Title

Non-canonical nematode endogenous retroviruses resulting from RNA virus glycoprotein gene capture by a metavirus.

Journal

Journal of General Virology, 103(5)

ISSN

0022-1317

Authors

Sacco, Melanie Ann

Lau, Jonathan

Godinez-Vidal, Damaris

et al.

Publication Date

2022-05-12

DOI

10.1099/jgv.0.001739

Peer reviewed

JOURNAL OF GENERAL VIROLOGY

RESEARCH ARTICLE

Sacco *et al.*, *Journal of General Virology* 2022;103:001739
DOI 10.1099/jgv.0.001739

MICROBIOLOGY SOCIETY

Non-canonical nematode endogenous retroviruses resulting from RNA virus glycoprotein gene capture by a metavirus

Melanie Ann Sacco[1,*], Jonathan Lau[1], Damaris Godinez-Vidal[2] and Isgouhi Kaloshian[2]

Abstract

Reverse-transcribing retroviruses exist as horizontally transmitted infectious agents or vertically transmitted endogenous retroviruses (ERVs) resident in eukaryotic genomes, and they are phylogenetically related to the long terminal repeat (LTR) class of retrotransposons. ERVs and retrotransposons are often distinguished only by the presence or absence of a gene encoding the envelope glycoprotein (*env*). Endogenous elements of the virus family *Metaviridae* include the insect-restricted *Errantivirus* genus of ERVs, for which some members possess *env*, and the pan-eukaryotic *Metavirus* genus that lacks an envelope glycoprotein gene. Here we report a novel Nematoda endogenous retrovirus (NERV) clade with core retroviral genes arranged uniquely as a continuous *gag-env-pro-pol* ORF. Reverse transcriptase sequences were phylogenetically related to metaviruses, but envelope glycoprotein sequences resembled those of the *Nyamiviridae* and *Chrysoviridae* RNA virus families, suggesting *env* gene capture during host cell infection by an RNA virus. NERVs were monophyletic, restricted to the nematode subclass Chromadoria, and included additional ORFs for a small hypothetical protein or a large Upf1-like RNA-dependent AAA-ATPase/helicase indicative of viral transduction of a host gene. Provirus LTR identity, low copy number, ORF integrity and segregation of three loci in *Meloidogyne incognita,* taken together with detection of NERV transcriptional activity, support potential infectivity of NERVs, along with their recent emergence and integration. Altogether, NERVs constitute a new and distinct *Metaviridae* lineage demonstrating retroviral evolution through sequential heterologous gene capture events.

INTRODUCTION

The evolution of reverse transcriptase (RT), the enzyme that copies RNA into DNA, is an innovation that has occurred in nature only once [1]. All reverse-transcribing viruses therefore share evolutionary origins with the endogenous elements called retrotransposons that comprise a significant portion of all eukaryotic genomes [2]. Three families of retrotransposons (*Metaviridae*, *Belpaoviridae* and *Pseudoviridae*) have genomes that resemble reverse-transcribing RNA viruses of the family *Retroviridae*, with long terminal repeats (LTRs) that flank *gag*, *pro* and *pol* genes arranged in tandem [3]. Proteins encoded by *gag* are processed from a Gag polyprotein into the matrix (MA), capsid (CA) and nucleocapsid (NC) subunits by the *pro*-encoded viral protease (PR) [4, 5]. Depending on the viral genome organization, the Gag polyprotein may be translated alone or with PR in frame [6]; alternatively, PR may be translated as part of the *pol* ORF or in its own frame [7, 8]. The Pol polyprotein is also processed from a polyprotein precursor into the viral enzymes RT, RNaseH and integrase (IN) [3]. These genes are translated from the genome-length RNA as polyprotein precursors of different lengths. The *pol* coding sequence is typically translated as a longer Gag-PR-Pol polyprotein by read-through of Gag on the genome-length transcript using ribosomal frameshifting in some retroviruses or termination suppression in others [9], although the spumaviruses express *pol* as a separate spliced subgenomic transcript [10]. Retroviruses additionally possess an *env* gene encoding the envelope glycoproteins (Env) that are expressed from a spliced subgenomic transcript, a feature that is only shared with some *Metaviridae* members of the genus *Errantivirus*, for which the

Received 21 October 2021; Accepted 10 February 2022; Published 12 May 2022
Author affiliations: [1]Center for Applied Biotechnology Studies, Department of Biological Science, California State University, Fullerton, CA 92834-6850, USA; [2]Institute for Integrative Genome Biology, Department of Nematology, University of California, Riverside, CA, 92521, USA.
***Correspondence:** Melanie Ann Sacco, msacco@fullerton.edu
Keywords: endogenous retrovirus; nematode; metavirus; Upf1; env capture.
Abbreviations: BLAST, basic local alignment search tool; BLAST, basic local alignment search tool; ERV, endogenous retrovirus; EVE, endogenous viral element; HERV, human endogenous retrovirus; LTR, long terminal repeat; NC, nucleocapsid; NERV, Nematoda endogenous retrovirus; NMD, nonsense-mediated decay; ORF, open reading frame; PPN, plant-parasitic nematode; PR, protease; RT, reverse transcriptase; VSV, vesicular stomatitis virus; WGS, whole genome shotgun.
Four supplementary tables and nine supplementary figures are available with the online version of this article.

Drosophila melanogaster gypsy virus (DmeGypV) is the prototype [3]; however, the DmeGypV *env* gene has distinct origins from those of retroviruses and has been traced back to a dsDNA baculovirus [11]. The *gag*, *pro*, *pol* and *env* genes form the viral core blueprints that are essential for retrovirus infectivity; complex retroviruses possess additional accessory protein genes absent from simple retroviruses that may function in post-transcriptional regulation of viral RNA processing and export from the nucleus, or modulation of innate defences of host cells [12]. ERVs typically have only the core genes and frequently have large deletions of entire genes; however, expression of two accessory proteins, Rec and Np9, has been reported for the human ERV-K (HERV-K) from alternatively spliced *env* gene transcripts [13, 14].

The use of a replication strategy by retroviruses that includes integration of the DNA copy (provirus) into their host's genomes [15] presents the opportunity for new ERVs to form when germline cells are infected, allowing for vertical transmission to offspring through the gametes and for copy numbers to expand over time for ERVs to become repetitive elements [16, 17]. ERVs that have been endogenized recently are absent from individuals and segregating within populations, such as is seen in the case of the koala retrovirus, which is actively becoming endogenized [18]. Newly established ERVs possess intact coding sequences and LTRs that are perfectly identical, as the reverse transcription mechanism involves duplication of the retroviral genome's unique 5′ (U5) and unique 3′ (U3) sequences to produce LTRs in the DNA provirus with the structure of U3-R-U5 [19]. This duplication is significant to prevent loss of genetic information during RNA genome transcription and for positioning U3 as the promoter for RNA polymerase II for transcription of both viral genomic and subgenomic RNAs [20]. The duplication of U3 and U5 at the time of replication can be useful for estimating the age of recently established ERVs as LTRs diverge from each other over time [21].

Ancient endogenization events allowed diverse retrovirus relics to become established in the genomes of the ancestors to modern species; older elements existing as orthologues in diverged species provide a kind of 'fossil' record of infections by now extinct retroviruses spanning millions of years [22]. The ERV and retrotransposon-like elements constitute a significant amount of the eukaryotic genome, contributing to about 8% of the genome in the case of humans [23]. The HERVs are predominantly ancient proviruses with defective ORFs that have accumulated mutations and deletions rendering them inactive, although younger elements may retain some coding potential but are probably not infectious [24]. In some cases, ancient retroviral elements have been captured by their hosts to evolve new functions [25]: retroviral LTRs may influence gene expression regulation of nearby genes [26], while viral proteins have also been adapted to other purposes, as seen in the case of the syncytins that evolved from an envelope glycoprotein to function in placental development [27], or the tetrapod retroviral capsid-like ARC protein that functions in neuronal intercellular signalling to regulate synaptic function [28]. Conversely, ERVs tend to become degraded and lose genes over time, which confounds attempts to determine which reverse-transcribing elements in the genome may have originated first. Defective ERVs exist that resemble retrotransposons with the LTR-*gag-pro-pol*-LTR structure due to the loss of *env,* making it difficult in some cases to discern if that element is a remnant of a past retroviral infection or represents a bona fide retrotransposon that has never existed as an infectious virus [21].

The ERV compositions of model vertebrate species have been studied in depth with the availability of many complete, high-quality genome sequences, while studies in invertebrates have been more limited. The free-living nematode *Caenorhabditis elegans* is an important model for development and was the first multicellular eukaryote to have its genome sequenced and released [29]. Twelve lineages of viral elements were identified in *C. elegans,* designated as Cer1–Cer12 [30], of which only Cer1 and Cer7 possessed envelope sequences downstream of *pol* indicative of an ERV [30, 31]. The retrotransposon-like elements Cer1–Cer6 are classified within the family *Metaviridae* [30, 32], while Cer7 belongs to the Tas clade of the genus *Belpaoviridae* [33]. The recent explosion in eukaryotic genome sequencing initiatives makes it possible to extend studies of retroviral elements into non-model species of significance that are not tractable in the laboratory, including invertebrate organisms such as nematode species that are parasitic to plants, insects and vertebrates, for which endogenous viruses have been unexplored.

Here we describe recently integrated endogenous retroviruses in the phylum Nematoda (NERVs) with a non-canonical genome arrangement that places the gene for the envelope glycoproteins between the *gag* gene for the structural polyprotein and *pro-pol* gene encoding the viral enzymes for proteolytic processing, reverse transcription and integration. Strikingly, in most complete NERV proviruses, the four core virus genes are in a single, continuous ORF (ORF1), while additional ORFs appear at the 3′ end, which in some proviruses showed homology to the Upf1 AAA-ATPase and helicase. NERVs were identified in nematodes of the subclass Chromadoria that are either free-living or parasites of plants, animals or insects. Phylogenetic analysis with RT sequences placed these retroelements in a clade with the *C. elegans* Cer2 in the family *Metaviridae*, while the Env sequence most closely matched the glycoprotein from the families *Nyamiviridae* and *Chrysoviridae*. NERVs have a low copy number and showed perfect or high LTR sequence identity indicative of recent integration, which, together with evidence that they are transcriptionally active in the plant-parasitic nematodes (PPNs), suggests that these retroelements evolved recently and could be actively expanding in copy number.

METHODS

Detection and analysis of novel ERV elements in nematode whole-genome shotgun (WGS) databases

NERV sequences were identified using the Basic Local Alignment Search Tool (BLAST) suite from the NCBI to search WGS contigs, initially using amino acid sequences of representative alpharetrovirus and class III avian ERV RT sequences (Table S1, available in the online version of this article) using Translated BLAST (tblastn) using BLOSUM62 matrix with default parameters for word size and gap costs, and low-complexity regions filtered [34]. Genomes were subject to repeated rounds of search using standard nucleotide BLAST (blastn) with NERV DNA sequences, and deduced amino acid sequences (tblastn) determined using the Expasy Translate tool, and aligned using Clustal Omega version 1.2.4 [35], and phylogenetic trees were generated using Distance Methods in SeaView version 4.6.4, with the BioNJ algorithm and Poisson distances, with 100 bootstrap replicates and gaps ignored [36]. Loci were only considered as possessing bona fide NERV sequences for further study by detection of the acquired *env* gene positioned atypically between *gag* and *pro*. Accession numbers of all published viral sequences used in our analyses and their abbreviations are provided in Table S2. To identify polypeptide regions by homology, NERV amino acid sequences were submitted to the PHYRE^2 (Protein Homology/analogy Recognition Engine V 2.0) web portal using the normal modelling mode [37]. Suborders of nematodes in the subclass Chromadoria were identified as described by Smythe *et al.* [38] in a cladogram generated using an 18S rRNA segment from one nematode species representing each genus using Distance Methods in SeaView version 4.6.4.

DNA isolation from *M. incognita*

Meloidogyne incognita race 3 was maintained on tomato plants (*Solanum lycopersicum*) cv. Red Cherry in a greenhouse. Nematode eggs were extracted by shaking infected roots in a 10% bleach solution and sieving [39]. The eggs were then floated on a 35% sucrose cushion, rinsed with sterile water and hatched using a modified Baermann method in sterile distilled water at 25 °C in the dark. Eggs and infective second stage juveniles (J2) were used for genomic DNA extraction by CTAB according to Doyle and Doyle [40] with some modifications. J2 were incubated in 3% CTAB supplemented with 100 µg proteinase K ml^{-1} for 3 min at 100 °C followed by chloroform extraction and ethanol precipitation.

PCR analysis of *M. incognita* NERV loci

PCR was performed in 50 µl reaction mixtures on genomic DNA extracted from *M. incognita* eggs and J2 using Herculase II Fusion DNA polymerase (Agilent) according to the manufacturer's protocol. Oligonucleotides were synthesized by Integrated DNA Technologies (IDT). For MincNERV$_1$ analysis, primer pair M1IntFor (5′-AGCGTGACGAGATTCAACG-3′) and M1U3PBSRev (5′-ACCAATATTAGGTCCCGGTTC-3′) was used to detect the integrated provirus at the 5′ end. The presence of the preintegration state locus was amplified with primer pair M1IntFor plus M1IntRev (5′-TGGCCCAGACCTAATTCATC-3′). Similarly, the MincNERV$_2$ preintegration locus was assessed with primer pair M2IntFor (5′-TGGGATAATCGACT ATTTGTATAAC-3′) and M2IntRev (5′-ATAGGATATTGGTG GATTTGGAC-3′) and the 5′-integration site with M2IntFor plus M2UTRRev (5′-GAGT TCTCAAAAAGCACACAG-3′). Reactions were mixed with 250 ng DNA, 0.25 µM primers and 0.2 mM dNTPs in 1× Herculase reaction buffer and thermocycling was performed with one cycle of 95 °C for 2 min and 30 cycles of 95 °C for 30 s, 53 °C for 30 s and 72 °C for 30 s, followed by one cycle of 72 °C for 3 min in an MJ Mini thermocycler (Bio-Rad). PCRs were separated on 1.2% agarose gels in TBE stained with ethidium bromide and photographed using a Gel Imager Plus (VWR). Amplicons of the expected size ranges were excised, and DNA was extracted using the Qiaquick gel extraction kit (Qiagen) for sequencing by Eton Bioscience using the same primers as for the PCR amplification.

RESULTS

Unexpected genome organization of a novel NERV in PPN species

Because of our research interests in PPNs and ERVs, we initiated a study to leverage the recent National Center for Biotechnology Information (NCBI) whole-genome shotgun (WGS) DNA sequencing databases published for nematodes to identify intact ERV elements. A TBLASTN search of the NCBI databases for cyst nematodes of the genus *Globodera* using representative RT sequences from the three ERV classes and *Retroviridae* genera identified endogenous elements in the datasets for *Globodera pallida* strain Lindley and *Globodera rostochiensis* strain Ro1 (Table 1). *Globodera* proviruses were intact but lacked the typical 4–6 bp direct repeats of host sequences flanking the LTRs that are the hallmark of retrovirus integration [41], and the LTRs showed some divergence, which could result from incorrect assembly of reads in a contig. Therefore, these proviruses were in turn used to retrieve related ERVs from the root-knot nematode *M. incognita*. Three intact proviruses with identifiable flanking direct repeats were identified from *M. incognita* with perfect or almost perfect LTR identity (Table 1), and two proviruses (MincNERV$_1$ and MincNERV$_2$) shared 99% DNA sequence identity to each other (Table S1).

In silico translation of these NERVs revealed two ORFs (Fig. 1). The short polypeptide (ORF2a) had no sequence homology identified using BLASTP searches or using secondary structure prediction with PHYRE2 but was predicted to be nuclear-localized. The deduced translation of ORF1 yielded an unusually long retroviral polyprotein that was analysed for the presence

Table 1. NERV sequences identified from WGS databases of nematodes from the suborder Tylenchina

Species	Provirus	NCBI sequence ID	Coordinates	Length (bp)	LTRs*	ORF1 (aa)	ORF2 (aa)
Globodera pallida	GpalNERV$_1$	CBXT010010034.1	3656–12331	8676	96%	2240	245
	GpalNERV$_2$	CBXT010007201.1	8941–14985	8941	98%	2240	238
Globodera ellingtonae	GellNERV$_1$	MEIZ01000033.1	37156–45004	7849†	ND	2157	258
Globodera rostochiensis	GrosNERV$_1$	FKKZ01000366.1	1–9146	9146†	ND	2170	N/A†
Heterodera glycines	HglyNERV$_1$	PTLT01000575.1	121472–130140	8669	99%‡	2173	262
	HglyNERV$_2$	PTLT01000045.1	74488–83245	8758	98%‡	2262	248
Hoplolaimus columbus	HcolNERV$_1$	JAEHOM010000475.1	1476–10032	8558	94%	2140	196
Hoplolaimus galeatus	HgalNERV$_1$	JAEHOL010000297.1	1–8195	8195†	ND	2183	251
Meloidogyne enterolobii	MentNERV$_1$	CAJEWN010000063.1	131360–140082	8713	99%‡	2162	238
Meloidogyne floridensis	MfloNERV$_1$	RCFN01000091.1	2476–11102	8627	100%‡	2111	250
Meloidogyne incognita	MincNERV$_1$	RCFL01010719.1	612–9237	8626	99.8%‡	2111	250
	MincNERV$_2$	BLLR01000009.1	2175568–2184191	8624	100%‡	2074	250
	MincNERV$_3$	BLLR01000042.1	911600–920001	8402	100%‡	2189	215
Meloidogyne luci	MlucNERV$_1$	CACSLI010000217.1	272300–281150	8712	100%‡	2,161	238
Bursaphelenchus xylophilus	BxylNERV$_1$	CAJFDI010000004.1	7903789–7917291	13 503	99%‡	2409	393
Ditylenchus destructor	DdesNERV$_1$	LSTP01000477.1	494–14028§	ND†	ND	3179‖	829
Rotylenchulus reniformis	RrenNERV$_1$	LDKF01000328.1	47750–57436	9687	100%	2388	393
Steinernema feltiae	SfelNERV$_1$	MQUG01001037.1	7913–20480	12 568	99%	2527	752

*LTR identity determined only for proviruses with complete 5′ and 3′ LTRs.
†Incomplete provirus in contig, preventing complete identification of both LTRs.
‡LTRs with 3–5 bp flanking host DNA direct repeat identified.
§Coordinates beginning with ORF1 start codon and ending at ORF2 stop codon.
‖Includes stop codon located between *gag* and *pro* genes in the same reading frame.
¶ORF1 approximate total length for polypeptide coding sequence with one or more frameshifts.

of conserved domains and motifs that are signatures of retroviral protein subunits. PHYRE2 modelling of the amino terminus defined an approximately 150 aa residue region that aligned with the Ty3 retrotransposon capsid protein encoded by the *gag* gene, including the conserved zinc knuckle motif ($CX_2CX_4HX_4C$) typical of NC proteins (Fig. S1). Additionally, the highly conserved aspartic proteinase signature motif (DTG) of the viral PR and RT YΨDD motif (where Ψ represents the aliphatic amino acids M, V, L or I) were identified in the carboxy-terminal third of the ORF1 polyprotein sequence (Fig. S1). Positioning the conserved elements of the NC and PR subunits, which are normally in close proximity, revealed a large insertion between these subunits in the deduced polyprotein sequence with no BLAST similarity to known proteins of *Retroviridae*. However, the PHYRE2-predicted secondary structure for the MincNERV$_3$ polypeptide insertion modelled onto a region of 431 aa residues of the glycoprotein B from Herpes simplex virus type I template with 99.5% confidence. This putative NERV glycoprotein sequence followed a potential start methionine and amino-terminal signal peptide sequence for an envelope glycoprotein polypeptide (Fig. S1). Hydrophobicity analysis defined two stretches of hydrophobic residues separated by a single lysine residue on the carboxy-terminal side of the candidate Env sequence in the expected location for the transmembrane-spanning region. Overall, the insertion region encoded a typical retroviral Env structure with a surface protein (SU) ectodomain and transmembrane (TM) peptide that is separated by the furin cleavage site (Fig. S1). The NERV gene arrangement contrasts with typical retrovirus and ERV genomes that generally have the order *gag-pro-pol* with stop codons following *gag* and/or *pro* as well as *pol*, followed the *env* gene, which is expressed as a spliced subgenomic transcript (Fig. 1).

NERV provirus loci in diverse nematode genera of the subclass Chromadoria

The discovery of an ERV lineage with a novel gene arrangement prompted the question of whether NERVs were solely a curiosity of closely related nematodes that parasitize plants or whether these retroviruses may be more widespread, including species of animal-infecting and free-living nematodes, which might indicate that their unusual genome structure could support infection and

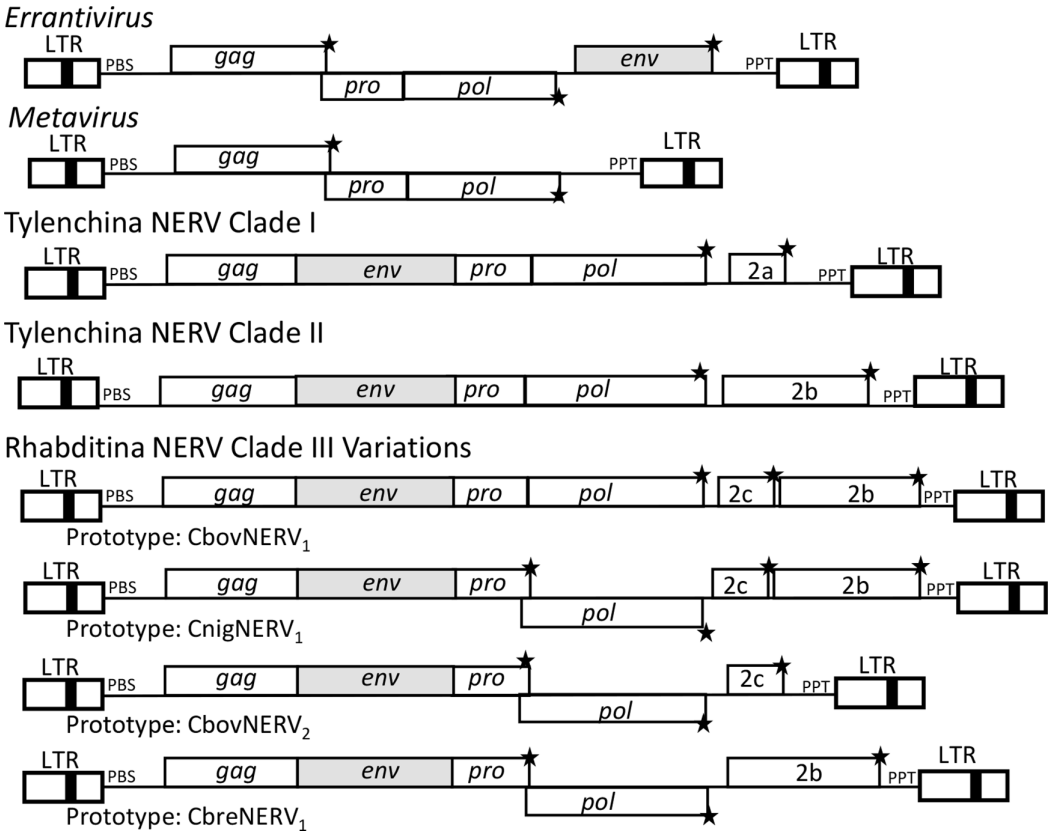

Fig. 1. Novel, non-canonical genome organization of NERVs. Schematic diagram comparing distinct NERV genomes to prototype members of the two Metaviridae genera: *Errantivirus*, which includes an *env* gene, and *Metavirus*, which includes retrotransposons without an *env* gene. NERV clade I is the predominant genome structure found in closely related plant-parasitic cyst and root-knot nematodes of the suborder Tylenchina, with ORF2a encoding a small protein of unknown function. Clade II proviruses carry ORF2b encoding a large AAA-ATPase/helicase related to the Upf1 protein or a truncated gene. Clade III proviruses exclusively found in the genus *Caenorhabditis* had a continuous *gag-env-pol* ORF punctuated by a stop codon (star) and the *pol* gene in a different reading frame, and carried ORF2b and/or ORF2c.

horizontal spread. NERV elements were identified by a TBLASTN search with the long ORF1 deduced translation product from proviruses MincNERV$_1$ and MincNERV$_3$ in most of the available WGS databases for PPN belonging to the nematode suborder Tylenchina (Fig. S2). However, in some cases, matching contigs in a dataset were lower quality and/or short in length, so provirus sequences were incomplete, or the LTRs diverged and lacked a flanking direct repeat of the host DNA generated by integration, which could indicate *in silico* contig assembly errors (Table 1). Many PPN NERV sequences retrieved in the TNBLASTN search had complete coding sequences, but some proviruses were incomplete due to the small contig sizes, particularly for the genus *Hoplolaimus*; however, ORF sequences could still be determined for incorporation into phylogenetic analyses (Table 1). Tylenchina includes the entomopathogenic genus *Steinernema* (Fig. S2), for which a WGS was available for *Steinernema feltiae* in which a single intact provirus was identified (Table 1). Additional sequences retrieved from other Tylenchina genera *Panagrolaimus*, *Panagrellus, Radopholus*, *Strongyloides*, *Acrobeloides*, *Deladenus*, and *Halicephalobus* were excluded from analysis as only provirus fragments were found or sequences were unclear due to low-quality contigs with base ambiguities. Nonetheless, detection of partial NERV proviruses in these genera showed widespread distribution of these ERV elements in the suborder Tylenchina as determined by detection of the presence of the novel *env* sequences positioned uniquely between *gag* and *pro*. Comparison of the Gag polyprotein sequences of the Tylenchina prototype sequences showed low identity and similarity between elements from more distantly related species, in contrast to pairwise comparisons of closely related species of the cyst and root-knot nematodes (Table S1).

The existence of NERVs in genera across all clades of Tylenchina suggested that these retroelements could be more pervasive in nematodes generally. Tylenchina is one of three suborders in the order Rhabditida in the subclass Chromadoria and contains all the PPN genera (Fig. S2). A TBLASTN search of 44 available WGS assemblies representing 29 different species in the order Spirurina yielded only a single full-length NERV from the herring worm *Anisakis simplex* (AsimNERV$_1$; Table 2). A similar search of 44 available WGS assemblies representing 33 different species in the suborder Rhabditina identified grossly intact NERV loci

Table 2. Prototype NERV sequences identified from WGS databases of nematode suborders Rhabditina and Spirurina

Species	Provirus	NCBI sequence ID	Coordinates	Length (bp)	LTRs*	Polyprotein (aa)	Accessory ORFs (aa)
Anisakis simplex	AsimNERV$_1$	UYRR01031084	91724–100428	ND†	ND†	2475	None
Caenorhabditis bovis	CbovNERV$_1$	CADEPM010000005.1	256789–268 751	11 963	99%‡	2431	246/831
	CbovNERV$_2$	CADEPM010000003.1	2677018–2 686 375	9358	99%‡	1187/1076§	176
	CbovNERV$_3$	CADEPM010000001.1	660896–670 228	9333	99%‡	2289	176
Caenorhabditis brenneri	CbreNERV$_1$	ABEG02001423.1	104003–116150	12 148	100%‡	1337/1052‖	1071
Caenorhabditis nigoni	CnigNERV$_1$	PDUG01000002.1	37156–45004	13 629	100%‡	1376/1061§	164/1135
	CnigNERV$_2$	PDUG01000021.1	124912–139892	15 037	99%	1376/1061§	164/809
Caenorhabditis remanei	CremNERV$_1$	WUAV01000060	31333–30 953	12 148	100%‡	1331/1024‖	1079
	CremNERV$_2$	LFJK02000008.1	7362835–7373142	ND†	ND†	1414/1043‖	163

ND, Not determined.
*LTR identity determined only for proviruses with complete 5′ and 3′ LTRs.
†Incomplete provirus in contig.
‡LTRs with 4 bp flanking host DNA direct repeat identified.
§Pol protein length beginning from potential methionine start codon following Pro stop codon.
‖Pol protein length beginning from region of potential translational frameshift (without methionine start codon present) based on alignment to CnigNERV$_1$ Pol amino terminus.

in four species of nematodes in only one genus, *Caenorhabditis* (Table 2). Most NERVs in this genus had the ORF1 sequences disrupted by a frameshift positioned between *pro* and *pol*, in a configuration similar to other retroviral genes that use ribosomal frameshifting for extension of the Gag-PR polyprotein into the polymerase-coding region. Overall, intact NERVs in all nematode species were present as a low number of copies (Fig. S3), many of which have LTRs with 100% DNA sequence identity, which is an indicator of recent integration.

RT phylogeny shows NERVs as a new member of the *Metaviridae* with distinct lineages

The unusual gene arrangement in the NERV proviruses suggests that these viruses emerged independently of other modern virus lineages with *gag-pro-pol-env* genome structures. This idea was explored further by retrieving RT sequences from representative members of diverse genera in the order Ortervirales and the three ERV classes (Table S2), as RT is the slowest evolving gene of the reverse-transcribing viruses, allowing comparison across the order [42]. Our initial analysis indicated that NERVs were most closely related to the family *Metaviridae* that includes DmeGypV, the yeast Ty3 and the *C. elegans* Cer1 retrotransposons. Therefore, RT sequences from additional members of this family were retrieved, including another nematode element from *C. elegans*, Cer2. NERVs formed a distinct subclade with Cer2 within the larger clade of *Metaviridae* sequences (Fig. 2).

An analysis of RT and Gag sequences was conducted to examine the relationships of all complete NERV proviruses identified in this study (Tables 1 and 2). We also included additional retrotransposon-like sequences (lacking *env*) retrieved from four nematode species with homology to Cer1 (Table S3) for phylogenetic analysis. The *env*-possessing NERVs clustered as three distinct clades separate from the *Metaviridae* sequences (Fig. S4A). Two of the candidate retrotransposons (MenCerL-RT and AceyCerL-RT) clustered with DmeGypV and Cer1 and may represent bona fide retrotransposons. Cer2 and the other candidate retrotransposons (GpaCerL-RT and HglyCerL-RT) were distributed in the three NERV clades and could represent remnants of past NERV integrations with the type of gene deletions typically seen in ERVs. RT and Gag trees showed similar topologies to each other and to the nematode 18S rRNA tree (Fig. S2) with the exception of the Gag protein from *Bursaphelenchus xylophilus* (BxlNERV$_1$), which was highly diverged from all NERVs and probably represents a recombinant provirus (Fig. S4b).

Phylogenetic and homology-based inference of NERV evolution by capture of an RNA virus *env* gene by a new member of the *Metaviridae*

While RT analysis suggested NERVs originated from a nematode retrotransposon (Fig. 2), the position and sequence of the NERV *env* gene suggested that it has a distinct origin from the *env* gene of the genus *Errantivirus*, for which DmeGypV is the

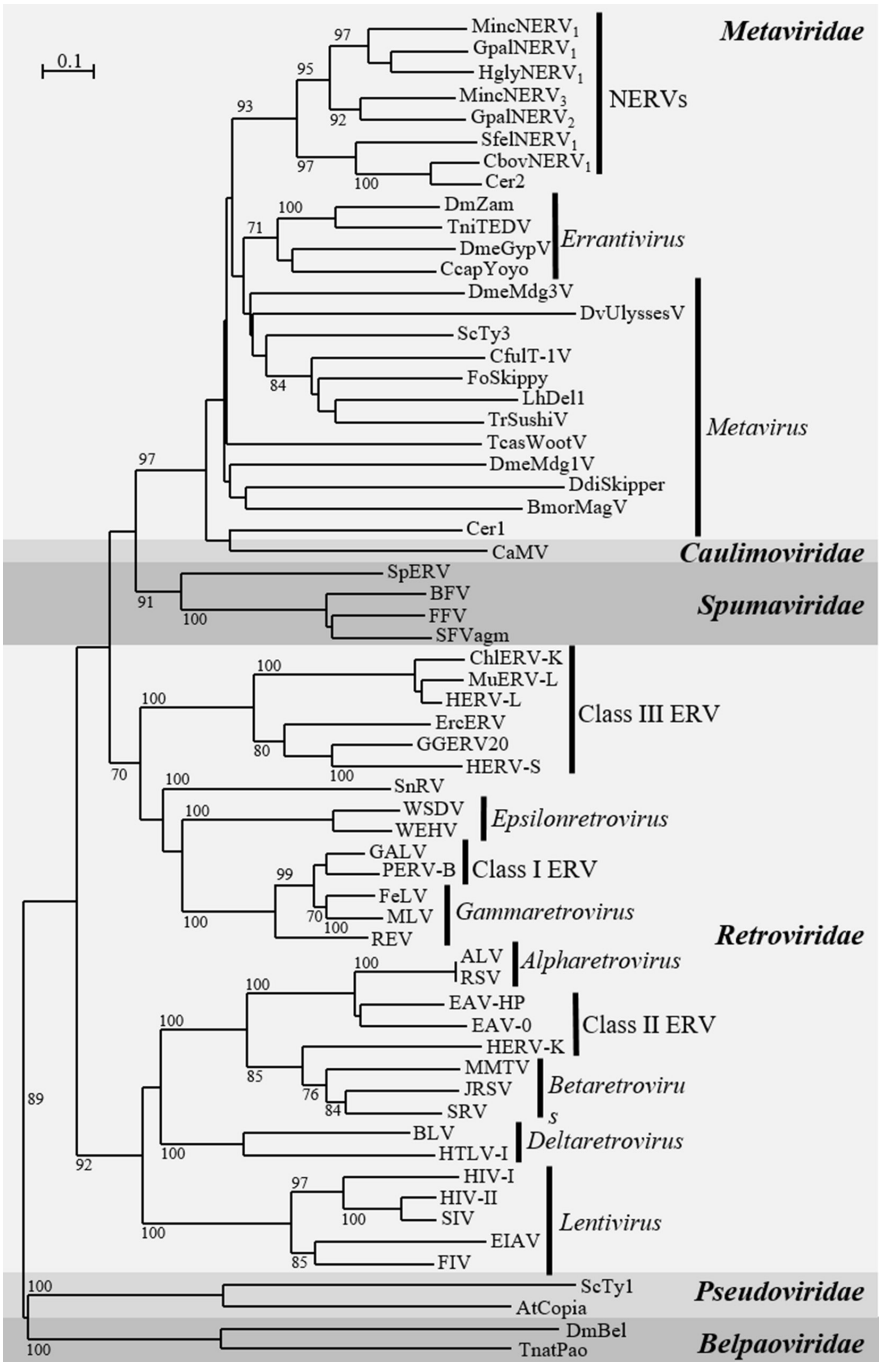

Fig. 2. Phylogenetic relationship of NERVs to members of the order Ortervirales. RT sequences were aligned for a conserved region of the RT corresponding to 232 aa of the MincNERV$_3$ prototype sequence (Fig. S1) from representative sequences across families in Ortervirales and classes of endogenous retroviruses. NERVs formed a distinct, monophyletic clade within a larger clade defining the family *Metaviridae* that includes the genera *Metavirus* and *Errantivirus*. The tree was generated using distance methods with BioNJ bootstrap values of 70 or more shown for 100 replicates. The scale bar represents 0.1 expected amino acid replacements per site. Accession numbers and full names for virus abbreviations are listed in Table S2.

prototype. DmeGypV is thought to have emerged through the acquisition of the fusion (F) glycoprotein gene from a baculovirus, a DNA virus that infects insects [11, 43]. An initial BLASTP search with putative Env sequences from GpalNERV$_1$ matched the glycoprotein (G) protein from Sierra Nevada virus, a member of the genus *Nyamiviridae* from the minus-sense RNA virus order *Mononegavirales*, with 62% query coverage, 22% identity and 37% similarity to the GpalNERV$_1$ amino-terminal region corresponding to the SU protein subunit of the glycoprotein precursor polypeptide and the amino-terminal portion of the TM

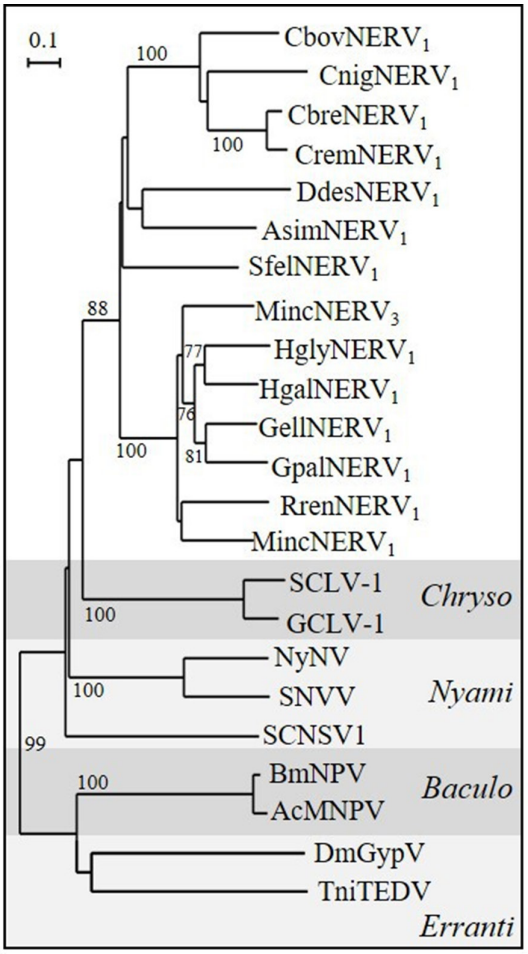

Fig. 3. Phylogenetic relationship of NERV envelope glycoproteins with RNA virus glycoproteins from *Chrysoviridae* and *Nyamiviridae*. Glycoprotein sequences were aligned in Clustal Omega for the first 490 aa of MincNERV$_3$ that includes the entire predicted SU, the conserved furin cleavage site, and the amino-terminal portion of TM before the putative transmembrane region (Fig. S1). Fusion (F) proteins from two baculoviruses, AcMNV and BmNPV, were included as glycoproteins that share a common ancestor with the errantiviruses, DmeGypV and TED virus from *Trichoplusia ni* (TniTEDV). Accession numbers and full names for virus abbreviations are listed in Table S2. High-confidence bootstrap values (threshold=70) for 100 replicates are shown and the scale bar represents 0.1 expected amino acid replacements per site.

protein released by furin cleavage. Further searches with MincNERV$_3$ retrieved matches to dsRNA viruses from the family *Chrysoviridae*, with 76% query coverage for a protein from Shuangao chryso-like virus 1 (SCLV-1) showing 23% identity and 38% similarity. Alignment of candidate envelope glycoprotein sequences representing NERVs and members of the families *Nyamiviridae* and *Chrysoviridae* (Fig. S5) showed a great deal of divergence, which was expected for proteins that are under pressure from host immunity to evolve; however, multiple sites showed strong conservation, including cysteine residues probably involved in disulphide bond linkages, and other residues that may affect tertiary structure, such as Gly and Pro. For phylogenetic analysis, envelope glycoprotein sequences from prototype NERVs and representative nyamiviruses and chrysoviruses were aligned with the errantivirus sequences and related F proteins from two baculoviruses (Table S2). The phylogenetic tree showed a distinct clade for the baculovirus-derived sequences that excluded the NERVs, which clustered with the nyamiviruses and chrysoviruses (Fig. 3). This analysis supports our hypothesis that a metavirus-like retrotransposon gave rise to the NERVs by acquiring an *env* gene that originated from an RNA virus in the kingdom Orthornavirae rather than from a DNA virus, as in the case of DmeGypV. Furthermore, the families *Chrysoviridae* and *Nyamiviridae* probably share a common ancestor for their envelope glycoproteins.

Evidence of host gene transduction by NERV

The genome structures of the NERVs showed multiple lineages with regard to the provirus 3′ ends (Fig. 1). Proviruses clustering in Clade I based on Gag and RT phylogeny (Fig. S4) all contain a short ORF (ORF2a) encoding a hypothetical protein of unknown function, but which nonetheless shows a degree of conservation among the PPN (Fig. S6a). Some members of Clade III also

possessed a short second ORF (ORF2c) that was distinct from ORF2a and had a high degree of divergence among proviruses, but highly conserved residues are apparent in the sequence alignment (Fig. S6B), suggesting ORF2c also encodes a small protein of unknown function. Clade II proviruses and some Clade III members possessed a much larger ORF (ORF2b) that matched the Upf1 AAA-ATPase and helicase. Alignment of the deduced ORF2b amino acid sequence with Upf1 and related AAA-ATPase/helicases allowed identification of highly conserved signature motifs of these enzymes (Fig. S7), such that the ORF2b protein could be a functional ATPase and helicase used in NERV infection of host cells.

Transcriptionally active NERVs in *Meloidogyne* and their alternative splicing potential

The location of the SU and TM coding sequences as part of a continuous ORF with Gag and Pol proteins is inconsistent with the translation of these proteins on the rough endoplasmic reticulum (ER), requiring recognition and processing of an amino-terminal signal peptide, furin cleavage during transit through the Golgi, and localization to the plasma membrane [44, 45]. Known retroviral *env* genes are translated from a spliced subgenomic transcript that allows translation of the Env polypeptide precursor separately from Gag, PR and Pol. To investigate NERV expression and potential splicing of *env*, PPN and *Caenorhabditis* NERVs were used to search SRA datasets to first identify transcriptionally active proviruses, as ERVs are prone to transcriptional gene silencing. For the PPN, reads were observed along the provirus lengths for GpalNERV$_1$ and GpalNERV$_2$ or MincNERV$_{1/2}$ and MincNERV$_3$ in *G. pallida* and *M. incognita*, respectively (Fig. S8A–C). For the Rhabditina, reads aligning with CnigNERV$_1$ were seen in *C. nigoni* (Fig. S8D). None of our searches of SRA data retrieved reads derived from a splice junction; however, the low number of reads relative to the provirus length may have precluded their detection. Notably, higher numbers of reads were seen in regions corresponding to the candidate accessory genes for MincNERV$_3$ and CnigNERV$_1$ (*orf2a* and *orf2b*, respectively), consistent with accumulation of subgenomic transcripts; however, a read bias may also be generated from expression of defective elements with large deletions of the genome in the genetic backgrounds used for generating the RNA-sequencing data, such as can be seen for elements matching MincNERV$_1$ and MincNERV$_2$ but not MincNERV$_3$ (Fig. S3). Further investigation of potential splicing may require cloning of a provirus for expression in a different species for a clean background for analysis by RT-PCR.

Integration of *Meloidogyne* NERVs pre-speciation and continued segregation of loci

A key method of determining the age of ERVs is to examine related animal species for the presence of a specific provirus integration that would indicate integration prior to speciation, or alternatively, to detect the presence of the preintegration state locus. We used PCR of genomic DNA to amplify MincNERV$_1$ and MincNERV$_2$ from *M. incognita* race 3 and only amplified sequences corresponding to the provirus locus (Fig. S9), which was confirmed by sequencing of the DNA extracted from agarose gels to match the published sequence (data not shown). We extended this line of investigation by searching the *Meloidogyne* WGS for preintegration state and proviral loci for the three intact NERVs from *M. incognita*. For *M. incognita*, only the MincNERV$_1$ proviral locus was found and, taken together with our PCR analysis, suggests this allele may have become fixed in *M. incognita*, although additional populations and races need to be examined for this conclusion to be more certain. The MincNERV$_1$ locus, but not the preintegration state, was also found in *M. arenaria, M. javanica* and *M. enterolobii* sequences (Fig. 4), indicating that it pre-dates separation of these species, although the MincNERV$_1$ proviruses in *M. javanica* and *M. enterolobii* have undergone subsequent distinct internal deletions, which probably occurred after separation of those species. MincNERV$_2$ and MincNERV$_3$ loci were both found in *M. incognita* and *M. enterolobii*, the two most distantly related root-knot nematode species, although the preintegration

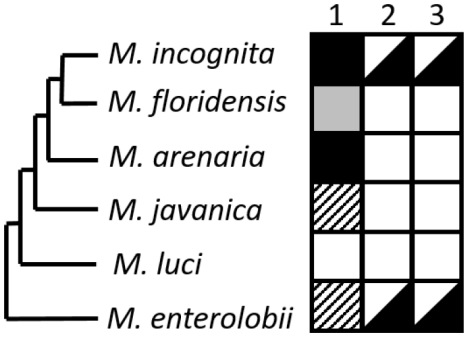

Fig. 4. Integration prior to separation of species and segregation of intact NERVs in *Meloidogyne*. NERVs from *M. incognita* and a reconstituted preintegration sequence were used to search the WGS databases for the six species of *Meloidogyne* available at NCBI. Relationships of species are based on Álvarez-Ortega *et al*., all of which belong to *Meloidogyne* clade I [71]. Shading indicates the presence of an intact provirus in the corresponding locus (black) or a disrupted provirus with an identical provirus insertion site and host sequence duplication (hatched) versus the preintegration state sequence (white), with boxes split for species databases in which both the preintegration and the proviral loci were present. Neither the complete preintegration site nor a provirus were conclusively retrieved for MincNERV$_1$ in *M. floridensis* (grey). Preintegration and proviral state sequences retrieved from WGS databases are provided in Table S4.

state loci were also present, indicating that these loci are still segregating. The presence of only the preintegration state locus in the other four *Meloidogyne* species could indicate that these loci have been lost from their lineages as they descended from a common ancestor or they are segregating in those species but were not present in the individuals sampled for genome analysis. Proviral loci that have become fixed in a population tend to be older than proviruses that are still segregating. The patterns of segregation for the three NERV elements in the genus *Meloidogyne* is consistent with the LTR identities that indicate MincNERV$_1$ is older because its LTRs are not perfectly identical as they are for MincNERV$_2$ and MincNERV$_3$ (Table 1).

DISCUSSION

The NERVs represent a unique monophyletic evolutionary branch of the Ortervirales tree that is as distinct from the genera *Metavirus* and *Errantivirus* as these two genera appear from each other in our phylogenetic analysis (Fig. 2). NERVs therefore probably constitute a new genus within the family *Metaviridae*, for which we propose the name *Nematavirus*. While it remains to be determined if these retroelements are fully competent for infection with their unusual gene arrangement, the existence of intact provirus loci across three suborders of the phylum Nematoda, detection of transcriptional activity and the presence of multiple copies for many proviruses, as well as high sequence identity of LTRs, suggest that NERVs have recently been active and may be infectious. In support of this hypothesis, expansion of the copy number of the HERVs following independent integration events and endogenization was deduced to have occurred predominantly through germline reinfection of replication-competent elements for the HERV-K.HML2 lineage [46], although other mechanisms such as retrotransposition have made other contributions in HERV copy expansion [47]. Follow-up in-depth study of NERVs will be important for elucidating all the mechanisms these endogenous elements employ to transcribe and translate their proteins, to assess their infectivity and to define their host range. Such an undertaking will require cloning proviruses for study in a different genetic background so observations are not potentially confounded by the presence of related proviruses, particularly those with large deletions that might resemble spliced transcripts. An investigation of NERV gene expression is likely to reveal strategies for gene expression that are as distinct from other ERVs as these endogenous elements appear at the viral genome level, but may more closely resemble the foamy viruses of the family *Spumaviridae* for which both PR-Pol and ENV polyproteins are expressed independently from spliced subgenomic transcripts instead of as a read-through Gag-PR-Pol polyprotein [10, 48–50]. The potential translation of an NERV PR-Pol polyprotein from subgenomic transcript could, however, also involve ribosomal frameshifting for clade III NERVs that possess a stop codon at the *pro* coding sequence terminus followed by *pol* in a different reading frame (Fig. 1).

The ability of retroviruses to capture genes is a phenomenon that has been well documented since the discovery of retroviral transduction of cellular oncogenes [51]. Viral transduction of host oncogenes may occur through multiple mechanisms, including illegitimate recombination between the DNA provirus and host genomic DNAs [52] or a multi-step RNA recombination mechanism requiring: (1) integration of a provirus upstream of a proto-oncogene, (2) DNA rearrangement that confers expression of a hybrid oncogene transcript with retroviral signals for co-packaging with viral genome, and (3) generation of a recombinant provirus by strand transfer reactions involving the heterodimeric co-packaged RNAs [53]. This last mechanism requires two recombination events that act at the DNA level and subsequently at the RNA level [54], but is probably the typical mechanism due to the frequent transcriptional read-through at the 3′ LTR [55]. Consequently, viral oncogenes such as *v-src* and *v-erbB* that have been transduced by avian leucosis viruses are frequently present in the 3′ end of retrovirus genomes [56, 57]. The Upf1 coding sequence resident in the 3′ end of some NERVs probably arose by similar viral transduction of a cellular gene. Upf1 is an ATP-dependent RNA helicase of the SFI superfamily functioning in nonsense-mediated decay (NMD) that has been identified as a component of the HIV-1 ribonucleoprotein (RNP) [58]. While Upf1 plays a key role in NMD to recruit enzymes that bring about degradation of mRNAs with premature termination codons and long 3′ UTRs [59], its association with HIV-1 serves a distinct function that stabilizes genomic viral RNA in both the cytoplasm and the nucleus, and promotes nuclear export of the unspliced HIV-1 genomic RNA [58, 60]. The NERV Upf1 may represent capture of a host factor for viral genomic RNA protection and trafficking in the same manner that HIV-1 hijacks cellular Upf1.

Previously, acquisition of an envelope by a retrotransposon from a heterologous virus has been inferred from the phylogenetic relationship of the baculovirus F protein and the *Errantivirus* Env [11], and may have occurred through a similar multistep mechanism of DNA integration. Errantiviruses reside in insect genomes and one member, the TED virus of *T. ni* (TniTEDV), has been found integrated into a mutant AcMNPV genome [61]. The mechanism for acquisition of the NERV *env* is probably distinct, however, as nyamiviruses and chrysoviruses are strictly RNA viruses without a DNA phase in which an integration event can initiate gene capture. Cellular mRNAs have been detected in retroviral virions, and reverse transcription of co-packaged mRNAs can generate DNA copies that form pseudogenes that are integrated within the genome [62]. NERV *env* capture could have occurred through a co-packaging mechanism and recombination of the heterologous viral RNA sequences during the reverse transcription strand transfer reactions, possibly through pairing of very short sequences with microhomology, as previously demonstrated for HIV-1 RT [63]. This possibility is further supported by the recent observation of sequences encoding an RNA-dependent RNA polymerase from a nodavirus, a positive-sense RNA virus, detected in the nematode *B. xylophilus* that were captured by a retrotransposon [64]. The ability of *env*-deleted retroviral genomes to be packaged into infectious virions with envelope glycoproteins from an RNA virus has been long known from pseudotyping experiments with the vesicular

stomatitis virus (VSV) G protein [65, 66]. VSV is a member of Mononegavirales, the order of minus-sense RNA viruses with a non-segmented genome that includes the *Nyamiviridae* [67]. Moreover, a lentivirus vector pseudotyped with VSV-G was able to transduce the animal-parasitic nematode *Nippostrongylus brasiliensis* [68], further supporting the potential functionality of the minus-sense RNA virus-derived Env in NERVs. The existence of RNA virus sequences as endogenous viral elements (EVEs) in DNA genomes has also been reported, with vertebrate genomes harbouring virus genes captured from ancient infections of viruses related to two Mononegavirales families, the *Filoviridae* and *Bornaviridae* [69], and a group of EVEs identified in the crustacean *Armadillidium vulgare* in a lineage related to the *Nyamiviridae* [70]. The vertebrate EVEs were suggested to have resulted from reverse transcription involving non-LTR retrotransposon LINE elements [69]. However, the growing resources of sequenced genomes of diverse organisms may present an unexplored treasure trove of EVEs that may include elements with hallmarks of retroviral reverse transcription as seen with the NERVs.

The apparent restriction of NERVs to nematodes may indicate that these endogenous elements are young, although we cannot exclude the possibility that older and/or divergent NERVs are present in other taxa of non-model organisms that are absent from the genome databases. Our search of WGS databases revealed instances of integration pre-speciation in the genus *Meloidogyne*, but we found no evidence of shared loci across genera that would be expected of more ancient ERVs. The presence of the MincNERV$_1$, MincNERV$_2$ and MincNERV$_3$ loci in *M. incognita* and *M. enterolobii* but their absence in *M. luci* indicates that these loci were present and undergoing vertical transmission, but still segregating in the last common ancestor of these modern species. These loci may have been lost in the lineages leading to some species in the genus *Meloidogyne*, but this cannot be determined from a single genome. NERVs were seen to have disseminated widely in PPNs of the suborder Tylenchina, as well as to the entomopathogenic nematode *Steinernema feltiae,* which has a soil-dwelling free-living form, and the genus *Caenorhabditis* that includes free-living nematodes with a natural habitat of decaying organic material. The absence of these retrovirus sequences from the well-studied model organism *C. elegans*, however, is probably why NERVs were not previously discovered. The distribution of NERVs could suggest a primary mode of cross-genera transmission in the soil and explain the large absence of these recently evolved NERVs from animal-parasitic nematodes. Indeed, the only instance of NERV detection in an animal-parasitic nematode species was in *A. simplex*, a medically important parasite of fish and marine mammals that can infect humans when fish is consumed raw. Although *A. simplex* is a marine nematode, transmission from soil may have been possible through run-off into ocean waters. The detection of NERV sequences in *A. simplex* but no other nematodes of the suborder Spirurina further supports the infectivity of NERVs.

Funding information
This work was made possible by funds to M.A.S. from California State University Fullerton Department of Biological Science. Research in I.K.'s lab is funded by USDA-AFRI (2017-67014-26699), Hatch project (1017522) and USDA Multistate project (1018832).

Acknowledgements
We acknowledge the Entomological Society of America for recently announcing their intention to rename the gypsy moth and gypsy ants due to the use of this word as a pejorative term for Romani people. In consideration for this term's use as a racial slur, we named DmeGypV in full only once for context and only used an abbreviation thereafter to minimize our use of a hurtful term, while also recognizing the importance of DmeGypV on our current understanding of retrotransposon evolution.

Conflicts of interest
The authors declare that there are no conflicts of interest.

References

1. Ellefson JW, Gollihar J, Shroff R, Shivram H, Iyer VR, *et al*. Synthetic evolutionary origin of a proofreading reverse transcriptase. *Science* 2016;352:1590–1593.
2. Goodier JL, Kazazian HH. Retrotransposons revisited: the restraint and rehabilitation of parasites. *Cell* 2008;135:23–35.
3. Krupovic M, Blomberg J, Coffin JM, Dasgupta I, Fan H, *et al*. Ortervirales: new virus order unifying five families of reverse-transcribing viruses. *J Virol* 2018;92:e00515-18.
4. Stewart L, Vogt VM. trans-acting viral protease is necessary and sufficient for activation of avian leukosis virus reverse transcriptase. *J Virol* 1991;65:6218–6231.
5. Craven RC, Bennett RP, Wills JW. Role of the avian retroviral protease in the activation of reverse transcriptase during virion assembly. *J Virol* 1991;65:6205–6217.
6. Bennett RP, Rhee S, Craven RC, Hunter E, Wills JW. Amino acids encoded downstream of gag are not required by Rous sarcoma virus protease during gag-mediated assembly. *J Virol* 1991;65:272–280.
7. Jacks T, Townsley K, Varmus HE, Majors J. Two efficient ribosomal frameshifting events are required for synthesis of mouse mammary tumor virus gag-related polyproteins. *Proc Natl Acad Sci U S A* 1987;84:4298–4302.
8. Moore R, Dixon M, Smith R, Peters G, Dickson C. Complete nucleotide sequence of a milk-transmitted mouse mammary tumor virus: two frameshift suppression events are required for translation of gag and pol. *J Virol* 1987;61:480–490.
9. Jacks T. Translational suppression in gene expression in retroviruses and retrotransposons. *Curr Top Microbiol Immunol* 1990;157:93–124.
10. Yu SF, Baldwin DN, Gwynn SR, Yendapalli S, Linial ML. Human foamy virus replication: a pathway distinct from that of retroviruses and hepadnaviruses. *Science* 1996;271:1579–1582.
11. Rohrmann GF, Karplus PA. Relatedness of baculovirus and gypsy retrotransposon envelope proteins. *BMC Evol Biol* 2001;1:1.
12. Bolinger C, Boris-Lawrie K. Mechanisms employed by retroviruses to exploit host factors for translational control of a complicated proteome. *Retrovirology* 2009;6:8.
13. Armbruester V, Sauter M, Krautkraemer E, Meese E, Kleiman A, *et al*. A novel gene from the human endogenous retrovirus K expressed in transformed cells. *Clin Cancer Res* 2002;8:1800–1807.

14. Magin C, Löwer R, Löwer J. cORF and RcRE, the Rev/Rex and RRE/RxRE homologues of the human endogenous retrovirus family HTDV/HERV-K. *J Virol* 1999;73:9496–9507.
15. Temin HM. Reverse transcription in the eukaryotic genome: retroviruses, pararetroviruses, retrotransposons, and retrotranscripts. *Mol Biol Evol* 1985;2:455–468.
16. Weiss RA. The discovery of endogenous retroviruses. *Retrovirology* 2006;3:67.
17. Mager DL, Henthorn PS. Identification of a retrovirus-like repetitive element in human DNA. *Proc Natl Acad Sci U S A* 1984;81:7510–7514.
18. Tarlinton RE, Meers J, Young PR. Retroviral invasion of the koala genome. *Nature* 2006;442:79–81.
19. Arkhipova IR, Mazo AM, Cherkasova VA, Gorelova TV, Schuppe NG, *et al*. The steps of reverse transcription of Drosophila mobile dispersed genetic elements and U3-R-U5 structure of their LTRs. *Cell* 1986;44:555–563.
20. Dhar R, McClements WL, Enquist LW, Vande Woude GF. Nucleotide sequences of integrated Moloney sarcoma provirus long terminal repeats and their host and viral junctions. *Proc Natl Acad Sci U S A* 1980;77:3937–3941.
21. Hayward A. Origin of the retroviruses: when, where, and how? *Curr Opin Virol* 2017;25:23–27.
22. Lee A, Nolan A, Watson J, Tristem M. Identification of an ancient endogenous retrovirus, predating the divergence of the placental mammals. *Philos Trans R Soc Lond B Biol Sci* 2013;368:20120503.
23. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, *et al*. Initial sequencing and analysis of the human genome. *Nature* 2001;409:860–921.
24. Boller K, Schönfeld K, Lischer S, Fischer N, Hoffmann A, *et al*. Human endogenous retrovirus HERV-K113 is capable of producing intact viral particles. *J Gen Virol* 2008;89:567–572.
25. Wang J, Han GZ. Frequent retroviral gene co-option during the evolution of vertebrates. *Mol Biol Evol* 2020;37:3232–3242.
26. Wang Z, Qu L, Yao J, Yang X, Li G, *et al*. An EAV-HP insertion in 5' Flanking region of SLCO1B3 causes blue eggshell in the chicken. *PLoS Genet* 2013;9:e1003183.
27. Mi S, Lee X, Li X, Veldman GM, Finnerty H, *et al*. Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. *Nature* 2000;403:785–789.
28. Pastuzyn ED, Day CE, Kearns RB, Kyrke-Smith M, Taibi AV, *et al*. The neuronal gene arc encodes a repurposed retrotransposon gag protein that mediates intercellular RNA transfer. *Cell* 2018;172:275–288.
29. C. elegans Sequencing Consortium. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* 1998;282:2012–2018.
30. Bowen NJ, McDonald JF. Genomic analysis of *Caenorhabditis elegans* reveals ancient families of retroviral-like elements. *Genome Res* 1999;9:924–935.
31. Britten RJ. Active gypsy/Ty3 retrotransposons or retroviruses in *Caenorhabditis elegans*. *Proc Natl Acad Sci U S A* 1995;92:599–601.
32. Llorens C, Soriano B, Krupovic M, Ictv Report C. ICTV virus taxonomy profile: metaviridae. *J Gen Virol* 2020;101:1131–1132.
33. Soriano B, Krupovic M, Llorens C. ICTV virus taxonomy profile: belpaoviridae 2021. *J Gen Virol* 2021;102.
34. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol* 1990;215:403–410.
35. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, *et al*. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* 2011;7:539.
36. Gouy M, Guindon S, Gascuel O. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol Biol Evol* 2010;27:221–224.
37. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJ. The phyre2 web portal for protein modeling, prediction and analysis. *Nat Protoc* 2015;10:845–858.
38. Smythe AB, Holovachov O, Kocot KM. Improved phylogenomic sampling of free-living nematodes enhances resolution of higher-level nematode phylogeny. *BMC Evol Biol* 2019;19:121.
39. Hussey RS, Barker K. A comparison of methods of collecting inocula of *Meloidogyne* spp., including A new technique. *Plant Disease Reporter* 1973;57:1025–1028.
40. Doyle JJ, Doyle JL. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemical Bulletin* 1987;19:11–15.
41. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, *et al*. A unified classification system for eukaryotic transposable elements. *Nat Rev Genet* 2007;8:973–982.
42. McClure MA, Johnson MS, Feng DF, Doolittle RF. Sequence comparisons of retroviral proteins: relative rates of change and general phylogeny. *Proc Natl Acad Sci U S A* 1988;85:2469–2473.
43. Malik HS, Henikoff S, Eickbush TH. Poised for contagion: evolutionary origins of the infectious abilities of invertebrate retroviruses. *Genome Res* 2000;10:1307–1318.
44. Snapp EL, McCaul N, Quandte M, Cabartova Z, Bontjer I, *et al*. Structure and topology around the cleavage site regulate post-translational cleavage of the HIV-1 gp160 signal peptide. *Elife* 2017;6:e26067.
45. Rein A. Across the hall from pioneers. *Viruses* 2021;13:491.
46. Belshaw R, Pereira V, Katzourakis A, Talbot G, Paces J, *et al*. Long-term reinfection of the human genome by endogenous retroviruses. *Proc Natl Acad Sci U S A* 2004;101:4894–4899.
47. Belshaw R, Katzourakis A, Paces J, Burt A, Tristem M. High copy number in human endogenous retrovirus families is associated with copying mechanisms in addition to reinfection. *Mol Biol Evol* 2005;22:814–817.
48. Bodem J, Löchelt M, Winkler I, Flower RP, Delius H, *et al*. Characterization of the spliced pol transcript of feline foamy virus: the splice acceptor site of the pol transcript is located in gag of foamy viruses. *J Virol* 1996;70:9024–9027.
49. Jordan I, Enssle J, Güttler E, Mauer B, Rethwilm A. Expression of human foamy virus reverse transcriptase involves a spliced pol mRNA. *Virology* 1996;224:314–319.
50. Enssle J, Jordan I, Mauer B, Rethwilm A. Foamy virus reverse transcriptase is expressed independently from the Gag protein. *Proc Natl Acad Sci U S A* 1996;93:4137–4141.
51. Stehelin D, Varmus HE, Bishop JM, Vogt PK. DNA related to the transforming gene(s) of avian sarcoma viruses is present in normal avian DNA. *Nature* 1976;260:170–173.
52. Goodrich DW, Duesberg PH. Retroviral transduction of oncogenic sequences involves viral DNA instead of RNA. *Proc Natl Acad Sci U S A* 1988;85:3733–3737.
53. Swanstrom R, Parker RC, Varmus HE, Bishop JM. Transduction of a cellular oncogene: the genesis of Rous sarcoma virus. *Proc Natl Acad Sci U S A* 1983;80:2519–2523.
54. Stuhlmann H, Dieckmann M, Berg P. Transduction of cellular neo mRNA by retrovirus-mediated recombination. *J Virol* 1990;64:5783–5796.
55. Herman SA, Coffin JM. Efficient packaging of readthrough RNA in ALV: implications for oncogene transduction. *Science* 1987;236:845–848.
56. Raines MA, Maihle NJ, Moscovici C, Crittenden L, Kung HJ. Mechanism of C-erbB transduction: newly released transducing viruses retain poly(A) tracts of erbB transcripts and encode C-terminally intact erbB proteins. *J Virol* 1988;62:2437–2443.
57. Coffin JM. Genes responsible for transformation by avian RNA tumor viruses. *Cancer Res* 1976;36:4282–4288.
58. Ajamian L, Abrahamyan L, Milev M, Ivanov PV, Kulozik AE, *et al*. Unexpected roles for UPF1 in HIV-1 RNA metabolism and translation. *RNA* 2008;14:914–927.
59. Lykke-Andersen S, Jensen TH. Nonsense-mediated mRNA decay: an intricate machinery that shapes transcriptomes. *Nat Rev Mol Cell Biol* 2015;16:665–677.

60. Ajamian L, Abel K, Rao S, Vyboh K, García-de-Gracia F, *et al*. HIV-1 Recruits UPF1 but excludes UPF2 to promote nucleocytoplasmic export of the genomic RNA. *Biomolecules* 2015;5:2808–2839.
61. Miller DW, Miller LK. A virus mutant with an insertion of copia-like transposable element. *Nature* 1982;299:562–564.
62. Linial M. Creation of a processed pseudogene by retroviral infection. *Cell* 1987;49:93–102.
63. Oz-Gleenberg I, Herschhorn A, Hizi A. Reverse transcriptases can clamp together nucleic acids strands with two complementary bases at their 3'-termini for initiating DNA synthesis. *Nucleic Acids Res* 2011;39:1042–1053.
64. Cotton JA, Steinbiss S, Yokoi T, Tsai IJ, Kikuchi T. An expressed, endogenous Nodavirus-like element captured by a retrotransposon in the genome of the plant parasitic nematode *Bursaphelenchus xylophilus*. *Sci Rep* 2016;6:39749.
65. Love DN, Weiss RA. Pseudotypes of vesicular stomatitis virus determined by exogenous and endogenous avian RNA tumor viruses. *Virology* 1974;57:271–278.
66. Weiss RA, Wong AL. Phenotypic mixing between avian and mammalian RNA tumor viruses: I. Envelope pseudotypes of Rous sarcoma virus. *Virology* 1977;76:826–834.
67. Dietzgen RG, Ghedin E, Jiāng D, Kuhn JH, Song T, *et al*. ICTV virus taxonomy profile: nyamiviridae. *J Gen Virol* 2017;98:2914–2915.
68. Hagen J, Sarkies P, Selkirk ME. Lentiviral transduction facilitates RNA interference in the nematode parasite *Nippostrongylus brasiliensis*. *PLoS Pathog* 2021;17:e1009286.
69. Belyi VA, Levine AJ, Skalka AM. Unexpected inheritance: multiple integrations of ancient bornavirus and ebolavirus/marburgvirus sequences in vertebrate genomes. *PLoS Pathog* 2010;6:e1001030.
70. Thézé J, Leclercq S, Moumen B, Cordaux R, Gilbert C. Remarkable diversity of endogenous viruses in a crustacean genome. *Genome Biol Evol* 2014;6:2129–2140.
71. Álvarez-Ortega S, Brito JA, Subbotin SA. Multigene phylogeny of root-knot nematodes and molecular characterization of *Meloidogyne nataliei* Golden, Rose & Bird, 1981 (Nematoda: Tylenchida). *Sci Rep* 2019;9:11788.