Title: An endogenous amdoparvovirus in the genome of the Transcaucasian mole vole (*Ellobius lutescens*): implications for ecology and evolution of carnivore amdoparvoviruses.

Authors: Soledad Marsile-Medun[1,2] and Robert James Gifford[1]

Affiliations:

1. MRC-University of Glasgow Centre for Virus Research, 464 Bearsden Road, Glasgow, UK
2. Agrocampus Ouest, 65 Rue de Saint-Brieuc, 35000 Rennes, France

Corresponding author: Robert J. Gifford, MRC-University of Glasgow Centre for Virus Research, Glasgow, UK. E-mail: robert.gifford@glasgow.ac.uk

Abstract

Sequences derived from parvoviruses (family *Parvoviridae*) occur relatively frequently in eukaryotic genomes, and can be used to investigate the co-evolutionary history of parvoviruses and their hosts. Here, we report the identification of sequences derived from amdoparvoviruses in the genome of a rodent - the Transcaucasian mole vole (*Ellobius lutescens*). We recovered the putative ancestral coding sequences of these endogenous viral elements, and showed that they group robustly with exogenous amdoviruses in phylogenetic trees. We identified the corresponding empty genomic integration sites in the genome of a sister species - the northern mole vole (*Ellobius talpinus*) - indicating that both elements were integrated into the *Ellobius lutescens* germline within the last 10 million years. Our findings extend the host range of amdoparvoviruses to a second mammalian order (Rodentia), and indicate that the various carnivore amdoviruses that have so far been identified are likely to derive from independent cross-species transmission events.

Background

Amdoparvovirus is a newly defined parvoviral genus in the family *Parvoviridae* [1]. The type species - Aleutian mink disease virus (AMDV) - causes an immune complex-associated progressive syndrome in mink (Family *Mustelidae*: genera Neovison and Mustela) called Aleutian disease or plasmacytosis. First described in 1956, Aleutian disease is presently considered the most important infectious disease affecting farm-raised mink [2, 3]. Relatively little is known about the evolutionary history and natural biology of amdoparvoviruses. However,

AMDV infection is known to be widespread in wild mink as well as in farmed animals [2]. In recent years, amdoparvovirus infections have also been identified in raccoon dogs [4]. In addition, amdoparvovirus sequences have been recovered from tissues sampled from other carnivore species, including foxes and skunks [5-8].

Endogenous viral elements (EVEs) are sequences derived from virus genomes that have been integrated into the nuclear genome of germline cells. These sequences enable calibration of virus evolution across macroevolutionary timescales, and can provide insights into the long-term evolutionary associations between viruses and hosts [9, 10]. Progress in genome sequencing has revealed that EVEs derived from parvoviruses occur relatively frequently in eukaryotic genomes [11-14]. So far, however, no EVEs have been described that group closely with parvoviruses in phylogenetic trees. Here we describe the first clear example of an EVE derived from an amdoparvovirus, in the genome of the Transcaucasian mole vole (*Ellobius lutescens*).

Results

We screened 362 vertebrate genome assemblies (**Table S1**) for sequences closely related to Amdoparvoviruses. Searches using the replicase (NS) and capsid (VP) proteins of AMDV identified six distinct sequences that exhibited greater similarity to amdoparvoviruses than to other parvovirus groups (**Table 1**). One of these - in the Cape hyrax (*Procavia capensis*) - has been described previously [13] and spans a relatively short genomic region: ~100 amino acid (aa) residues in length. Similar length matches were identified in the Tasmanian devil (*Sarcophilus harrisii*), aardvark (*Orycteropus afer*), and pit viper (*Protobothrops mucrosquamatus*) genomes. The short length of these sequences - all of which grouped outside the Amdoparvovirus clade in phylogenies (**Figure S1**) - meant that it was not possible to determine whether they derive from viruses that would be considered *bona fide* members of the Amdoparvovirus genus, or from viruses belonging to a more distantly related, potentially extinct parvovirus lineage.

However, we identified two longer sequences in the genome of the Transcaucasian mole vole (*Ellobius lutescens*) that exhibited a much higher degree of similarity to AMDV. The first comprised a near complete genome (with both the viral replicase (NS) and capsid protein (VP) genes present), and the second spanned the majority of the NS gene (**Figure 1a**). We constructed maximum likelihood (ML) phylogenies using conserved regions of the putative NS and VP peptide sequences encoded by these elements. Phylogenies showed that both elements were relatively closely related, and grouped robustly within the clade defined by

exogenous amdoparvoviruses (**Figure 1b, Figure S1**). They were therefore labelled *Ellobius lutescens* endogenous amdoparvovirus (AmdoPVe-EllLut) 1 and 2.

In between the two major open reading frames (ORFs), NS and VP, amdoparvovirus genomes have a short middle ORF (M-ORF) of unknown function. As shown in the **Figure 1a**, AmPVe-EllLut.1 encodes both the NS and VP genes. A region of potentially protein-coding sequence that corresponds to the M-ORF of AMDV is also present. A methionine residue that might represent the start codon of an M-ORF gene product could not be identified (notably, this is also the case for several exogenous amdovirus isolates). Both the NS and VP ORFs of AmPVe-EllLut.1 have gaps relative to AMDV. The gap in NS is apparently due to a deletion (since the corresponding region is present in AmPVe-EllLut.2). The VP pseudogene encoded by AmPVe-EllLut.1 has a gap relative to the AMDV VP that spans most of the 5' region of the gene, presumably due to these sequences having been deleted. Frameshifting mutations are present in both the NS and VP pseudogenes of AmPVe-EllLut.1. The element is integrated into a locus that is homologous to mouse chromosome 12.

The AmPVe-EllLut.2 element comprises the NS gene alone (**Figure 1a**). This element is integrated into a locus immediately adjacent to the sequences encoding the MAF BZIP transcription factor G (MAFG) gene, which in the mouse genome is located in the 11qE2 region of chromosome 11. The NS gene encoded by AmPVe-EllLut.2 lacks a methionine start codon, but intriguingly, encodes a complete NS sequence that is otherwise intact.

The Transcaucasian mole vole (*E. lutescens*) is a species of cricetid rodent inhabiting semi-arid or grassland areas in Central Asia, and is notable for its unusual karyotype - all individuals possessing a diploid number of 17 chromosomes. Only a single sex chromosome [15] is present - with the Y chromosome having been eliminated. This interesting characteristic has motivated the sequencing of the *E.lutescens* genome, as well as that of a sister species - the northern mole vole (*E. talpinus*) [16]. We identified empty integration sites in the *E. talpinus* genome at the loci where the AmdoPVe-EllLut-1 and AmdoPVe-EllLut-2 elements are integrated in *E.lutescens*. This indicates that both elements were integrated into the *E.lutescens* germline after these two species diverged ~10 million years ago (MYA) [17, 18]. It seems unlikely, however, that either element has been integrated in very recent times. Firstly, the degraded nature of AmPVe-EllLut.1 indicates that it has been resident in the germline for some time. Furthemore, two *E.lutescens* individuals have been sequenced (a male and a female). We identified both insertions in both individuals, suggesting they are likely to be fixed, or to occur at high frequency in the species gene pool.

The identification of empty integration in *E. talpinus* facilitated the identification of EVE-genome junctions for both the AmdoPVe elements. In the case of AmdoPVe-EllLut-2, genomic flanks occur closely flank the NS gene, suggesting that it was likely derived from an mRNA that was reverse transcribed and integrated into the nuclear genome of an ancestral germline cell. By contrast, AmdoPVe-EllLut-1 appears to be derived from genome-length nucleic acid. In addition to containing regions of both NS and VP, this element includes a region of 3' sequence between the end of VP and the beginning of the genomic flanking sequence that exhibits similarity to the 3' untranslated region of AMDV, and contains inverted repeats capable of folding into a stem loop structure (data not shown). We could not, however, identify sequences corresponding to the 5' UTR in the AmdoPVe-EllLut-1 element. No methionine start codon could be identified in at the start of the NS pseudogene encoded by this element, but based on comparisons to AMDV, it appeared to be close to, or contiguous with the 5' genomic flanking sequence.

Discussion

So far, all Amdoparvoviruses have only been identified infecting carnivores (order Carnivora). In this study, we identified two EVEs derived from amdoparvoviruses in the genome of the Transcaucasian mole vole, extending the amdoparvovirus host range to a second mammalian order (Rodentia). Recently, a partial amdoparvovirus capsid sequence was identified via metagenomic screening of samples derived from the least horseshoe bat (*Rhinolophus pusillus*), suggesting that bats (order Chiroptera) may also harbour amdoparvoviruses. However, as with all virus sequences recovered via metagenomic sequencing, there remains a degree of uncertainty regarding host associations [19], whereas the presence of an EVE in the germline of a given species unambiguously demonstrates an association between that species and the virus group from which the EVE was derived.

Unlike the short amodparvovirus-related EVEs described previously [13], the elements identified here spanned near-complete amdoparvovirus genes or genomes. Moreover, in ML phylogenies, both elements grouped robustly within the amdoparvovirus clade as defined by exogenous isolates. Our data indicate that both elements – which share ~80% identity at the amino acid level - were integrated into the *E. lutescens* germline within the last 10 million years (**Figure S2**). One element (AmdoPVe-EllLut-1) was relatively degraded, indicative of a long period of residence within the host germline. However, the AmdoPVe-EllLut-2 element comprises a single NS gene that is not interrupted by stop codons and frameshifts. Amongst the various parvovirus-derived EVEs that have been reported so far, several have exhibited such

intact or nearly intact replicase genes. Furthermore, in a previous study, we showed tissue-specific transcription of an intact, replicase-encoding EVE in the genome of the degu (*Octodon degus*) [20]. The identification of yet another, relatively intact and independently acquired parvovirus replicase in a rodent genome deepens the mystery surrounding these elements. Given that fixation of EVEs is extremely unlikely, the independent acquisition of intact parvovirus replicase genes by several mammalian species suggests that these elements are being functionalised and selected for in some as yet undetermined way.

The phylogenetic relationships of the endogenous elements described here relative to exogenous carnivore amdoparvoviruses are quite revealing. Remarkably, we found that elements from the *E. lutescens* genome robustly separate carnivore isolates from skunks from those found in raccoon dogs, mink and foxes. Prior to the identification of amdovirus EVEs, all isolates in the *Amdoparvovirus* genus might have been considered to represent a complex of closely related virus species circulating in carnivores. The observation that an ancient EVE is more closely related to AMDV than AMDV is to amdoparvoviruses infecting raccoon dogs suggests that multiple, introductions of amdoparvoviruses into carnivore species have likely occurred, potentially from rodent hosts.

Methods

Genome screening and sequence analysis

Vertebrate genome assemblies were obtained from NCBI genomes (**Table S1**). Screening was performed using the database-integrated genome-screening tool (available from http://giffordlabcvr.github.io/DIGS-tool/). ORFs were inferred by manual comparison of putative peptide sequences to those of closely related exogenous parvoviruses. The putative peptide sequences of Amdoparvovirus-related EVEs were aligned with NS and VP sequences of representative Amdoparvoviruses and Protoparvoviruses using MUSCLE [21] and PAL2NAL [22]. Phylogenies were reconstructed from this alignment, using maximum likelihood as implemented in RaxML [23], and the RtEV protein substitution model [24] as selected using ProTest [25].

Table 1. Amdoparvovirus hits from *in silico* genome screening.

Element	Organism	Genes	Scaffold*	Start	End	O*	Length**
AmdoPVe-EllLut.1	*Mole vole*	NS-M-VP	LOEQ01006026.1	6398	10207	+ve	839
AmdoPVe-EllLut.2	*Mole vole*	NS	LOEQ01001077.1	110870	112639	+ve	589
PVe-ProCap.1	*Cape hyrax*	M-VP	ABRQ02031156.1	7932	8225	+ve	98
PVe-ProMuc.1	*Pit viper*	VP	BCNE02035092.1	59618	60139	-ve	184
PVe-SarHar.1	*Tasmanian devil*	VP	AFEY01431940.1	6434	6646	-ve	71
PVe-OryAfe.1	*Aardvark*	VP	ALYB01102612.1	8482	8882	-ve	100

Footnote: * Genbank accession numbers are shown for genomic scaffolds *O=orientation in scaffold. ** length is shown in codons. Abbreviations: NS=replicase; M=middle ORF; VP=capsid protein.

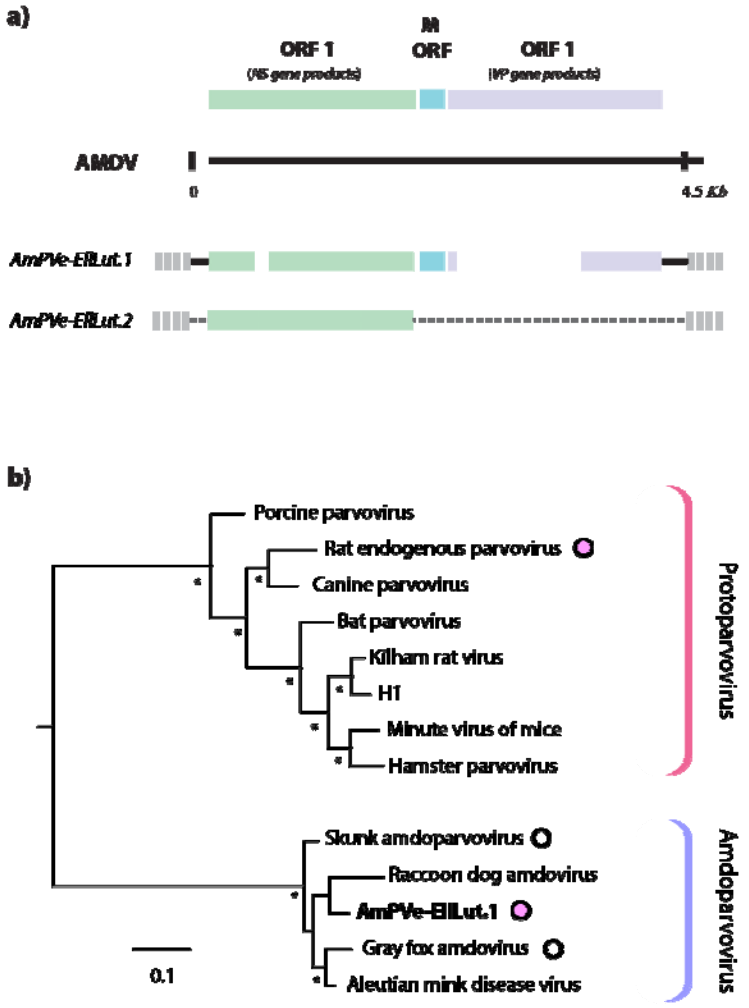

Figure 1. (a) Genome structure of AmPVe-EllLut elements. Asterisks (*) indicate stop codons. Green and orange arrowheads on non-structural (NS) protein indicate the position of conserved amino acid motifs of parvoviruses. AmdoPVe-EllLut = *Ellobius lutescens* endogenous amdoparvovirus. **(b) Phylogenetic relationships of amdoparvovirus and protoparvovirus sequences.** Maximum likelihood phylogeny showing the estimated evolutionary relationships between amdoparvoviruses and protoparvoviruses. Brackets to the right indicate the relevant genera. The scale bar shows evolutionary distance in substitutions per amino acid site. Pink-filled circles indicate endogenous taxa. White-filled circles indicate taxa identified via sequencing. Asterisks indicate nodes with maximum likelihood bootstrap support >80%. Genbank acession numbers are listed in **Table S2**.

Acknowledgements

RJG was funded by the Medical Research Council of the United Kingdom (MC_UU_12014/12). We thank Andrew Davison and Joseph Hughes for their comments and feedback on the manuscript.

References

1. Cotmore, S.F., et al., *The family Parvoviridae.* Arch Virol, 2014. **159**(5): p. 1239-47.
2. Canuti, M., H.G. Whitney, and A.S. Lang, *Amdoparvoviruses in small mammals: expanding our understanding of parvovirus diversity, distribution, and pathology.* Front Microbiol, 2015. **6**: p. 1119.
3. Canuti, M., et al., *Driving forces behind the evolution of the Aleutian mink disease parvovirus in the context of intensive farming.* Virus Evol, 2016. **2**(1): p. vew004.
4. Li, L., et al., *Novel amdovirus in gray foxes.* Emerg Infect Dis, 2011. **17**(10): p. 1876-8.
5. Pennick, K.E., et al., *Aleutian disease in two domestic striped skunks (Mephitis mephitis).* Vet Pathol, 2007. **44**(5): p. 687-90.
6. Shao, X.Q., et al., *Novel amdoparvovirus infecting farmed raccoon dogs and arctic foxes.* Emerg Infect Dis, 2014. **20**(12): p. 2085-8.
7. LaDouceur, E.E., et al., *Aleutian Disease: An Emerging Disease in Free-Ranging Striped Skunks (Mephitis mephitis) From California.* Vet Pathol, 2015. **52**(6): p. 1250-3.
8. Canuti, M., et al., *Full genetic characterization and epidemiology of a novel amdoparvovirus in striped skunk (Mephitis mephitis).* Emerg Microbes Infect, 2017. **6**(5): p. e30.
9. Feschotte, C. and C. Gilbert, *Endogenous viruses: insights into viral evolution and impact on host biology.* Nat Rev Genet, 2012. **13**(4): p. 283-96.
10. Holmes, E.C., *The evolution of endogenous viral elements.* Cell Host Microbe, 2011. **10**(4): p. 368-77.
11. Belyi, V.A., A.J. Levine, and A.M. Skalka, *Sequences from ancestral single-stranded DNA viruses in vertebrate genomes: the parvoviridae and circoviridae are more than 40 to 50 million years old.* J Virol, 2010. **84**(23): p. 12458-62.
12. Kapoor, A., P. Simmonds, and W.I. Lipkin, *Discovery and characterization of mammalian endogenous parvoviruses.* J Virol, 2010. **84**(24): p. 12628-35.
13. Katzourakis, A. and R.J. Gifford, *Endogenous viral elements in animal genomes.* PLoS Genet, 2010. **6**(11): p. e1001191.
14. Liu, H., et al., *Widespread endogenization of densoviruses and parvoviruses in animal and human genomes.* J Virol, 2011. **85**(19): p. 9863-76.
15. Matthey, R., *La formule chromosomique et le problème de la détermination sexuelle chez Ellobius lutescens,.* Arch Klaus-Stift Vererb-Forsch, 1953. **28**: p. 65–73.
16. Mulugeta, E., et al., *Genomes of Ellobius species provide insight into the evolutionary dynamics of mammalian sex chromosomes.* Genome Res, 2016. **26**(9): p. 1202-10.
17. Fabre, P.H., et al., *A glimpse on the pattern of rodent diversification: a phylogenetic approach.* BMC Evol Biol, 2012. **12**: p. 88.
18. Pisano, J., et al., *Out of Himalaya: the impact of past Asian environmental changes on the evolutionary and biogeographical history of Dipodoidea (Rodentia).* Journal of Biogeography, 2015. **42**(5): p. 856-870.
19. Bexfield, N. and P. Kellam, *Metagenomics and the molecular identification of novel viruses.* Vet J, 2011. **190**(2): p. 191-8.
20. Arriagada, G. and R.J. Gifford, *Parvovirus-derived endogenous viral elements in two South American rodent genomes.* J Virol, 2014. **88**(20): p. 12158-62.

21. Edgar, R.C., *MUSCLE: a multiple sequence alignment method with reduced time and space complexity.* BMC Bioinformatics, 2004. **5**: p. 113.
22. Suyama, M., D. Torrents, and P. Bork, *PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments.* Nucleic Acids Res, 2006. **34**(Web Server issue): p. W609-12.
23. Stamatakis, A., *RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies.* Bioinformatics, 2014. **30**(9): p. 1312-3.
24. 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.
25. Darriba, D., et al., *ProtTest 3: fast selection of best-fit models of protein evolution.* Bioinformatics, 2011. **27**(8): p. 1164-5.