An endogenous lentivirus in the germline of a rodent.

Roziah Kambol[1], Anna Gatseva[2], and Robert J. Gifford[2]

[1]*School of Biological Sciences, Faculty of Applied Sciences, University Teknologi MARA, Shah Alam, 40450 Selangor*

[2]*MRC-University of Glasgow Centre for Virus Research, 464 Bearsden Rd, Bearsden, Glasgow, UK, G61 1QH*

ABSTRACT

Lentiviruses (genus *Lentivirus*) are complex retroviruses that infect a broad range of mammals, including humans. Unlike many other retrovirus genera, lentiviruses have only rarely been incorporated into the mammalian germline. However, a small number of endogenous retrovirus (ERV) lineages have been identified, and these rare genomic "fossils" can provide crucial insights into the long-term history of lentivirus evolution. Here, we describe a previously unreported endogenous lentivirus lineage in the genome of the South African springhare (*Pedetes capensis*), demonstrating that the host range of lentiviruses has historically extended to rodents (order Rodentia). Furthermore, through comparative and phylogenetic analysis of lentivirus and ERV genomes, considering the biogeographic and ecological characteristics of host species, we reveal broader insights into the long-term evolutionary history of the genus.

INTRODUCTION

The lentiviruses (genus *Lentivirus*) are an unusual group of retroviruses (family *Retroviridae*) that infect mammals and are associated with a range of slow, progressive diseases in their respective host species groups [1] (**Table 1**). They are most familiar as the genus of retroviruses that includes human immunodeficiency virus type 1 (HIV-1), but the group also includes viruses that infect a broad range of other mammalian groups. Lentiviruses are distinguished from other retroviruses by several characteristic features, including several unique accessory genes, a characteristic nucleotide composition [2, 3], and the capacity to infect non-dividing target cells [4].

All retroviruses replicate via an obligate step in which a DNA copy of the viral genome is integrated into a host cell chromosome [5]. The integrated viral genome is flanked at either side by identical long terminal repeat (LTR) sequences (a form referred to as a 'provirus'), each composed of functionally distinct U3, R and U5 regions. Occasionally, germline cells may be infected and subsequently go on to form viable progeny, so that integrated retroviral proviruses are vertically inherited as host alleles [6]. Such endogenous retroviruses (ERV) insertions are relatively common features in vertebrate genomes [7, 8]. Phylogenetic studies indicate that, following genome invasion, ERVs can increase their germline copy number through a variety of mechanisms, including active replication [9]. However, reflecting their ancient origins, most ERV insertions are genetically fixed and highly degraded by germline mutation. Furthermore, deletion of the entire internal region occurs frequently via homologous recombination between proviral LTRs, leaving behind a single LTR sequence or 'solo LTR' [10]. Despite being highly degraded, however, ERVs provide a useful source of retrospective information about the long-term evolutionary interactions between retroviruses and their hosts [11]. For example, identification of orthologous ERV insertions in related species provides a robust means of deriving minimum age calibrations for retrovirus groups, based on host species divergence estimates (which are in part informed by the fossil record) [12]. More broadly, ERV sequences can be used to explore the long-term evolutionary history of ancient - presumably extinct - retrovirus groups

[13, 14], and to inform our understanding of their interactions with host genes [15]. ERV sequences can even be used to guide the reconstitution of ancient retrovirus proteins so that their biological properties may be empirically investigated *in vitro* [16-18].

Lentiviruses have only rarely been incorporated into the germline of host species. However, a handful of Lentivirus-derived ERV lineages have now been identified (**Table 1**), and these sequences demonstrate that viruses clearly recognisable as lentiviruses circulated in mammals many millions of years ago. For example, rabbit endogenous lentivirus K (RELIK) insertions were found to occur at orthologous positions in the rabbit (*Oryctolagus cuniculus*) and hare (*Lepus europaeus*) genomes, demonstrating that genome invasion occurred prior to divergence of these species ~12 million years ago (Mya) [12, 19]. Endogenous lentiviruses have also been identified in lemurs (family Lemuridae) [20, 21]; mustelids (family Mustelidae) [22, 23]; and dermopterans (order Dermoptera - a group of arboreal gliding mammals native to Southeast Asia) [24-26]. Together, these sequences provide a range of minimum age calibrations in the Miocene epoch (23.5-5.3 Mya), based on host species divergence date estimates derived from the fossil record [11, 22, 25]. Widespread circulation among mammals is further supported by estimates derived via application of a molecular clock, some of which extend into the Eocene epoch (56-33.9 Mya) [24, 26].

In this study we perform comprehensive screening of published mammalian genomes and identify a previously unreported endogenous lentivirus lineage in the genome of the South African springhare (*Pedetes capensis*), demonstrating that lentivirus host range extends to rodents. Furthermore, through comparative and phylogenetic analysis, incorporating all available data, we provide broader insight into the origins and long-term evolutionary history of lentiviruses.

MATERIALS & METHODS

Genome screening in silico

We used database-integrated genome screening (DIGS) [27] to derive a non-redundant database of lentivirus-derived ERV loci contained in published genome sequence assemblies. In DIGS, the output of systematic, sequence similarity search-based 'screens' is captured in a relational database. The DIGS tool [27] is a Perl-based framework in which the Basic Local Alignment Search Tool (BLAST) program suite (version 2.2.31+) [28] is used to perform systematic similarity searches of sequence databases (e.g., genome assemblies) and the MySQL relational database management system (MySQL Community Server version 8.0.30) is used to record and organise output data. WGS data of 431 mammalian species were obtained from the National Center for Biotechnology Information (NCBI) genome database [29] (**Table S1**). Query polypeptide sequences were derived from representative lentivirus species (**Table 1**). DNA sequences in WGS assemblies that disclosed significant similarity to lentivirus queries (as determined by BLAST *e*-value) were classified via comparison to published retrovirus genome sequences (again using BLAST). Consensus genome sequences for endogenous lentivirus lineages were extracted from the supplementary material of associated publications, as follows: RELIK [19]; PSIV1 [20]; PSIV2 [21]; MELV [22]; DELV [24].

We compiled a set of endogenous lentivirus loci (**Table S2**) by using structured query language) to filter screening the classified, non-redundant results of >130,000 searches, selecting matches based on their degree of similarity to lentivirus reference sequences, or the taxonomic characteristics of the species in which they occur. Using this approach we separated putatively novel lentivirus ERV loci from both (i) orthologs or paralogs of previously characterised lentivirus ERVs, and (ii) non-lentiviral sequences that cross-matched to lentivirus probes due to shared ancestry (e.g., clade II ERVs) [30, 31]. We confirmed that putative novel lentivirus ERVs were indeed derived from lentiviruses (rather than other, related retroviruses) through phylogenetic and genomic analysis as described below.

Phylogenetic and genomic analysis

Nucleotide and protein phylogenies were reconstructed using maximum likelihood (ML) as implemented in RAxML (version 8.2.12) [32]. Protein substitution models were selected via hierarchical maximum likelihood ratio test using the PROTAUTOGAMMA option in RAxML. To estimate the ages of solo LTRs we measured divergence from an LTR consensus sequence and applied a neutral rate calibration, as described by Subramanian *et al.* [33]. We used Se-Al (version 2.0) to visualise alignments and create consensus sequences [34].

RESULTS & DISCUSSION

We systematically screened WGS data representing 431 mammalian species (**Table S1**) for endogenous lentivirus loci using similarity search-based approaches We identified a total of 842 distinct lentivirus-derived ERV loci, most of which represented members of previously described lentivirus ERV lineages (**Table 2,** [35]). However, we also identified lentivirus-derived sequences in the genome of a species group in which they have not previously been described – rodents (order Rodentia).

Matches to lentiviral Gag and Pol proteins were identified in WGS data of the South African springhare (*Pedetes capensis*), and phylogenetic analysis of the reverse transcriptase (RT) coding region encoded by these ERVs demonstrates that they group within the diversity of previously described lentivirus species (**Fig. S1a**). Initially, only four copies of Springhare endogenous lentivirus (SpELV) were identified in the *P. capensis* genome. However, we were able to identify the 5' LTR of a partial provirus sequence by using upstream flanking sequence as a query in BLASTn-based searches of the *P. capensis* genome assembly. This revealed the presence of a repetitive sequence showing the characteristic features of a retroviral LTR (i.e., ~500 nucleotides in length with terminal TG and CA dinucleotides) in the expected position upstream of the Gag ORF. Using this LTR sequence as input for screening enabled us to identify another 10 SpELV loci represented by solo LTR sequences (**Table 3)**. We generated a consensus SpELV genome using all fourteen loci identified in our screen (**Fig. S2**). We did not identify an envelope (*env*) gene

associated with any SpELV insertions, nor did we identify any contigs containing complete proviruses with paired LTR sequences. Furthermore, because the longest provirus sequence we identified was truncated in *pol* we could not determine whether any accessory genes might have been encoded downstream of this gene. Nonetheless, the partial genome obtained in our analysis exhibits the characteristic features of lentivirus genomes, including (i) a primer-binding site specific for tRNA Lysine (**Fig. S3**); (ii) a Pro-Pol ORF expressed via -1 ribosomal frameshifting (**Fig. S3**); (iii) an adenine-rich (34%) genome (**Fig. S4**) containing few CpG dinucleotides (0.29%); (iv) a putative *trans*-activator response (TAR) element (**Fig. S2, Fig. S3**). We estimated the age of the SpELV lineage utilising a molecular clock-based approach in which divergence is calculated by comparing individual LTR sequences to an LTR consensus [33]. We obtained age estimates in the range of 8-18 Mya for SpELV loci (**Table 3**), consistent with an origin in the Middle Miocene.

We used maximum likelihood-based phylogenetic approaches to reconstruct the evolutionary relationships between contemporary lentiviruses and the extinct lentiviruses represented by ERVs. Phylogenetic trees clearly separate the Lentiviruses into two robustly supported subclades (**Fig. 1**). One (here labelled 'Archaeolentivirus') contains SpELV together with dermopteran endogenous lentivirus (DELV) which occurs in the germline of colugos [24-26]. A second (here labelled 'Neolentivirus') contains all other endogenous lentivirus lineages and all known contemporary lentiviruses. We obtained relatively high support for internal branching relationships within the Neolentivirus clade – reconstructions support the existence of a distinct 'primate' group of neolentiviruses containing both simian and prosimian sub-lineages, and an 'artiodactyl' group incorporating both the bovine lentiviruses and the small ruminant lentiviruses. In addition, the primate lentiviruses group separately from all other neolentiviruses, which together constitute a 'grasslands-associated' clade comprised of lentiviruses that infect(ed) grassland-adapted host species.

To examine the distribution and diversity of lentiviruses in the context of host evolution, we plotted information related to (i) lentivirus distribution and (ii) host biogeographic range onto a time-calibrated phylogeny of boreoeutherian hosts (**Fig. 2**). This

revealed that age estimates obtained from the genomic fossil record (either through the identification of ancient orthologs, or via the application of a molecular clock) are consistent with other calibrations in deep time that can be tentatively inferred from ancestral biogeographic distributions by parsimoniously assuming limited transfer of virus between major biogeographic regions and distantly related host groups. Lentiviruses are known to cross species barriers quite frequently [36, 37], but transmission between large phylogenetic distances (e.g., distinct taxonomic orders of mammals) has never been reported and is unlikely to be common based on current understanding of the barriers to zoonotic transfer [38]. Evidence from orthology and molecular clock-based analyses supports the presence of DELV in Asia (the only region where colugos occur) up to 60 Mya – i.e., throughout most of the Cenozoic Era [26]. identification of a DELV-related virus in springhares – which evolved in the African subcontinent – implies the presence of archaeolentiviruses in ancestral mammals >80 Mya [39]. Notably, other groupings within the Neolentivirus subclade are also consistent with late Cretaceous origins predating the subordinal divergences of major placental mammal groups. For example, the existence of an ancient primate lineage, incorporating both lemurs, apes, and monkeys (**Fig. 1**) is consistent with a parsimonious scenario under which lentiviruses were present in the common ancestor of all primates and arrived in Madagascar with founder populations of ancestral lemurs ~60 Mya [40, 41].

At the same time, it seems clear that transmission of lentiviruses between phylogenetically distant mammal groups has occurred in the past. For example, the 'grasslands-associated' subclade contains viruses and paleoviruses that infect (or infected) phylogenetically distinct host species groups that share grassland habitat. It includes equine infectious anaemia virus (EIAV) which infects horses, and two ERV lineages - RELIK (found in leporids) and MELV (found in mustelids) [42-44]. Notably, the grassland adaptation of these three host species groups took place in a similar time-period (early-to-middle Miocene) in interconnected biogeographic areas (Laurasia and Africa) [42-44] (**Fig. 2**), suggesting that the connections between the viruses in this clade could reflect inter-order transmission events that took place in a shared habitat.

CONCLUSIONS

We describe a novel endogenous lineage in the genome of the South African springhare. The identification of SpELV demonstrates that lentivirus host range has historically extended to rodents. Through comparative and phylogenetic analysis of modern and ancient lentivirus genomes, we show that the *Lentivirus* genus incorporates at least two major subclades and reveal evidence it originated >80 Mya. In addition, we reveal phylogenetic evidence that transmission of lentiviruses between distantly related mammalian groups (i.e., distinct orders) has historically occurred in shared habitats.

Table 1. Reference genome sequences of representative lentivirus species

Species	Host species		Abbreviation / sub-strain	Source[a]
Exogenous viruses				
Jembrana disease virus	Gaur	*Bos gaurus*	JDV	NC_001654
Bovine immunodeficiency virus	Domestic cattle	*Bos taurus*	BIV	M32690
Small ruminant lentivirus genotype A	Goats & sheep		SRLV-A	NC_001452
Small ruminant lentivirus genotype B	Goats & sheep		SRLV-B	NC_001463
Equine infectious anemia virus	Domestic horse	*Equus cabalus*	EIAV	M16575
Feline immunodeficiency virus	Domestic cat	*Felis catus*	FIV-fca	M25381
	Pallas's cat	*Otocolobus manul*	FIV-oma	U56928
	Puma	*Puma concolor*	FIV-pco	EF455613
Simian immunodeficiency virus	Spot-nosed monkeys	*Cercopithecus nictitans*	SIV-gsn	AF468659
	Colobus monkey	*Colobus guereza*	SIV-col	AF301156
	Sykes' monkey	*Cercopithecus albogularis*	SIV-syk	L06042
	Dent's monkey	*Cercopithecus denti*	SIV-den	AJ580407
	Drill monkey	*Mandrillus leucophaeus*	SIV-drl	AY159321
	Green monkey	*Chlorocebus sabaeus*	SIV-agm-sab	U04005
	Mangabey	*Cercocebus torquatus*	SIV-rcm	HM803689
	Central chimpanzee	*Pan troglodytes*	SIV-cpz-ptt	AF103818
	Sooty mangabey	*Cercocebus atys atys*	SIV-smm	X14307
	Sun-tailed monkey	*Cercopithecus solatus*	SIV-sun	AF131870
	Mandrill	*Mandrillus sphinx*	SIV-mnd-1	M27470
Human immunodeficiency virus 1*	Human	*Homo sapiens*	HIV-1	AF033819
Human immunodeficiency virus 2	Domestic cattle	*Bos taurus*	HIV-2A	X05291
Endogenous viruses				
Rabbit endogenous lentivirus K	Leporids (rabbits & hares)		RELIK	[19]
Prosimian immunodeficiency virus 1	Mouse lemurs		PSIV1	[20]
Prosimian immunodeficiency virus 2	Dwarf lemurs		PSIV2	[21]
Mustelid endogenous lentivirus	Mustelids (subclade)		MELV	[22]
Dermopteran endogenous lentivirus	Colugos		DELV	[24]

Footnote: [a] GenBank accession numbers are given for exogenous viruses. For endogenous lentivirus lineages consensus genome sequences were extracted from the publication shown.

Table 2. Endogenous lentivirus loci detected via screening.

Organism	Common name	ERV lineage	Count
Oryctolagus cuniculus	European rabbit	RELIK	203
Lepus americanus	Snowshoe hare	RELIK	212
Lepus timidus	European hare	RELIK	121
Sylvilagus bachmani	Brush rabbit	RELIK	159
Microcebus griseorufus	Reddish-grey mouse lemur	PSIV1	20
Microcebus mittermeieri	Mittermeier's mouse lemur	PSIV1	20
Microcebus murinus	Grey mouse lemur	PSIV1	11
Microcebus ravelobensis	Golden-brown mouse lemur	PSIV1	21
Microcebus tavaratra	Northern rufous mouse lemur	PSIV1	20
Cheirogaleus medius	Fat-tailed dwarf lemur	PSIV2	1
Mustela erminea	Stoat	MELV	14
Mustela putorius	Ferret	MELV	18
Neovison vison	Mink	MELV	12
Galeopterus variegatus	Sunda colugo	DELV	6
Pedetes capensis	South African springhare	SpELV	14

Footnote. Abbreviations**:** RELIK=Rabbit endogenous lentivirus K; PSIV=Prosimian immunodeficiency virus ; MELV=Mustelid endogenous lentivirus; DELV=Dermopteran endogenous lentivirus; SpELV=springhare endogenous lentivirus.

Table 3. Springhare endogenous lentivirus loci

ERV locus[a]	Structure	Age (Mya)[b]	Contig	Orientation	Start	End
SpELV.1-PedCap	LTR-Gag-Pol	8,750,000	VMDO01011022.1	+	63971	68252
SpELV.2-PedCap	LTR-Gag	12,500,000	VMDO01050306.1	+	2112	3978
SpELV.3-PedCap	Gag	*ND*	VMDO01082106.1	-	284	997
SpELV.4-PedCap	Gag	*ND*	VMDO01088624.1	-	445	1155
SpELV.6-PedCap	LTR	10,084,925	VMDO01001729.1	-	65134	65652
SpELV.14-PedCap	LTR	11,194,029	VMDO01006229.1	+	41369	41857
SpELV.15-PedCap	LTR	11,460,554	VMDO01001857.1	+	174371	174854
SpELV.16-PedCap	LTR	11,460,554	VMDO01001488.1	+	108014	108496
SpELV.17-PedCap	LTR	8,528,784	VMDO01000902.1	+	172296	172776
SpELV.18-PedCap	LTR	11,143,410	VMDO01003412.1	+	3534	4013
SpELV.21-PedCap	LTR	9,594,882	VMDO01035581.1	-	11304	11783
SpELV.24-PedCap	LTR	11,388,286	VMDO01046849.1	+	1434	1906
SpELV.25-PedCap	LTR	13,592,750	VMDO01015080.1	+	21755	22226
SpELV.28-PedCap	LTR	17,665,952	VMDO01001033.1	-	148080	148542

Footnote: [a] Loci were assigned unique identifiers (IDs) following a standard nomenclature system [45]. PedCap=*Pedetes capensis*. The ages of LTR-encoding elements was estimated by measuring divergence from an LTR consensus sequence and applying a neutral rate calibration, as described by Subramanian *et al.* [33]. *ND*=not done. SpELV=Springhare endogenous lentivirus.

FIGURE LEGENDS

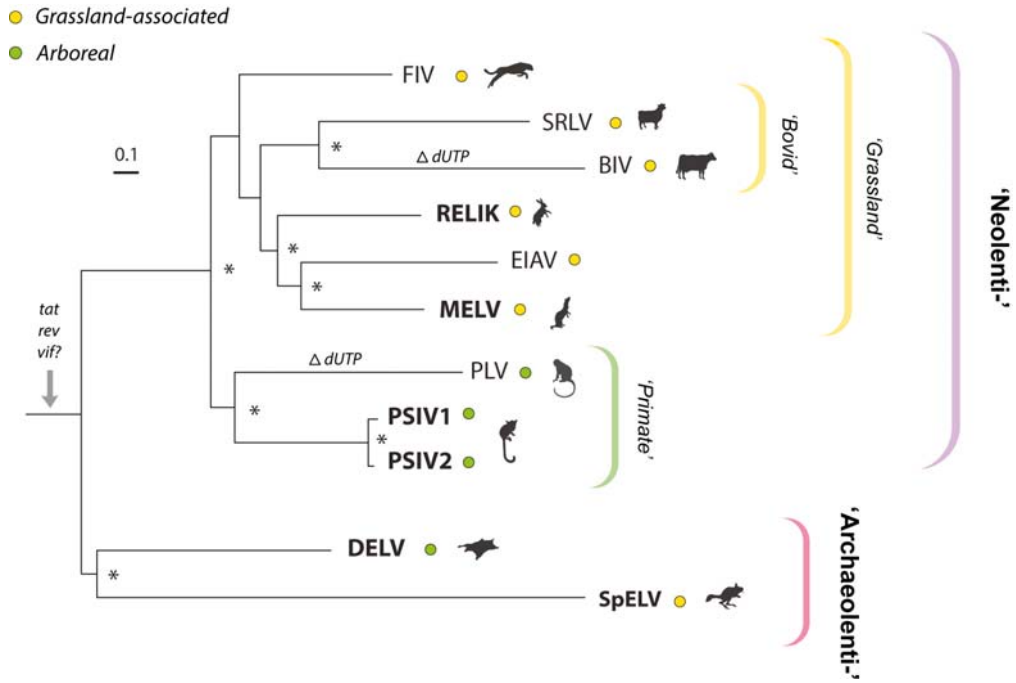


Figure 1. Phylogenetic relationships within the Lentivirus genus.

Maximum likelihood phylogeny showing reconstructed evolutionary relationships between all known lentivirus species, including the extinct species represented by endogenous lentiviruses. Brackets to the left indicate proposed subgroupings within genus *Lentivirus*. Coloured circles adjacent virus taxa labels indicate the ecological characteristics of the associated host species (grassland-dwelling or arboreal) as shown in the key top right. The phylogeny is midpoint rooted for display purposes and was reconstructed using a multiple sequence alignment spanning 1405 amino acid residues of the Gag-Pol polyprotein and the RT-Rev substitution model [46]. The scale bar shows evolutionary distance in substitutions per site. Asterisks indicate nodes with bootstrap support >70% (1000 replicates). **Abbreviations**: DELV=Dermopteran endogenous lentivirus; SpELV=springhare endogenous lentivirus. RELIK=Rabbit endogenous lentivirus type K; Mustelidae endogenous lentivirus (MELV); BIV=Bovine immunodeficiency virus; SIV=Simian immunodeficiency virus; EIAV=equine infectious anaemia virus; FIV=Feline immunodeficiency virus; Human immunodeficiency virus=HIV; SRLV=small ruminant lentivirus; PSIV=Prosimian immunodeficiency virus.

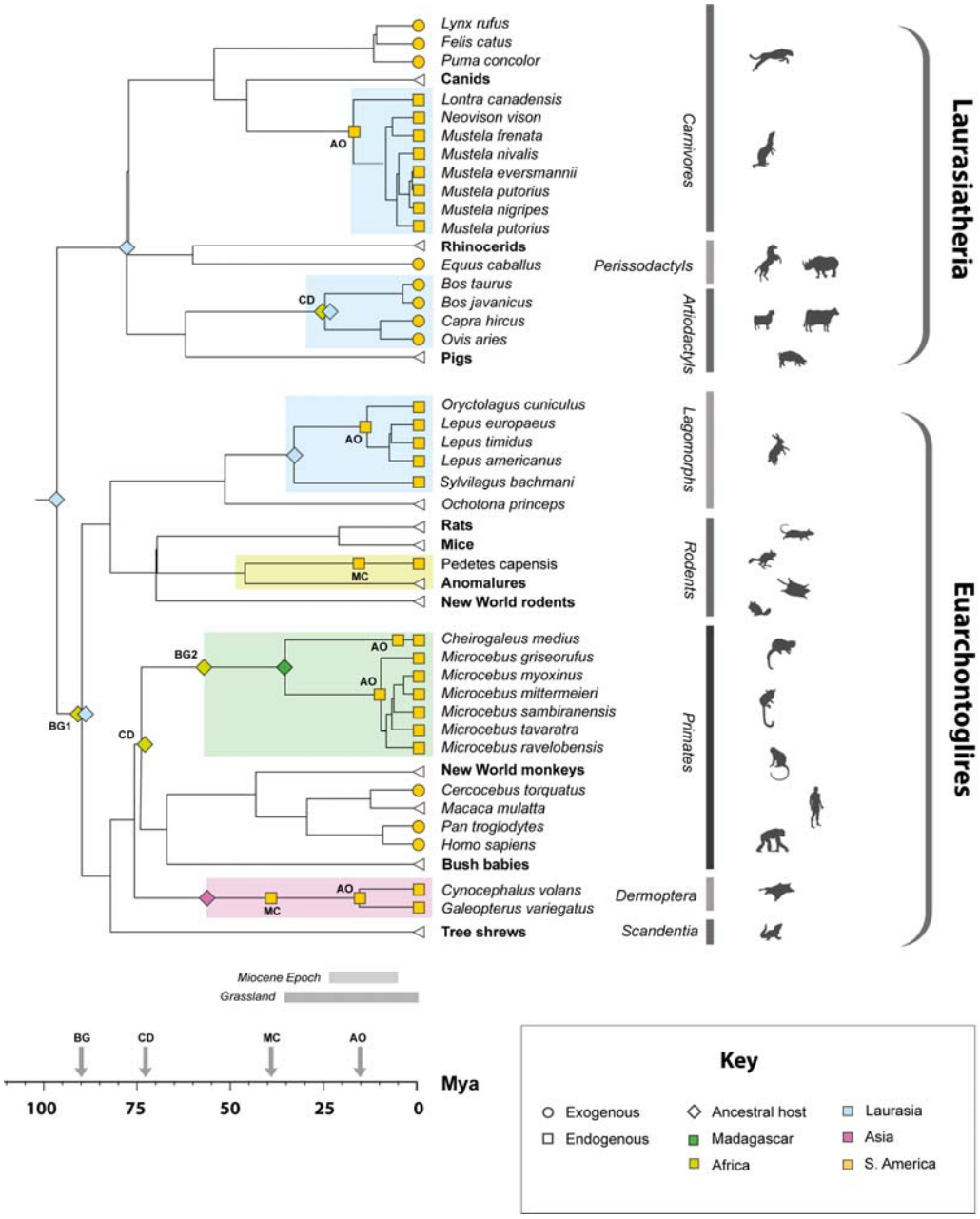


Figure 2. An updated timeline of lentivirus evolution. A time-calibrated phylogeny of mammalian species showing the known extent of association between lentiviruses and mammals, based on data obtained from TimeTree [47]. The scale bar shows time in millions of years before present. Brackets and bars to the right of taxa labels indicate host taxonomic groups. Coloured squares on terminal nodes indicate that host species associated with endogenous lentiviruses (squares) or exogenous lentiviruses (circles). The timeframe of endogenous lentivirus presence in each mammalian lineage is indicated by shaded boxes underneath clades, with colours indicating biogeographic associations of hosts within each clade following the key. White triangles at tree tips indicate host species or groups that have not yet been associated with any lentiviruses (endogenous or exogenous). Two-letter codes adjacent internal markers indicate the type of calibration being shown, as follows (AO=identification of an ancient ortholog; MC=application of a molecular clock to neutrally diverging sequences; CD=assumption of codivergence with hosts; BG=assumption of presence in biogeographic area inhabited by ancestor of species groups that are now biogeographically separated – note that this assumes no transfer between the respective regions identified by derived host species). Colours on diamond-shaped node markers indicate the known biogeographic range of ancestral hosts, as indicated in the key. The biogeographic range of the springhare-colugo ancestor (BG1) is uncertain (hence two regions are shown). The colonisation of isolated Madagascar by lemurs (BG2) is thought to have occurred ~60 million years ago (Mya) [40, 41].

SUPPLEMENTARY FIGURE LEGENDS

Figure S1. Phylogenetic and genomic characteristics of springhare endogenous lentivirus.

(a) Maximum likelihood (ML) phylogeny based on an alignment of reverse transcriptase (RT) protein sequences and showing the reconstructed evolutionary relationships between lentiviruses and other retroviruses. Asterisks indicate nodes with bootstrap support >70% (1000 replicates). The scale bar shows evolutionary distance in substitutions per site.

(b) ML phylogeny showing reconstructed evolutionary relationships between SpELV long terminal repeat (LTR) sequences. Numbers next to nodes indicate bootstrap support values (1000 replicates). The scale bar shows evolutionary distance in substitutions per site.

(c) Consensus genome structures of ancient lentiviral paleoviruses. **Abbreviations:** DELV=Dermopteran endogenous lentivirus; RELIK=Rabbit endogenous lentivirus type K; Mustelidae endogenous lentivirus (MELV); BIV=Bovine immunodeficiency virus; SIV=Simian immunodeficiency virus; FIV=Feline immunodeficiency virus; Human immunodeficiency virus=HIV; Prosimian immunodeficiency virus=PSIV; RV=Retrovirus; LV=Leukemia virus.

Figure S2. The SpELV consensus sequence.

Inverted repeats present at the ends of the 5' long terminal repeat (LTR) sequence are highlighted in light grey. Regions of nucleic acid secondary structure, the transactivation responsive (TAR) element and primer binding site (PBS) are highlighted in dark grey. The locations of the proteins encoded by the *gag* and *pol* genes were determined by homology to the DELV consensus sequence [24-26].

Figure S3. The putative SpELV TAR (transactivation responsive region) element

Secondary structures were predicted using the MFOLD thermodynamic folding algorithm [48] and assessed by comparison to well-characterised examples in other lentiviruses.

Figure S4. Nucleotide compositional bias in lentivirus genomes.

Nucleotide composition of whole genomes of Lentiviruses were normalised to length and plotted as percentages using R in R Studio (version 4.2.1). Reference genome sequences for each virus correspond to those given in **Table 1**. Bovine immunodeficiency virus (BIV), Dermopteran endogenous lentivirus (DELV), Equine infectious anaemia virus American strain (EIAV_Am), Feline immunodeficiency virus (FIV), Human immunodeficiency virus 1 (HIV_1M), Mustelidae endogenous lentivirus (MELV), Prosimian immunodeficiency virus 2 (PSIV); Rabbit endogenous lentivirus type K (RELIK), Springhare endogenous lentivirus (SpELV), Small ruminant lentivirus A (SRLV_A); Adenine (A), Guanine (G), Cytosine (C), Thymine (T).

DECLARATIONS

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

All data are openly available in the Lentivirus-GLUE project hosted on GitHub: *https://giffordlabcvr.github.io/Lentivirus-GLUE/*

Competing interests

None declared.

Funding

RJG is funded by the Medical Research Council of the United Kingdom (MC_UU_12014/12). NIH funding. The funding bodies had no role in the design of the study and collection, analysis, and interpretation of data, or in writing the manuscript.

Acknowledgments

We thank Daniel Blanco-Melo, Anne Emory, Ron Swanstrom and Greg Towers for helpful discussions.

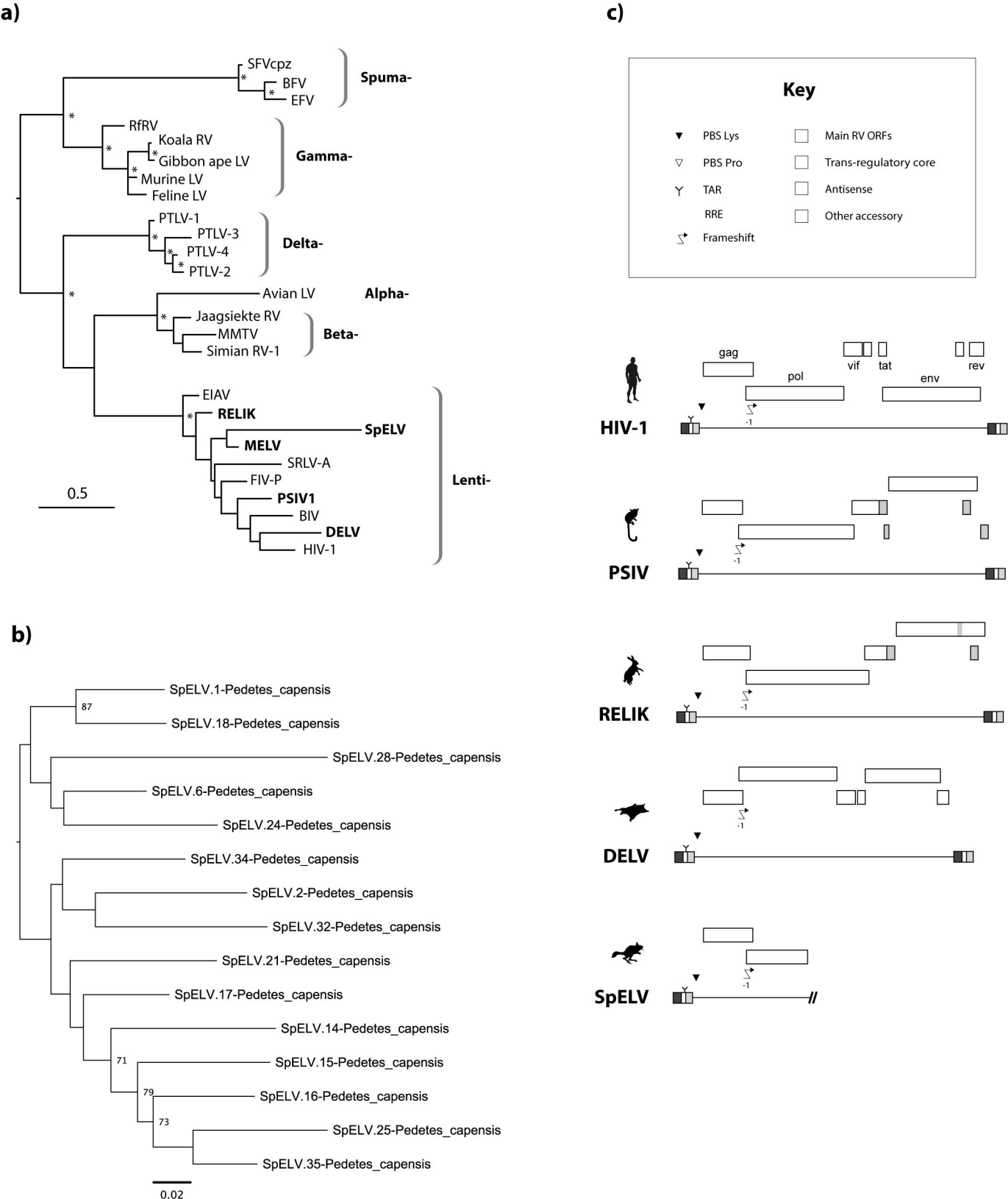
Authors' contributions

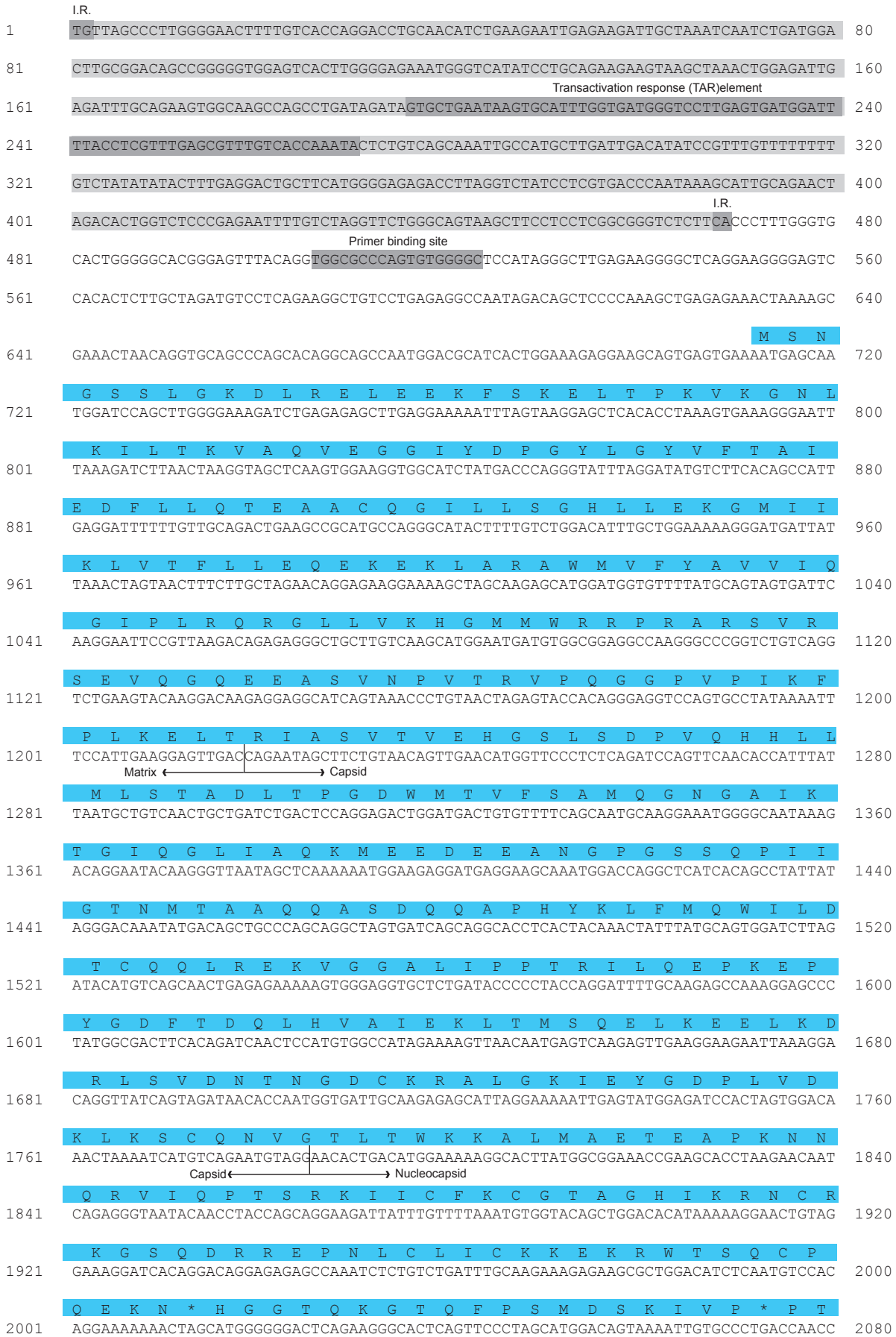
Conceptualization, R.J.G.; methodology and validation, A.G., R.K., and R.J.G.; formal analysis, A.G., R.J.G.; writing—original draft preparation, R.J.G.; writing—review and editing, A.G., R.J.G., and R.K.; visualization, A.G., R.J.G.; supervision, R.J.G.; project administration, R.J.G.; data curation, R.J.G. All authors have read and agreed to the published version of the manuscript.

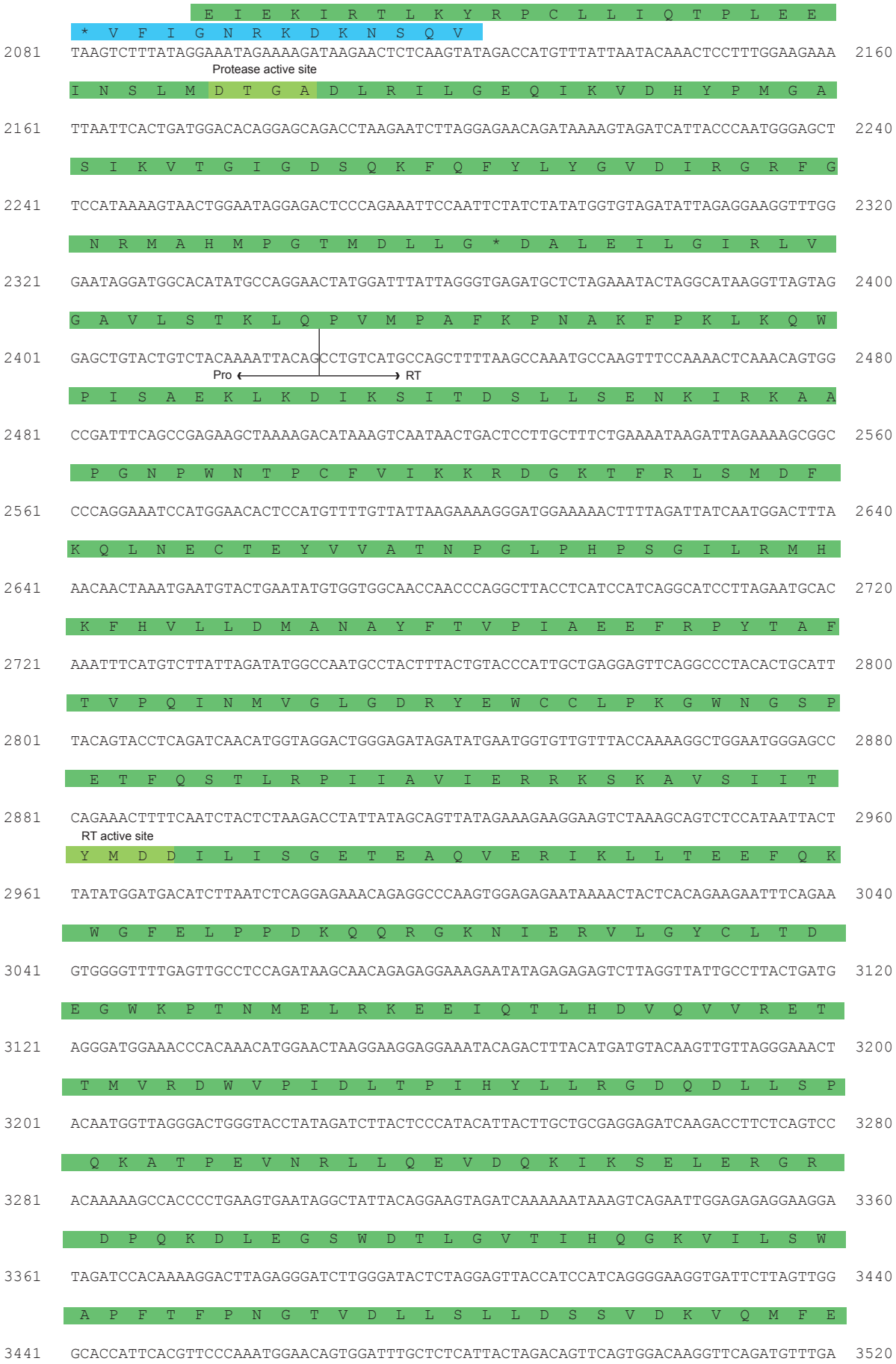
REFERENCES

1. Narayan, O. and J.E. Clements, *Biology and pathogenesis of lentiviruses.* J Gen Virol, 1989. **70 (Pt 7)**: p. 1617-39.
2. van der Kuyl, A.C. and B. Berkhout, *The biased nucleotide composition of the HIV genome: a constant factor in a highly variable virus.* Retrovirology, 2012. **9**: p. 92.
3. van Hemert, F., A.C. van der Kuyl, and B. Berkhout, *On the nucleotide composition and structure of retroviral RNA genomes.* Virus Res, 2014. **193**: p. 16-23.
4. Yamashita, M. and M. Emerman, *Retroviral infection of non-dividing cells: old and new perspectives.* Virology, 2006. **344**(1): p. 88-93.
5. Brown, P.O., *Integration*, in *Retroviruses*, J.M. Coffin, S.M. Hughes, and H.E. Varmus, Editors. 1997, Cold Spring Harbor Laboratory Press: Cold Spring Harbor.
6. Gifford, R. and M. Tristem, *The evolution, distribution and diversity of endogenous retroviruses.* Virus Genes, 2003. **26**(3): p. 291-315.
7. Hayward, A., C.K. Cornwallis, and P. Jern, *Pan-vertebrate comparative genomics unmasks retrovirus macroevolution.* Proc Natl Acad Sci U S A, 2015. **112**(2): p. 464-9.
8. Johnson, W.E., *Origins and evolutionary consequences of ancient endogenous retroviruses.* Nat Rev Microbiol, 2019. **17**(6): p. 355-370.
9. Stoye, J.P., *Studies of endogenous retroviruses reveal a continuing evolutionary saga.* Nat Rev Microbiol, 2012. **10**(6): p. 395-406.
10. Belshaw, R., et al., *Rate of recombinational deletion among human endogenous retroviruses.* J Virol, 2007. **81**(17): p. 9437-42.
11. Gifford, R.J., *Viral evolution in deep time: lentiviruses and mammals.* Trends Genet, 2012. **28**(2): p. 89-100.
12. Keckesova, Z., et al., *Identification of a RELIK orthologue in the European hare (Lepus europaeus) reveals a minimum age of 12 million years for the lagomorph lentiviruses.* Virology, 2009. **384**(1): p. 7-11.
13. Halo, J.V., et al., *Origin and recent expansion of an endogenous gammaretroviral lineage in domestic and wild canids.* Retrovirology, 2019. **16**(1): p. 6.
14. Diehl, W.E., et al., *Tracking interspecies transmission and long-term evolution of an ancient retrovirus using the genomes of modern mammals.* Elife, 2016. **5**: p. e12704.
15. Compton, A.A., H.S. Malik, and M. Emerman, *Host gene evolution traces the evolutionary history of ancient primate lentiviruses.* Philos Trans R Soc Lond B Biol Sci, 2013. **368**(1626): p. 20120496.
16. Dewannieux, M., et al., *Identification of an infectious progenitor for the multiple-copy HERV-K human endogenous retroelements.* Genome Res, 2006. **16**(12): p. 1548-56.
17. Blanco-Melo, D., R.J. Gifford, and P.D. Bieniasz, *Co-option of an endogenous retrovirus envelope for host defense in hominid ancestors.* Elife, 2017. **6**.
18. Goldstone, D.C., et al., *Structural and functional analysis of prehistoric lentiviruses uncovers an ancient molecular interface.* Cell Host Microbe, 2010. **8**(3): p. 248-59.
19. Katzourakis, A., et al., *Discovery and analysis of the first endogenous lentivirus.* Proc Natl Acad Sci U S A, 2007. **104**(15): p. 6261-5.
20. 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 U S A, 2008. **105**(51): p. 20362-7.
21. Gilbert, C., et al., *Parallel germline infiltration of a lentivirus in two Malagasy lemurs.* PLoS Genet, 2009. **5**(3): p. e1000425.
22. Han, G.Z. and M. Worobey, *Endogenous lentiviral elements in the weasel family (Mustelidae).* Mol Biol Evol, 2012. **29**(10): p. 2905-8.
23. Cui, J. and E.C. Holmes, *Endogenous lentiviruses in the ferret genome.* J Virol, 2012. **86**(6): p. 3383-5.
24. Hron, T., et al., *Endogenous lentivirus in Malayan colugo (Galeopterus variegatus), a close relative of primates.* Retrovirology, 2014. **11**(1): p. 84.

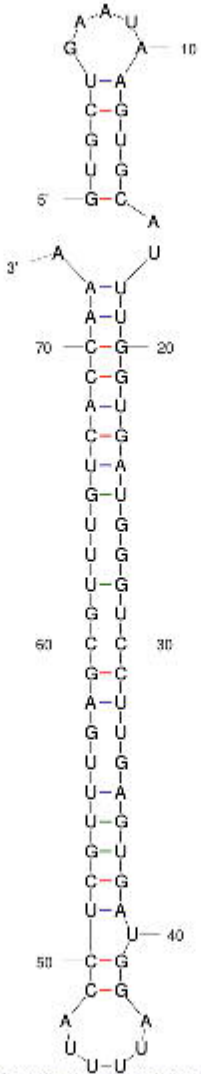
25. Han, G.Z. and M. Worobey, *A primitive endogenous lentivirus in a colugo: insights into the early evolution of lentiviruses.* Mol Biol Evol, 2015. **32**(1): p. 211-5.
26. Hron, T., et al., *Life History of the Oldest Lentivirus: Characterization of ELVgv Integrations in the Dermopteran Genome.* Mol Biol Evol, 2016. **33**(10): p. 2659-69.
27. Zhu, H., et al., *Database-integrated genome screening (DIGS): exploring genomes heuristically using sequence similarity search tools and a relational database.* bioRxiv, 2018.
28. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nuc. Acids Res., 1997. **25**: p. 3389-3402.
29. Kitts, P.A., et al., *Assembly: a resource for assembled genomes at NCBI.* Nucleic Acids Res, 2016. **44**(D1): p. D73-80.
30. Dewannieux, M. and T. Heidmann, *Endogenous retroviruses: acquisition, amplification and taming of genome invaders.* Curr Opin Virol, 2013. **3**(6): p. 646-56.
31. Gifford, R., et al., *Evolution and distribution of class II-related endogenous retroviruses.* J Virol, 2005. **79**(10): p. 6478-86.
32. Stamatakis, A., *RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies.* Bioinformatics, 2014. **30**(9): p. 1312-3.
33. Subramanian, R.P., et al., *Identification, characterization, and comparative genomic distribution of the HERV-K (HML-2) group of human endogenous retroviruses.* Retrovirology, 2011. **8**: p. 90.
34. Rambaut, A., *Se-Al: sequence alignment editor*. 1997: Edinburgh.
35. Gifford, R.J. *Lentivirus-GLUE*. 2021; Available from: https://giffordlabcvr.github.io/Lentivirus-GLUE.
36. Sharp, P.M. and B.H. Hahn, *The evolution of HIV-1 and the origin of AIDS.* Philos Trans R Soc Lond B Biol Sci, 2010. **365**(1552): p. 2487-94.
37. Kosugi, Y., et al., *Comprehensive Investigation on the Interplay between Feline APOBEC3Z3 Proteins and Feline Immunodeficiency Virus Vif Proteins.* J Virol, 2021. **95**(13): p. e0017821.
38. Albery, G.F., et al., *The science of the host-virus network.* Nat Microbiol, 2021. **6**(12): p. 1483-1492.
39. Springer, M.S., et al., *The historical biogeography of Mammalia.* Philos Trans R Soc Lond B Biol Sci, 2011. **366**(1577): p. 2478-502.
40. Karanth, K.P., et al., *Ancient DNA from giant extinct lemurs confirms single origin of Malagasy primates.* Proc Natl Acad Sci U S A, 2005. **102**(14): p. 5090-5.
41. Poux, C., et al., *Asynchronous colonization of Madagascar by the four endemic clades of primates, tenrecs, carnivores, and rodents as inferred from nuclear genes.* Syst Biol, 2005. **54**(5): p. 719-30.
42. Ge, D., et al., *Evolutionary history of lagomorphs in response to global environmental change.* PLoS One, 2013. **8**(4): p. e59668.
43. Toljagić, O., et al., *Millions of Years Behind: Slow Adaptation of Ruminants to Grasslands.* Systematic Biology, 2017. **67**(1): p. 145-157.
44. Law, C.J., *Evolutionary shifts in extant mustelid (Mustelidae: Carnivora) cranial shape, body size and body shape coincide with the Mid-Miocene Climate Transition.* Biol Lett, 2019. **15**(5): p. 20190155.
45. Gifford, R.J., et al., *Nomenclature for endogenous retrovirus (ERV) loci.* Retrovirology, 2018. **15**(1): p. 59.
46. Dimmic, M.W., et al., *rtREV: an amino acid substitution matrix for inference of retrovirus and reverse transcriptase phylogeny.* J Mol Evol, 2002. **55**(1): p. 65-73.
47. Kumar, S., et al., *TimeTree: A Resource for Timelines, Timetrees, and Divergence Times.* Mol Biol Evol, 2017. **34**(7): p. 1812-1819.
48. Zuker, M., *Mfold web server for nucleic acid folding and hybridization prediction.* Nucleic Acids Research, 2003. **31**(13): p. 3406-3415.


a)
SFVcpz
BFV
EFV
Spuma-
RfRV
Koala RV
Gibbon ape LV
Murine LV
Feline LV
Gamma-
PTLV-1
PTLV-3
PTLV-4
PTLV-2
Delta-
Avian LV
Alpha-
Jaagsiekte RV
MMTV
Simian RV-1
Beta-
EIAV
RELIK
SpELV
MELV
SRLV-A
FIV-P
PSIV1
BIV
DELV
HIV-1
Lenti-
0.5
b)
SpELV.1-Pedetes_capensis
87
SpELV.18-Pedetes_capensis
SpELV.28-Pedetes_capensis
SpELV.6-Pedetes_capensis
SpELV.24-Pedetes_capensis
SpELV.34-Pedetes_capensis
SpELV.2-Pedetes_capensis
SpELV.32-Pedetes_capensis
SpELV.21-Pedetes_capensis
SpELV.17-Pedetes_capensis
SpELV.14-Pedetes_capensis
71
SpELV.15-Pedetes_capensis
79
SpELV.16-Pedetes_capensis
73
SpELV.25-Pedetes_capensis
SpELV.35-Pedetes_capensis
0.02
c)
Key
PBS Lys
PBS Pro
TAR
RRE
Frameshift
Main RV ORFs
Trans-regulatory core
Antisense
Other accessory
gag
pol
vif
tat
env
rev
HIV-1
-1
PSIV
RELIK
DELV
SpELV
//





```
        L  L  R  Y  G  Y  E  S  K  I  I  N  Q  S  G  S  *  K  E  L  K  S  L  Q  M  Q

3521    ATTATTAAGATATGGATATGAGTCCAAGATAATAAACCAGTCAGGGTCCTAGAAAGAATTAAAAAGTTTACAAATGCAGG  3600

        D  V  W  P  T  R  W  T  Y  K  K  F  I  C  K  N  G  Q  G  W  T  I  G  L  S  N  L

3601    ATGTGTGGCCTACAAGATGGACATATAAGAAGTTTATTTGTAAAAATGGTCAGGGTTGGACTATAGGATTATCCAATCTA  3680

        L  Q  I  R  R  I  E  K  T  P  I  V  G  G  E  T  V  Y  T  D  A  S  R  L  R  K  T

3681    CTTCAGATTAGGAGGATTGAAAAAACACCTATTGTAGGAGGAGAGACTGTCTATACAGATGCATCGAGGTTGCGGAAAAC  3760

        N  H  K  R  I  A  W  Y  N  T  T  T  G  D  T  H  S  M  E  V  S  T  E  T  G  H

3761    TAACCACAAGAGAATAGCTTGGTACAACACAACCACGGGAGACACACATAGTATGGAAGTAAGCACAGAAACAGGACATG  3840

        A  *  Q  A  E  L  L  A  I  I  G  V  L  I  N  H  P  R  S  L  N  I  V  T  H  S  K

3841    CATGACAAGCAGAGCTCTTAGCAATTATAGGAGTACTAATTAATCATCCTAGGTCACTAAATATAGTAACACATAGCAAA  3920

        Y  I  A  A  F  L  P  K  I  G  G  H  E  R  N  N  L  W  Q  E  V  I  A  M  L  A  E

3921    TATATTGCAGCTTTTCTACCAAAAATAGGAGGACATGAGAGAAATAACCTATGGCAAGAAGTAATAGCCATGCTGGCAGA  4000

        R  V  K  Q  R  Y  R  T  F  V  S  W  V  P  G  H  S  E  V  Q  E  M

4001    AAGGGTAAAGCAAAGATATAGAACATTTGTTTCTTGGGTTCCTGGACACAGTGAGGTCCAGGAAATGA  4068
```



dG = -25.90 [initially -25.90] 22May23-17-26-10