappellini E, Petersen B, Moltke I, Johnson PL, Fumagalli M, Vilstrup JT, Raghavan M, Korneliussen T, Malaspinas AS, Vogt J, Szklarczyk D, Kelstrup CD, Vinther J, Dolocan A, Stenderup J, Velazquez AM, Cahill J, Rasmussen M, Wang X, Min J, Zazula GD, Seguin-Orlando A, Mortensen C, Magnussen K, Thompson JF, Weinstock J, Gregersen K, Roed KH, Eisenmann V, Rubin CJ, Miller DC, Antczak DF, Bertelsen MF, Brunak S, Al-Rasheid KA, Ryder O, Andersson L, Mundy J, Krogh A, Gilbert MT, Kjaer K, Sicheritz-Ponten T, Jensen LJ, Olsen JV, Hofreiter M, Nielsen R, Shapiro B, Wang J, Willerslev E. 2013. Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse. Nature 499:74-8.
26. Martin J, Herniou E, Cook J, Waugh O'Neill R, Tristem M. 1997. Human endogenous retrovirus type I-related viruses have an apparently widespread distribution within vertebrates. J Virol 71:437-43.
27. Lower R, Lower J, Kurth R. 1996. The viruses in all of us: characteristics and biological significance of human endogenous retrovirus sequences. Proc Natl Acad Sci U S A 93:5177-84.
28. Benit L, Lallemand JB, Casella JF, Philippe H, Heidmann T. 1999. ERV-L elements: a family of endogenous retrovirus-like elements active throughout the evolution of mammals. J Virol 73:3301-8.

29. Lee A, Nolan A, Watson J, Tristem M. 2013. Identification of an ancient endogenous retrovirus, predating the divergence of the placental mammals. Philos Trans R Soc Lond B Biol Sci 368:20120503.
30. Belshaw R, Watson J, Katzourakis A, Howe A, Woolven-Allen J, Burt A, Tristem M. 2007. Rate of recombinational deletion among human endogenous retroviruses. J Virol 81:9437-42.
31. Grandi N, Cadeddu M, Blomberg J, Mayer J, Tramontano E. 2018. HERV-W group evolutionary history in non-human primates: characterization of ERV-W orthologs in Catarrhini and related ERV groups in Platyrrhini. BMC Evol Biol 18:6.
32. Sookdeo A, Hepp CM, Boissinot S. 2018. Contrasted patterns of evolution of the LINE-1 retrotransposon in perissodactyls: the history of a LINE-1 extinction. Mob DNA 9:12.
33. Subramanian RP, Wildschutte JH, Russo C, Coffin JM. 2011. Identification, characterization, and comparative genomic distribution of the HERV-K (HML-2) group of human endogenous retroviruses. Retrovirology 8:90.
34. Stefanetti V, Marenzoni ML, Passamonti F, Cappelli K, Garcia-Etxebarria K, Coletti M, Capomaccio S. 2016. High Expression of Endogenous Retroviral Envelope Gene in the Equine Fetal Part of the Placenta. PLoS One 11:e0155603.
35. Scott EY, Mansour T, Bellone RR, Brown CT, Mienaltowski MJ, Penedo MC, Ross PJ, Valberg SJ, Murray JD, Finno CJ. 2017. Identification of long non-coding RNA in the horse transcriptome. BMC Genomics 18:511.
36. dos Reis M, Inoue J, Hasegawa M, Asher RJ, Donoghue PC, Yang Z. 2012. Phylogenomic datasets provide both precision and accuracy in estimating the timescale of placental mammal phylogeny. Proc Biol Sci 279:3491-500.
37. Vilstrup JT, Seguin-Orlando A, Stiller M, Ginolhac A, Raghavan M, Nielsen SC, Weinstock J, Froese D, Vasiliev SK, Ovodov ND, Clary J, Helgen KM, Fleischer RC, Cooper A, Shapiro B, Orlando L. 2013. Mitochondrial phylogenomics of modern and ancient equids. PLoS One 8:e55950.
38. Henzy JE, Gifford RJ, Johnson WE, Coffin JM. 2014. A novel recombinant retrovirus in the genomes of modern birds combines features of avian and Mammalian retroviruses. J Virol 88:2398-405.
39. Huder JB, Boni J, Hatt JM, Soldati G, Lutz H, Schupbach J. 2002. Identification and characterization of two closely related unclassifiable endogenous retroviruses in pythons (Python molurus and Python curtus). J Virol 76:7607-15.
40. Petropoulos C. 1997. Retroviral Taxonomy, Protein Structures, Sequences, and Genetic Maps. *In* Coffin JM, Hughes SH, Varmus HE (ed).
41. Katzourakis A, Magiorkinis G, Lim AG, Gupta S, Belshaw R, Gifford R. 2014. Larger mammalian body size leads to lower retroviral activity. PLoS Pathog 10:e1004214.
42. Aiewsakun P, Katzourakis A. 2016. Time-Dependent Rate Phenomenon in Viruses. J Virol 90:7184-95.
43. Herniou E, Martin J, Miller K, Cook J, Wilkinson M, Tristem M. 1998. Retroviral diversity and distribution in vertebrates. J Virol 72:5955-66.
44. Gifford R, Kabat P, Martin J, Lynch C, Tristem M. 2005. Evolution and distribution of class II-related endogenous retroviruses. J Virol 79:6478-86.
45. Hayward A, Cornwallis CK, Jern P. 2015. Pan-vertebrate comparative genomics unmasks retrovirus macroevolution. Proc Natl Acad Sci U S A 112:464-9.
46. Zhuo X, Feschotte C. 2015. Cross-Species Transmission and Differential Fate of an Endogenous Retrovirus in Three Mammal Lineages. PLoS Pathog 11:e1005279.

47. Wildschutte JH, Williams ZH, Montesion M, Subramanian RP, Kidd JM, Coffin JM. 2016. Discovery of unfixed endogenous retrovirus insertions in diverse human populations. Proc Natl Acad Sci U S A 113:E2326-34.
48. Dupressoir A, Vernochet C, Bawa O, Harper F, Pierron G, Opolon P, Heidmann T. 2009. Syncytin-A knockout mice demonstrate the critical role in placentation of a fusogenic, endogenous retrovirus-derived, envelope gene. Proc Natl Acad Sci U S A 106:12127-32.
49. Mi S, Lee X, Li X, Veldman GM, Finnerty H, Racie L, LaVallie E, Tang XY, Edouard P, Howes S, Keith JC, Jr., McCoy JM. 2000. Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. Nature 403:785-9.
50. Nethe M, Berkhout B, van der Kuyl AC. 2005. Retroviral superinfection resistance. Retrovirology 2:52.
51. Robinson HL, Lamoreux WF. 1976. Expression of endogenous ALV antigens and susceptibility to subgroup E ALV in three strains of chickens (endogenous avian C-type virus). Virology 69:50-62.
52. Spencer TE, Mura M, Gray CA, Griebel PJ, Palmarini M. 2003. Receptor usage and fetal expression of ovine endogenous betaretroviruses: implications for coevolution of endogenous and exogenous retroviruses. J Virol 77:749-53.
53. Blanco-Melo D, Gifford RJ, Bieniasz PD. 2017. Co-option of an endogenous retrovirus envelope for host defense in hominid ancestors. Elife 6.
54. Hizi A, Herzig E. 2015. dUTPase: the frequently overlooked enzyme encoded by many retroviruses. Retrovirology 12:70.
55. Yoon JH, Abdelmohsen K, Gorospe M. 2013. Posttranscriptional gene regulation by long noncoding RNA. J Mol Biol 425:3723-30.
56. MacFadden BJ. 1994. Fossil horses: systematics, paleobiology, and evolution of the family Equidae. Cambridge University Press.
57. Krueger F. 2017. Trim Galore!, v0.4.5. https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/.
58. Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nature Methods 9:357.
59. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078-9.
60. Zhu H, Dennis T, Hughes J, Gifford RJ. 2018. Database-integrated genome screening (DIGS): exploring genomes heuristically using sequence similarity search tools and a relational database. bioRxiv doi:10.1101/246835.
61. Ellinghaus D, Kurtz S, Willhoeft U. 2008. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinformatics 9:18.
62. Bao W, Kojima KK, Kohany O. 2015. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob DNA 6:11.
63. Steinbiss S, Willhoeft U, Gremme G, Kurtz S. 2009. Fine-grained annotation and classification of de novo predicted LTR retrotransposons. Nucleic Acids Res 37:7002-13.
64. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A, Salazar GA, Tate J, Bateman A. 2016. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res 44:D279-85.
65. Chan PP, Lowe TM. 2009. GtRNAdb: a database of transfer RNA genes detected in genomic sequence. Nucleic Acids Res 37:D93-7.
66. Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312-3.

67. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods 14:587-589.
68. Dimmic MW, Rest JS, Mindell DP, Goldstein RA. 2002. rtREV: an amino acid substitution matrix for inference of retrovirus and reverse transcriptase phylogeny. J Mol Evol 55:65-73.
69. Blond JL, Beseme F, Duret L, Bouton O, Bedin F, Perron H, Mandrand B, Mallet F. 1999. Molecular characterization and placental expression of HERV-W, a new human endogenous retrovirus family. J Virol 73:1175-85.
70. Katzourakis A, Tristem M. 2005. Phylogeny of human endogenous and exogenous retroviruses. Retroviruses and primate genome evolution:186-203.
71. Brown K, Emes RD, Tarlinton RE. 2014. Multiple groups of endogenous epsilon-like retroviruses conserved across primates. J Virol 88:12464-71.
72. Vargiu L, Rodriguez-Tome P, Sperber GO, Cadeddu M, Grandi N, Blikstad V, Tramontano E, Blomberg J. 2016. Classification and characterization of human endogenous retroviruses; mosaic forms are common. Retrovirology 13:7.
73. Kumar S, Stecher G, Suleski M, Hedges SB. 2017. TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. Mol Biol Evol 34:1812-1819.

Journal of Virology

JVI

Journal of Virology