

The discovery of multiple active mys-related LTR-retroelements within the Neotominae subfamily of cricetid rodents

David H Kass (✉ dkass@emich.edu)

Eastern Michigan University

Sarah Beatty

Eastern Michigan University

Ashlee Smith

Eastern Michigan University

Megan Scott

Eastern Michigan University

Dishita Shah

Eastern Michigan University

Mary Czaplicki

Eastern Michigan University

Research Article

Keywords: Genome Evolution, LTR-retroelements, Peromyscus, Neotominae, Retrotransposons

Posted Date: November 16th, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-2129589/v2>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Genetica on March 4th, 2023. See the published version at <https://doi.org/10.1007/s10709-023-00183-z>.

Abstract

Retrotransposon families in the rodent family Cricetidae have been understudied in contrast to Muridae, both taxa classified within the superfamily Muroidea. Therefore, we pursued advancing our knowledge of the unique *mys*-1 LTR-retroelement identified in *Peromyscus leucopus* by utilizing intra-ORF PCR, quantitative dot blots, DNA and protein library screens, the generation of molecular phylogenies, and analyses of orthologous LTR-retroelement loci. This led to the discovery of three additional related families of LTR-retroelements. These include a 2900 bp full-length element of *mys*-related sequences (*mys*RS), an 8,000 bp element containing the *mys* ORF1 sequence (mORF1) with ERV-related sequences downstream in the reverse orientation, as well as an 1800 bp element primarily consisting of *mys* ORF2 (mORF2) related sequences flanked by LTRs. Our data revealed a very limited number of full-length *mys* elements among genera of the Neotominae subfamily of cricetid rodents. Additionally, although we estimated roughly 5,000–10,000 genomic copies of the *mys*RS element among Neotominae, most exist as partial copies. The mORF1 elements are also limited to the genomes of the Neotominae subfamily of cricetid rodents, whereas mORF2 appears limited to the *Peromyscus* genus. Molecular phylogenies demonstrating concerted evolution, along with an assessment of orthologous loci in *Peromyscus* for the presence or absence of elements, are consistent with recent activity of these novel LTR-retroelement families. Together with known activity of various families of non-LTR retroelements in *Peromyscus* species, we propose that retrotransposons have been continually contributing to the dynamics of *Peromyscus* genomes and are plausibly associated with the evolution of the over 50 identified species.

Introduction

Retrotransposons represent the class I group of transposable elements (TEs) that result from the spread of non-functional copies throughout genomes via an RNA intermediate derived from a limited number of source or master genes (Deininger et al. 1992). These include the non-long terminal repeat (LTR) long and short interspersed elements, LINEs and SINEs respectively, which comprise about 27% of the European house mouse genome (Mouse Genome Sequencing Consortium 2002) and over 33% in the human genome (Lander et al. 2001), thereby representing major contributors to the shaping and evolving of mammalian genomes. Autonomous elements either contain long terminal repeats (LTRs) such as those found in retroviruses and endogenous retroviral sequences (ERVs), or lack LTRs, such as LINE-1 (L1) which represents the predominant element in mammalian genomes (Ding et al. 2006). In contrast to L1, LTR-retrotransposons only comprise about 8% of the human genome, with no known active elements upon the recent extinction of HERV-K mobility (Holloway et al. 2019). In contrast, LTR-retroelements comprise about 10% of the house mouse genome with some, such as IAP and ETn elements, demonstrating recent activity (Mouse Genome Sequencing Consortium 2002). These elements, plus additional families of mammalian retrotransposons continue to be active (Woods-Samuels et al. 1989; Wallace et al. 1991; Kass et al. 1994; Roy et al. 1998; Kass et al. 2000; Kass and Jamison 2007; among others), with most integrations having a neutral impact. However, negative impacts via insertional mutagenesis have been responsible for various disorders in humans and mice (reviewed in Kass 2001;

Kazazian and Moran 2017). Alternatively retroelements have been co-opted by genes (Gifford et al., 2018) leading to genetic diversity as a result of mechanisms such as the formation of new genes by exon shuffling via 3' transduction (Moran et al. 1999), by altering the regulation of existing genes (reviewed in Richardson et al. 2015; Elbarbary et al. 2016), by potential involvement in X chromosome inactivation (Cantrell et al. 2009; Chow et al. 2010), by promoting DNA repair (Morrish et al. 2002), and by activation of miRNA (Piriyapongsa et al. 2007). Therefore, there is convincing support to assert that families of retroelements are major contributors to the dynamics of mammalian genomes thereby bearing a fundamental role in the evolution of species.

In contrast to the Muridae family of rodents, which include the species of mice and rats representing the commonly used model organisms, genomes, particularly their genomic components such as retroelements, members of the Cricetidae family (Fig. 1a) are far understudied. This investigation follows up on the study of *mys* elements that were identified by phylogenetic screening of a *Peromyscus leucopus* genomic library (Wichman et al., 1985), and represent a novel family of LTR-retroelements containing two open reading frames, instead of the standard retroviral *gag-pro-pol* genes found in endogenous retroviral (ERV) LTR-retroelements. Southern blot analyses based on banding patterns have supported this element as being highly active in the genus *Peromyscus* with multiple rounds of transposition (Lee et al. 1996) including the generation of five *mys* subfamilies in the *P. maniculatus*/*P. leucopus* clade (Lee et al. 1996). Additionally, the finding of elements present in specific loci limited to certain species (Cantrell et al. 2005) as well as within individuals of a particular species (Sawby and Wichman 1997) support recent activity. Estimates of *mys* elements in the *P. leucopus* genome have ranged from between 500-1,000

(Wichman et al. 1985) to about 4,100 copies (Janacek et al. 1993). The *mys*-9 member of this LTR-retroelement family, also isolated by phylogenetic screening differed from *mys*-1 in the restriction endonuclease pattern (Wichman et al. 1985) and is seemingly limited to the *P. maniculatus*-*P. leucopus* clade, *P. difficilis* and its sister species *P. truei* (Fig. 1b), as well as *P. crinitus* (Cantrell et al. 2001). Additionally, the size of this element among the orthologous loci of these *Peromyscus* species varies due to separate integrations of B1, B2, and ID SINE retrotransposons (Cantrell et al. 2001). However, to date, the *mys*-1 element from *P. leucopus* has been the only isolated full-length representative of the *mys* family with nearly intact open reading frames, and for that reason, *mys* and *mys*-1 terminology will be used interchangeably to represent this LTR-retroelement family in this study. Therefore, the lack of information on this LTR-retroelement family warranted a further detailed investigation, particularly in *Peromyscus* (deer mice and their relatives) as this genus represents the most populous species of New World Rodents and includes over 50 species as well as numerous subspecies (Carleton 1989), with *P. maniculatus* being the most common and widespread. In general, the *Peromyscus* genus offers a unique opportunity to further delve into the activity, genetic impact, and evolution of species resulting from retrotransposon activity based on the large number of species, as well as recognized subspecies.

We present a detailed investigation commencing with analysis into the relatively understudied, novel *mys* element, as (1) it does not correspond with the non-LTR mammalian L1 elements nor to LTR-ERV

elements, (2) there is a lack of identification of additional full-length *mys* elements, and (3) copy numbers of *mys* are potentially overestimated in *Peromyscus* genomes. This study led to the identification of three additional families of LTR-retrotransposons, including a more abundant family of *mys*-related sequences or *mysRS* elements, novel large full-length elements that contain only the ORF1 portion of *mys* (mORF1 family) but encompass ERV-related sequences in the opposite orientation, as well as elements containing only the ORF2 portion of *mys* (mORF2) flanked by LTRs. Additionally, the *mys*, *mysRS*, and mORF1 LTR-retroelement families appear limited to genomes of rodents designated within the Neotominae subfamily of the Cricetidae family of rodents, with mORF2 limited to the *Peromyscus* genus, thereby supporting relatively recent origins. In addition to providing support for relatively recent activity of the four retroelement families detailed in this study along with known activity of several retrotransposon families in *Peromyscus* (Kass et al 1992; Kass et al. 1996; Casavant et al. 1996; Rinehart et al. 2005; Yang et al. 2019) we propose an association between retrotransposon activity with the large number of *Peromyscus* species.

Materials And Methods

Dna And Tissue Samples

Genomic DNA samples of *Peromyscus aztecus* (AM), *P. boylii*, *P. californicus* (IS), *P. difficilis*, *P. eremicus* (EP), *P. gossypinus*, *P. leucopus* (LL), *P. maniculatus* (BW), *P. melanophrys* (XZ), *P. melanotis*, *P. polionotus* (PO), and *P. truei* were purchased through the *Peromyscus* Genetic Stock Center at the University of South Carolina. *Mus musculus* Balb/cJ DNA was purchased through the Jackson Laboratory. *Meriones unguiculatus* DNA was provided by Dr. M. Richard Shen at Lawrence Livermore National Laboratory, and DNA from *Pan troglodytes* provided by Dr. Mark Batzer at Louisiana State University. Tissues obtained through the Museum of Southwestern Biology (with loan numbers) include *Tylomys tumbalensis* (MSB 55652), *Neotomodon alstoni* (NK 9652), *Neotoma albigula* (MSB 140883), *Reithrodontomys fulvescens* (NK 43090), *Onychomys leucogaster* (NK 86784), *Oryzomys palustris* (MSB 57563), and *Sigmodon hispidus* (MSB 140165). *Octodon degus* tissue was provided by Dr. Theresa Lee at the University of Michigan. *Sciurus carolinensis* tissue was made available from a wild caught animal. Tissues from *Cavia porcellus*, and *Mesocricetus auratus* were purchased from Harlan Sprague Dawley, Inc. *Rattus rattus* tissue was purchased through Rockland Immunochemicals.

Isolation Of Dna

DNA was isolated from 20–50 mg of tissue finely chopped with a clean razor blade on a Petri dish on ice in chilled nuclei lysis solution from the Wizard Genomic DNA Purification Kit (Promega). The isolation of DNA was completed per manufacturer's protocol. Concentration of DNA and purity were assessed with a Nanodrop UV spectrophotometer.

Intra-orf Pcr

Throughout the experiment different sources of *Taq* polymerase were used, however the conditions were the same. In a 25 ml volume, the final concentrations of reagents included 1x PCR buffer (1.5mM MgCl₂), 200 mM dNTPs, 200 nM of both forward and reverse primers, 1 unit of *Taq* polymerase, and 25 ng of DNA. PCR conditions using an MJ thermocycler and 0.2 ml thin-wall tubes, were 94°C for two minutes, 30 cycles of 94°C for 30 s, 53°C for 30 s, and 72°C for 30 s, followed by a final extension step at 72°C for 5 m. Primers for *mys* ORF1 amplification were F671 (5'-CTGGACAGTCACCTGATGTTTC-3') and R1100 (5'-AAGTCTTCTAGGTATACATA-3'). For *mys* ORF2 amplification primers used were F1580 (5'-GAGCCAATTCCACCTCTAA-3') and R2210 (5'-CTGAGACTTGATGACTCCCA-3') with primer numbers corresponding to the *P. leucopus mys*-1 element. For *mys*RS amplification primers used were F60 (5'-TAAAGTTGGGGCACCTGTTTCA-3') and R430 (5'-TTTAAAAATTTAGACGTTCTG-3') with numbers corresponding to the isolated *mys*RS element PCR product from *N. alstoni*, that was generated using *mys* primers. Additional ORF1 PCR was performed based on highly conserved sequences from intra-ORF1 PCR DNA libraries using primer 1F (5'-AGCATTGTTGGTGCATT-3') with either R239 (5'-TCCAMTTGGAAGTTTGTCTTC-3') or R255 (5'-GCTTCTAAGACATTTCCAG-3').

Quantitative dot blot analyses for *mys* and *mys*RS elements

DNA stocks were analyzed for purity and concentration using a Nanodrop spectrophotometer. Two readings were taken for each sample. The DNA samples were diluted in purified, sterile water to create stocks of 100 ng/ml concentrations. Copy number standards were generated for *mys* ORF1 using a *P. gossypinus* clone containing an insert generated by amplification using F671-R1100 primers; for *mys* ORF2 a *P. polionotus* clone containing an insert generated by amplification with F1580-R2210 primers; for *mys*RS a *P. maniculatus (nubiterrae)* clone containing an insert generated by amplification with F60-R430 primers. The guinea pig L1 control was amplified using primers F5082 (5'-TATCACTATTTGCAGATGA-3') and R5670 (5'-TYTYTRTRCCARTACCA-3') based on highly conserved ORF2 sequences among mammals and amplified under the same intra-ORF PCR conditions (except 50°C annealing) as the other elements. Clones selected for dot blot probes and standards had been sequenced and chosen based on close sequence identity to a consensus sequence. Each clone was digested with *Eco*RI and size fractionated using a 1.5% agarose gel stained with 0.5x GelRed (Biotium, Fremont, CA) or ethidium bromide. To limit plasmid sequence background, the inserts were cut out of the gel, cleaned with a gel extraction kit (Bio Basic, Amherst, NY), reamplified using 1.0 ng and 0.1 ng of gel-extracted DNA by PCR with the original primer sets, run on 1.5% agarose gels, with isolation and extraction of the DNA repeated. The DNA for the standards were assessed for purity and concentration using the NanoDrop spectrophotometer. The DNA was diluted in water in sterile siliconized microcentrifuge tubes.

To estimate copy numbers for standards we based the size of the segment of DNA used as the probe as a proportion of the genome. For example, the generated *mys* ORF1 clones were 371 bp, and with an estimated genome size for the various rodents roughly 3×10^9 bp, then one copy of *mys* would correspond to $371 \text{ bp} / 3 \times 10^9 \text{ bp}$ representing a 1.24×10^{-7} proportion of the genome. Therefore, for 1.0 mg of total DNA, one copy would have a mass of 1.24×10^{-7} mg; 10 copies would then correspond to

1.24 x 10⁻⁶ mg, etc. Once these determinations were made for each set of standards, one mg of denatured total genomic DNA was loaded onto the blots, the number of copies contained in each genome would be based on the corresponding intensity of the dot to the intensities of the different standards.

Dot Blot Screening Procedures

DNA standards were constructed using the estimated amount of DNA corresponding to copy numbers and one mg of genomic DNA from various rodents in 500 ml volumes containing 10mM EDTA and 0.4 M NaOH. Samples were denatured by heating in a boiling water bath for 10 m, pulsed in a microcentrifuge and placed on ice. Two thin pieces of chromatography paper plus a precut 0.45mm nylon membrane (Biodyne, Fort Wayne, IN) were submerged in water for 10 m and placed on a dot blot apparatus (V & P Scientific, San Diego, CA) and clamped. 500 ml of water was added to the applicable wells and pulled through with vacuum. The entire amount of denatured DNA was added to the wells and “vacuum” absorbed onto the filter, followed by the addition of 400 ml of 0.4M NaOH. The membrane was briefly rinsed in 2xSSC, blotted on chromatography paper, air dried, placed in plastic wrap. The DNA was covalently bound to the membrane using a UV-crosslinker and followed by baking in a hybridization oven at 80°C for 30 m, then stored in a sealed bag until ready for use. Probes were generated using 150–250 ng of gel-extracted PCR-amplified DNA from selected clones using a random primer label kit (Takara Bio, San Diego, CA) according to the manufacturer’s protocol with dCTP labeled with ³²P (MP Biomedicals, Irvine, CA). Unincorporated ³²P-dCTP was removed utilizing a P-30 column (Bio-Rad, Brea, CA). A Beckman LS 6500 scintillation counter was used to determine level of radioactivity of the probe. Filters were prehybridized in a glass tube with Ambion ULTRAhyb (Thermo-Fisher, Walton, MA) solution at 42°C for 1 hour in roller hybridization oven, then denatured probe (approximately 2 x 10⁶ cpm) was added to the tube for overnight hybridization. The hybridization solution was removed, then filters were washed in a 10 ml prewarmed solution of 2xSSC, 0.1%SDS for 5 m at 42°C twice, followed by a 15 m wash in a solution of 0.2XSSC, 0.1%SDS at 42°C. Filters were blotted on chromatography paper, wrapped in plastic film, and placed on a reusable phosphorimager screen (blanked using a FujiFilm IP Eraser BAS) in an autoradiography cassette with at least three hours exposure. Results were visualized using a FujiFilm FLA-3000 phosphorimager.

Generating Intra-orf Pcr Dna Libraries

Mys intra-ORF1 and intra-ORF2 PCR DNA libraries, as well as *mysRS* intra-ORF1 PCR DNA libraries from various rodents were generated by direct ligation of isolated PCR products into the pGEM-T Easy vector using the pGEM-T Easy Vector System (Promega, Madison, WI) following the manufacturer’s protocol, followed by transformation of competent JM109 *E. coli* cells.

Isolation And Characterization Of Individual Intra-orf Clones

Transformed bacteria were grown on LB plates with 100 ng/ml ampicillin, with 100 ml of 100 mM IPTG and 20 ml of 50 g/mL X-gal and incubated at 37°C for 16–24 hours. Individual white colonies were selected and used to inoculate 1.5 ml liquid LB-ampicillin and grown at 37°C overnight. Plasmid DNA was isolated using the Wizard Plus SV Minipreps DNA Purification Systems (Promega). Clones were checked for inserts via *EcoRI* restriction enzyme digests and 1.5% agarose gel electrophoresis stained with 0.5x GelRed. Clones with inserts of roughly the expected size were submitted to the Core Sequencing site at the University of Michigan for Sanger dideoxy sequencing with T7 or SP6 primers.

Analysis Of Cloned Dna Sequences

Cloned intra-ORF PCR DNA sequences were incorporated into files using MacVector DNA sequencing software (MacVector.com) and chromatograms were checked to verify quality of the results. Sequence alignments using ClustalW were performed initially using default settings (pairwise- slow alignment speed, open gap penalty of 15, and extended gap penalty of 6.66; multiple alignment – same gap penalties, 30% delay divergence and weighted transitions), although various settings were modified to improve alignments. Additionally, 5' and 3' end sequences that extend past the other sequences were edited out for the alignments to contain the same regions and avoid skewing results.

Whole Genome Sequencing

High quality DNA isolated from *P. aztecus* was submitted to the University of Michigan Sequencing Core (AnnArbor, MI) for whole-genome sequencing. Two PacBio SMRT cells were used generating between 2.5×10^9 - 2.7×10^9 bp sequence of DNA each, which is in the range of 2X total coverage based on an estimate of roughly 2.7–2.8 billion base pairs in *Peromyscus maniculatus* genome (ncbi.nlm.nih.gov/assembly/GCF_000500345.1). DNA isolated from *P. eremicus* spleen was submitted to GeneWiz (South Plainfield, NJ) for whole-genome sequencing also using two PacBio SMRT cells with about 8 billion base pairs sequenced or roughly 3X coverage though with larger average read lengths to improve the chance of identifying full-length LTR-retrotransposons.

Database Screens for full-length *mys*-1 and *mysRS* elements

For further insights into the *Peromyscus* LTR-retroelements, we screened the genomic DNA libraries and public databases to identify full-length *mys* and *mysRS* elements either via alignment to a reference sequence using MacVector or using a BLASTn screening of NCBI (GenBank) and Ensembl databases. These would be characterized by containing ORF1, ORF2, LTR, and flanking direct repeat (FDR) sequences. Using the lone identified full-length *P. leucopus mys* element (*mys*-1; accession number X02855.1) and conserved sequences from the intra-*mysRS* ORF PCR clones, we screened the generated *P. aztecus* and *P. eremicus* genomic libraries using MacVector by alignment to a reference sequence. A complete draft genomic sequence of *P. maniculatus* is available in the Ensembl genome browser, and 14 chromosome assemblies of *Onychomys torridus* used for our analyses are available in GenBank

(accession numbers LR877197.1-LR877210.1). Additional draft genome assemblies of *P. leucopus* (GCA_004664715.2 UCL_PerLeu_2.1), *P. californicus* (GCA_007827085.3), and *P. polionotus* (GCA_003704135.2 HU_Ppol_1.3.3) are publicly available through the Baylor College of Medicine Human Genome Sequencing Center (<https://www.hgsc.bcm.edu/other-mammals/peromyscus-genome-project#:~:text=About%20the%20Project,and%20the%20white%2Dfooted%20mouse>), and available to search through NCBI. Sequences obtained in this study that are not already available in either database were submitted to GenBank via BankIt. Upon identification of elements in which only the ORF1 or ORF2 segments of *mys* or *mysRS* were detected, we rescreened libraries to identify full-length elements until LTR and FDR sequences were identified by incorporating into MacVector files and comparing similarities of 5' and 3' sequences of individual elements.

Orf Analysis

Full-length sequences were analyzed for ORFs for the six possible frames utilizing this tool in MacVector software, and in order to identify long ORFs in general without the requirement of a start codon we used the ORF Finder tool in the Sequence Manipulation Suite (SMS) web page (https://www.bioinformatics.org/sms2/orf_find.html).

Generation Of Molecular Phylogenies

To compare the relationships of the obtained sequences molecular phylogenies were generated using maximum likelihood (ML) with MEGA11 (Version 11.0.13) DNA Sequence Software (Tamura et al. 2021). The MacVector ClustalW alignments were exported into FASTA format, then into MEGA format. The Muscle program was then used in regenerating alignments. This was followed using the “models” tool in MEGA to determine the best method to generate alignments, which were typically either HKY + G or T92 + G parameter models, with 500 replicates. The bootstrap values and estimated pairwise distances were generated by applying NJ and BioNJ algorithms, with the method used and highest log likelihood values are provided in the figure legends of the generated molecular phylogenies.

Results

Analysis of *mys*-1 elements via intra-ORF1 and intra-ORF2 PCR amplification

To ascertain the evolution of the understudied *mys* elements in cricetid rodents, we utilized the *P. leucopus mys*-1 LTR-retrotransposon as the focal point as it represents the only full-length *mys* element with nearly intact open reading frames. Intra-ORF1 PCR products using primer sets corresponding to nucleotides of *P. leucopus mys*-1 620-1100 (data not shown) and 671-1100 were consistent as amplicons among *Peromyscus* species (Fig. 2a) were of the expected size, albeit more intense in members of the *P. maniculatus*/*P. leucopus* clade, which includes *P. gossypinus*. This was confirmed upon substituting a higher quality *P. gossypinus* DNA sample for PCR (Fig. 2b). Additionally, amplicons were only faintly

observed, but limited to, members of the Neotominae subfamily of cricetid rodents (Fig. 2b). Smaller sized amplicons were observed among *Peromyscus*, although considering that only full-length fragments were obtained via shotgun cloning, these size variants are potentially artifact bands, conformational variants, or primer-amplicon heteroduplexes. Additionally, we cloned and sequenced intra-ORF1 PCR products from *Peromyscus* and non-*Peromyscus* species (data not shown) and determined highly conserved sequences for primers to further enhance the verification for the presence of *mys* using intra-ORF1 PCR. Using the additional primers, an amplicon was generated for *N. albigula* (Fig. 2c) which is consistent with the restricted presence of *mys-1* in Neotominae rodents. Upon generating an intra-ORF (1F-239R) PCR DNA library for *N. albigula* we obtained sequences, that when queried to the NCBI collection of databases using BLASTn, most closely matched *P. leucopus mys-1* with 98% coverage (C) and 82% identity (I) consistent with the presence of *mys-1* among all analyzed rodents classified in the Neotominae subfamily including both the Reithrodontomyini and Neotomini tribes.

ORF2 intra-*mys* PCR among species of *Peromyscus* (Fig. 2d) generated intense bands of expected size for species of the *P. leucopus*/*P. maniculatus* clade, along with a main band of higher molecular weight in the sister species of *P. truei* and *P. difficilis*, and a main amplicon of higher molecular weight in *P. eremicus* and *P. californicus* of the subgenus *Haplomylomys*. This initial pattern is consistent concerted evolution of *mys* elements.

Identification of a novel *mys*-related (*mys*RS) family of retrotransposons.

Upon sequencing intra-ORF1 PCR-generated clones in *Peromyscus maniculatus* and *Neotomodon alstoni* (GenBank accession numbers ON778008-ON778016), we found all the *P. maniculatus* clones and two of four *N. alstoni* clones aligned well to *mys-1*, whereas the other two *N. alstoni* clones appeared similar but distinct from *mys-1* (Fig. 3), seemingly representing a novel discovery of a family of *mys*-related sequences (*mrs*), that we refer to as *mys*RS. elements. The origin and evolutionary history of the *mys*RS element in contrast to *mys-1* was therefore further investigated. Utilizing a primer set in which sequences were distinct from *mys-1* but highly conserved between the two *N. alstoni* clones, PCR products of the expected size were observed among representatives of the Reithrodontomyini tribe (Fig. 4), seemingly absent from *N. albigula*. This suggests *mys*RS may be of more recent origin than *mys*. It was observed through a non-quantitative analysis, that the relative intensity of the PCR products was more equivalent among Reithrodontomyini in contrast to *mys-1* suggesting less variation in *mys*RS sequences among the genera.

Individual intra-*mys*RS clones from *N. alstoni*, *P. leucopus*, *R. fulvescens*, and *O. leucogaster* (GenBank accession numbers ON677500-ON766521) were sequenced and a molecular phylogeny was generated (Fig. 5), with inclusion of the corresponding region of the *P. leucopus mys-1* element (accession no. X02855.1). From these data, it is highly conceivable that *mys*RS and *mys* represent distinct cricetid retrotransposon families. Some clustering is observed among *N. alstoni* intra-*mys*RS clones, which might have been the result of PCR bias due to using *N. alstoni*-based primers. Alternatively, there may be some level of concerted evolution, with amplified products representing both younger and older elements.

Estimation of copy numbers of *mys* and *mysRS* families in cricetid rodents

Estimated copy numbers were determined for *mys* and *mysRS* families of retrotransposons using quantitative dot blots, which were performed to provide a general picture of the level of genomic amplification of these elements to possibly assess activity over time. In turn, this would provide further information regarding the evolutionary time of origins of the elements, as *mys* was previously determined to date back to a shared ancestor with the Sigmodontinae rodent subfamily (Wichman et al. 1985). Probes used were gel-isolated inserts from clones that were reamplified to avoid cross hybridization of standards that would result from plasmid sequences. ORF1 analysis (Fig. 6a) using a *mys* probe from *P. gossypinus*, as well as ORF2 analysis (Fig. 6b) using a *mys* probe from *P. polionotus* appeared to show over 20,000 copies of *mys*-1 among *Peromyscus* species, with either fewer, or more divergent, elements in non-*Peromyscus* genera of the Neotominae subfamily. The restriction of *mys*-1 to members of the Neotominae subfamily of Cricetidae is consistent with the PCR results (Fig. 2). Additionally, there was no apparent cross-hybridization to *mysRS* elements further depicting distinct LTR-retroelement families. A control dot blot was constructed (Fig. S1) using guinea pig L1 (ORF2 region) to further verify that roughly similar amounts of DNA were loaded on the filter, as the guinea pig is most distantly related to all the other analyzed rodents (Fig. 1a). As anticipated, the guinea pig presented the most intense band based on concerted evolution of L1 sequences. Some variability was observed regarding signal intensities with the L1 probe albeit generally in a similar range among analyzed rodents including among *Peromyscus*, and therefore the results should be considered rough estimates. The non-*Peromyscus* Neotominae rodents displayed dots of intensities corresponding to the lower range found among the *Peromyscus* species, and there was clear hybridization of the L1 to *S. hispidus*, along with similarity of signal for *O. palustris* to those of the non-*Peromyscus* Neotominae. Along with the signal observed for the gerbil in the control blot, this additionally supports the lack of hybridization of the *mys* and *mysRS* probes to non-Neotominae rodents. The cross hybridization of guinea pig L1 ORF2 to high amounts of the *mys* ORF2 standards demonstrated that there is some level of similarity of the L1 ORF2 to *mys* ORF2, which could result in an overestimation of the number of *mys* elements in our study.

The *mysRS* quantitative dot blot (Fig. 6c) displays hybridization to *N. albigula*, which in contrast to the intra-*mysRS* PCR results, indicates *mysRS* is not limited to the Reithrodontomyini tribe but as with *mys* is found in all analyzed Neotominae genera. It appears that *mysRS* copy numbers are roughly 5,000-10,000 with less of a difference to the non-*Peromyscus* Neotominae than observed for *mys*.

DNA sequence analyses to identify additional full-length *mys* elements and further assess its evolutionary history

The discovery of *mys*-1 also included the previous finding of ten additional elements characterized as *mys*, based on restriction maps of clones isolated via screening a *P. leucopus* genomic library with a *mys*-1 probe (Wichman et al. 1985). However, beyond low stringency Southern blots indicating that *mys*-1 subfamilies have been active throughout the evolution of *Peromyscus* species (Lee et al. 1996), there is a lack of additional support at the DNA sequence level. We therefore queried the full chromosomal

assembly of *P. maniculatus*, a related species (Fig. 1), available in Ensembl, using BLASTn, which generated 3,199 hits at $E = 1e-50$. Among these, we identified five sequences (Fig. S2) corresponding to nucleotides 20 or 21 up through 2059 of *P. leucopus mys-1*, with additional similarity further downstream. We analyzed the sequences for LTRs and FDRs, and four were full-length elements, with possibly a fifth, but sequences downstream of the element were unavailable for verification. Of note, each contains a small insertion in relation to the *P. leucopus mys-1* within ORF2 explaining the loss of corresponding sequences at roughly nucleotide 2059. A ClustalW alignment similarity matrix showed a range of 76%-86% sequence identity to *mys-1*, and these sequences were further verified as *mys* via a screen of Repbase (Jurka, 2000), with the indel being the result of an insertion of a segment related to the ERV2-N1_Mio-I elements. The *P. maniculatus* sequences displayed 80-90% sequence identity to each other. The Pm *mys* 23-3312091 element (Fig. 7), labeled based on chromosome number and approximate location, contained the longest ORF of 162 amino acids, and when translated to protein to screen protein databases in NCBI via BLASTp, the only match was to a sequence referred to as ORF1 PI (accession 1110236A). A screen using BLASTp queried with the ORF2 region yielded a match to ORF2 PI (1110236B). ORF1 PI and ORF2 PI sequences correspond to the *P. leucopus mys-1* translated ORFs, and therefore *mys* proteins do not correspond to that of any known retrotransposon. We did not identify any full-length *mys-1* elements in *P. eremicus* and *P. aztecus*, but this could be a result of using low coverage genomic libraries. We did, however, identify numerous hits to the LTR sequences, as well as to either ORF1 or ORF2 but not both (see mORF1 and mORF2 sections). We translated and performed a BLASTp screen of a small ORF associated with the reverse complement of two *P. maniculatus mys-1* loci which each yielded a very weak hit to a *Cricetulus griseus* putative *pol* gene (61% C, 44% I $E = 0.017$; 68% C 47% I $E = 3e-05$). Therefore, there is an apparent paucity of full-length *mys* elements in *Peromyscus* genomes.

To address the presence of full-length *mys-1* in other genera of cricetid rodents and to gain insights into its origin and evolution, we screened the golden hamster and prairie vole assemblies in Ensembl, as well as performed an overall screen of the NCBI database collection with the full-length Pm *mys* 23-3312091 (Pm *mys* 23) element. We obtained at $E = 1e-50$, 3134 hits among the *P. maniculatus* assembly, albeit just five were greater than 2,000 bp, and most roughly corresponded either to the first half or the second half of the element. At $1e-10$, the 5,000 hit maximum was obtained although the matches were similar. At $1e-50$, no matches were obtained for the prairie vole as well as for the hamster. At $1e-10$, 876 matches were obtained for the hamster, albeit the largest segment was 285 bp, and although the 5,000 hit maximum was obtained for the prairie vole, the maximum length was only 279 bp, with very few close to that length. We therefore narrowed down the screen to ORF1- and ORF2-based sequences of Pm *mys* 23. For ORF1 at $1e-50$ there were 779 hits of full-length or near full-length ORF1 sequence matches in the *P. maniculatus* screen, with no matches to prairie vole or golden hamster. When reducing the expected value to $1e-10$, the number of matches to the deer mouse increased to 2509 with again most being full or near full length ORF1 sequences, whereas at $1e-50$ there were no hits to both prairie vole and golden hamster, and at $1e-10$ 338 partial hits to the prairie vole and 88 partial hits to the golden hamster. Querying the databases with Pm *mys* 23 ORF2 at $1e-50$, 1796 full or near full-length ORF2 hits aligned to the deer mouse with none to hamster or prairie vole, and when reduced to $1e-10$, 2539 hits aligned to deer mouse with again

no matches to the prairie vole or hamster. For a more broad-based analysis we used Pm *mys* 23 to screen the NCBI databases using BLASTn. The top hits, besides to the *P. leucopus mys*-1 sequence were to *Onychomys torridus*. There were 4,341 matches to the combined 14 chromosomal assemblies available for *O. torridus*, with the top matches from each being full or near full length at E=0. These results are consistent with our PCR and dot blot data supporting the limitation of the *mys* LTR-retroelement to the Neotominae subfamily of cricetid rodents. The best match (referred to as Ot (PI) *mys* 11; with 11 indicating the *O. torridus* chromosome and PI indicating that the *P. leucopus mys*-1 was used for screening) displayed 85% similarity aligning to *P. leucopus mys*-1 for nucleotides 85-2048, though further analysis showed high similarity to *mys*-1 downstream. Other matches displayed 80% or less sequence identity and therefore not classified until further distinguishing *mys* and *mysRS* elements (next section). The Ot (PI) *mys* 11 element was added to the alignment of the *P. leucopus mys*-1 element and the five *P. maniculatus* full-length or near full-length elements, with LTRs, ORF1, ORF2 and FDRs depicted in Fig S2.

Identifying full-length *mysRS* elements

To identify full-length *mysRS* elements, we utilized the intra-ORF1 *N. alstoni* sequences to screen the available *P. maniculatus* genetic database in Ensembl and were able to identify and isolate full-length elements based on inclusion of 5' and 3' LTRs and FDRs, as represented by Pm 12-54005644 (Fig. 8). As with *mys*, the *P. maniculatus* sequences were designated based on chromosome and approximate location. We utilized a *P. maniculatus* full-length *mysRS* element (Pm *mysRS* 3-2991789) to rescreen the *Peromyscus* database in contrast to the *mys* screen, there were numerous matches obtained that contained the region encompassing both the ORF1 and ORF2 sequences (approximate nucleotides 612-2259), including the first 38 of 100 hits, suggesting the *mysRS* family is notably more common than *mys*. However, at E=1e-50, only six were apparently full length or near full length among the 2775 total hits. Using *mysRS* full-length, ORF1, and ORF2 sequences we queried the hamster and prairie vole databases and obtained zero hits (1e-50). We also screened the *P. eremicus* partial genomic library by aligning to a *Peromyscus mysRS* sequence (Pm 12-54005644) as a reference sequence in MacVector and isolated full-length elements from this species as well. We additionally screened the collection of databases in NCBI to obtain *mysRS* elements from additional genera and identified full-length *mysRS* elements in *Onychomys torridus*. The *O. torridus* elements isolated using either the *P. maniculatus mys* or *mysRS* elements all contained the ERV2-N1_Mio-I insert within ORF2 (highlighted in Fig. 7) and therefore this indel cannot be used as a criterium to distinguish *mys* and *mysRS*, although the *O. torridus* elements isolated by screening with the *P. leucopus mys*-1 did not contain the insert and may simply represent a separate *Onychomys mysRS* subfamily.

Distinguishing *mys*-1 vs *mysRS* LTR-Retroelements

To establish criteria to distinguish between *mys* and *mysRS* elements, we generated ClustalW alignments of full-length or near full-length isolated sequences (Fig. S3), followed by Muscle alignments in Mega11 from which molecular phylogenies using maximum likelihood were generated, that also were broken

down into the components of the of the elements that included ORF1, ORF2 and the 5'LTRs (Fig. 9). The molecular phylogeny of full-length (or nearly full-length) elements displayed several notable groupings (Fig. 9a). At the broadest level, the *Peromyscus* sequences formed a distinct clade from the *Onychomys* sequences. *O. torridus* sequences queried with the *P. leucopus mys-1* sequence formed a distinct clade from *O. torridus* sequences queried with either *mys* or *mysRS* sequences from *P. maniculatus*. At this point, the *P. leucopus* (PI)-derive *O. torridus* (Ot) sequences could represent *mys* elements or a subfamily of *mysRS* elements as there is a lack of a conspicuous mechanism for distinguishing these LTR-retroelement families. However, there is a clear demarcation in *Peromyscus* distinguishing *mys* and *mysRS* elements. Additionally, the *P. eremicus* *mysRS* sequences form a distinct clade indicating concerted evolution of sequences and in turn, recent activity. Of note, both Pm 6-20830407 and Pm 6-25307443 appear highly related, and their sequences continued to align beyond the 3' end of the full-length element (data not shown but available in Ensembl).

To further ascertain any clear distinction between *mys* and *mysRS* elements, particularly as it relates to *O. torridus*, sequences were trimmed to align the various components of these LTR-retroelements. ORF1 sequences were aligned via ClustalW and a notable distinction was shared between the Ot (PI) *mys* 11 sequence and the *Peromyscus* *mys* sequences, outlined in Fig. S4. A molecular phylogeny from an ORF1 Muscle alignment was generated (Fig. 9b) with findings that include a monophyletic group consisting of the Ot (PI) *mys* 11 sequence (that we characterized as *mys*), with the *P. maniculatus* and *P. leucopus* *mys* elements; the single *O. torridus* *mys* element formed an outgroup to the *Peromyscus* sequences; the *O. torridus* sequences obtained via the *P. leucopus mys-1* sequence formed a clade (with one exception); and although with weak bootstrap values, *Peromyscus* and *O. torridus* *mysRS* sequences formed separate clades; and *P. eremicus* sequences formed a monophyletic group. These findings suggest that ORF1 sequences potentially distinguish *mys* and *mysRS* elements, that there is concerted evolution of elements, and, in turn, *mysRS* elements are more common than *mys* elements in Neotominae rodents.

The relationships of the ORF2 regions of the full-length elements (highlighted in Fig. S5) were additionally analyzed. A molecular phylogeny of ORF2 sequences (Fig. 9c) most notably places all the *Onychomys* sequences into a distinct clade from the *Peromyscus* sequences (100% bootstrap value) consistent with concerted evolution irrespective of the retrotransposon family. Among *Peromyscus*, the ORF2 sequence alignment yielded a paraphyletic group of *P. maniculatus* *mys*, with the *P. leucopus mys-1* element forming a clade with *mys*, albeit with a low bootstrap value. Notably, this is not the result of the ERV2-n1_Mio-I insertion, as both *P. maniculatus* *mys* and *mysRS* elements contain the inserts, but not the *P. leucopus mys-1*. However, the insertion may be responsible for the distinct *O. torridus* (PI) clade as the sequences obtained via the *P. leucopus mys-1* query lack the insertion in ORF2 (Fig. S5) and is apparently a feature distinguishing distinct *mysRS* subfamilies in *O. torridus*. Two *O. torridus* (Ot) *mysRS* elements (Ot *mysRS*15A and Ot *mysRS*19) derived via a *P. maniculatus* *mysRS* screen formed a clade with the Ot (PI) subfamily, albeit as outliers, with a 100% bootstrap value. In general, ORF2 may, in a limited and less conspicuous fashion, distinguish *mys* and *mysRS* elements, but only in *Peromyscus* and not *Onychomys*, and the result of the ERV2-n1_Mio-I indel seemingly contributes to delineating *mysRS* subfamilies in *Onychomys*.

The 5'LTRs of the elements were additionally analyzed to establish if these sequences can be recognized as criteria for assessing age and activity of elements, as well as an additional method for distinguishing *mys* and *mysRS* elements. We aligned and generated a molecular phylogeny of the individual presumptive *mys* and *mysRS* elements to assess their similarities (Fig. 9d) and found the 5'LTRs of *Peromyscus mys* elements formed a clade with the *O. torridus* elements obtained via a query of *P. leucopus mys-1*, which then separated into distinct monophyletic groups. There were no clear patterns for the 5'LTRs of other *Peromyscus* and *Onychomys mysRS* elements. With the objective of distinguishing *mys* and *mysRS* elements, it appears that the LTR sequence can be a supporting factor but limited to within a genus as the sequence of the presumptive *O. torridus mys* element (Ot (Pl) *mys* 11) did not align with *Peromyscus mys* elements. Therefore, the overall picture of characterizing *mys* and *mysRS* elements appears to primarily be associated with ORF1 sequences, with the LTRs being an additional, but not exclusive, distinguishing feature.

Identification and analysis of a novel *mys* “ORF1-only” or mORF1 LTR-retroelement

With distinction of *mys* and *mysRS* elements primarily by ORF1 sequences and to some extent the LTR sequences we performed additional analyses to address the inconsistency between *mys* copy number estimates and the lack of full-length elements identified in *Peromyscus* and *Onychomys* genomic DNA libraries. While screening for *mys* elements in both the *P. eremicus* and *P. aztecus* DNA libraries (accession PRJNA847558) we identified ORF1 sequences not adjacent to ORF2 or “ORF1-only” elements, that commence at the same position. There were no full-length elements identified as the sequences vary at different points downstream of ORF1, or individual sequences were not long enough to make conclusions. In contrast, by querying the assemblies of greater coverage with ORF1 sequences of the *P. leucopus mys* element, we isolated sequences matching to *O. torridus* and *P. maniculatus*, then continued rescreening with more downstream sequences and were able to subsequently identify similar sequences that eventually reached the 3' LTRs and the flanking direct repeats with full-length elements of approximately 8,000 nucleotides, yet lacking *mys*- or *mysRS*-like ORF2 sequences. We are tentatively classifying this element (represented in Fig. 10) as mORF1 (LTR-retroelement containing the *mys* ORF1 sequence), being nomenclature of LTR retroelements are not well established (Gifford et al. 2018). A ClustalW alignment of the mORF1 LTR-retroelements identified in *O. torridus* and *P. maniculatus* (Fig. S6), seemingly demonstrates a high-level of sequence conservation spanning the entire sequence. Based on pairwise analyses, the *O. torridus* mORF1 elements exhibited between 94.2-99% sequence identity to each other, with *Peromyscus* mORF1 elements displaying between 69.8-85.3% sequence identity to each other. Additionally, a molecular phylogeny (Fig. 11) supports the elements in *O. torridus* being more related to each other than to *Peromyscus* elements, consistent with concerted evolution. The ORF1-region nucleotides, 551-1140 of *P. leucopus mys-1*, appear to be the only region *mys-1* has in common with these elements and therefore we added this stretch of sequences to the mORF1 alignment (Fig. S6). When screening the NCBI genetic database collection queried with the full-length *O. torridus* mORF1 23 element (the number corresponds to the chromosome harboring the element), 7,486 BLAST hits for the 14 chromosome assemblies were indicated. The displayed top match for each assembly corresponded to a full-length element. Although unclear of the estimated proportion of matches that are full-length 8,000 bp

sequences, a BLAST of the ORF1 region (e.g. nucleotides 620-1166 of Ot mORF1 23) yielded 6,288 hits with the graphic summary spanning the entire region. These numbers, in contrast to *mys-1*, are more consistent with the dot blot copy number estimates for *mys-1* ORF1, and likely the result of a greater number of mORF1 elements within Neotominae genomes. Additionally, when translating the sequences in different phases within the ORF1 region, BLASTp screens of protein databases yielded matches strictly to ORF *P. leucopus* (1110236A), which corresponds to *P. leucopus mys-1* ORF1.

We performed ORF analyses of the remaining portion of mORF1 elements, which notably yielded long stretches of amino acids in the reverse strand corresponding to sequences roughly between 1500-6000 nucleotides downstream of the ORF1 region. One example was an ORF of 1022 amino acids located in similar regions among three *O. torridus* elements which include Ot mORF1 16 (Fig. 12), Ot mORF1 20, and Ot mORF1 20A. The 1022 AA ORF in Ot mORF1 16 corresponds to nucleotides 5668-2600 of the element. The translated amino acid sequences for these three elements displayed greater than 99% sequence identity to each other suggesting these represent “young” elements. A BLASTp screen of the NCBI protein databases yielded a primary hit to a reverse transcriptase domain-containing protein of *Pseudomonas* bacteria (WP_217682179.1) with up to 83% sequence identity, followed by low similarity matches to HERV-K (up to 48%). Just downstream and adjacent to this ORF was a second ORF in the reverse complement strand of 452 corresponding to nucleotides 7036-5667 in Ot mORF1 16 (Fig. 12), and up to 534 amino acids in Ot mORF1 20. A BLASTp screen of the NCBI protein databases of this region of mORF1 among the *O. torridus* and *P. maniculatus* translated sequences displayed low identities (30-40%) to portions of *C. griseus gag-pol-pro* related sequences. ORFs were of smaller lengths within the *P. maniculatus* mORF1 elements. A BLASTp screen of the 167 amino acid sequence associated with an ORF near the 3' end of OT mORF1 16 (Fig. 12) generated no significant protein matches. Overall, it appears that the region downstream of the *mys-1* ORF1 sequence corresponds to endogenous retroviral sequences, most closely related to a *Pseudomonas* ERV. We additionally screened Repbase using the *O. torridus* 23 and *P. maniculatus* 7 mORF1 elements and ascertained that large sections of these elements corresponded to the reverse complement of the ERV2-2_Mio-I element, inclusive of the ORF1 region. Since ERV2-2_Mio-I is an ERV in the prairie vole and there is some protein similarity to *C. griseus* (hamster) ERVK genes, then it is possible this element may have arisen in an ancestor of cricetid rodents and differentiated over time. Alternatively, a unique integration occurred from a *Pseudomonas* retrovirus into the genome of a Neotominae ancestor that recombined to incorporate the *mys* ORF1 sequence. The large reading frames of *O. torridus* mORF1 elements with high similarity and the finding of several full-length elements, along with decipherable direct repeats (Fig. S6), suggests these elements may still be actively contributing to shaping genomes among genera in the Neotominae subfamily of cricetid rodents.

Identification and analysis of a novel *mys* “ORF2-only” retroelement or mORF2

In addition to identifying elements lacking ORF2 sequences, our *mys-1* screening led to the discovery of full-length elements containing ORF2 but lacking ORF1 sequences in *P. maniculatus*, *P. eremicus* and *P. aztecus*, supported by displaying 5' and 3' LTRs and FDRs, and are roughly 1860 bp in length (Fig. 13). Translation of the ORF2 region used to query NCBI protein databases via BLASTp yielded hits to ORF2 PI,

which corresponds to the *mys*-1 ORF2. We refer to these elements as. mORF2, based on the rationale for mORF1. A molecular phylogeny of identified *Peromyscus* mORF2 sequences (Fig. 14) displayed concerted evolution and therefore suggestive of recent activity. We found the 5' and 3' LTRs, along with the sequences downstream of the 3' end of ORF2, to display greater similarity to *mys* elements than to *mys*RS elements. A ClustalW alignment incorporating additional presumptive *Peromyscus* mORF2 sequences is represented in Fig. S7. When screening the *P. maniculatus* library in Ensembl with Pm 6-7752989 as the query sequence, at 1e-10, 32 of the top 100 hits showed nearly complete coverage with 88% sequence identity or higher. In contrast, an Ensembl BLASTn query to voles and hamster displayed matches of only 137 nucleotides or fewer, primarily in the ORF2 region, with the vole displaying 280 hits at 1e-10 with none identified via screening the golden hamster. The NCBI database collection was also screened with the Pa 11410 mORF2 element (as it demonstrated the longest ORF) and numerous significant hits to *O. torridus* were obtained, but notably aligning at nucleotide position 440, roughly where ORF2 begins. Therefore, the sequences of *O. torridus* upstream of this site do not represent a common 5' LTR to the *Peromyscus* mORF2 and likely contain ORF1, and therefore mORF2 sequences are plausibly lacking in *Onychomys* and therefore originated in an ancestor of *Peromyscus*. Additionally, the presence of this element, large numbers of segments of this element, and similarity to the *mys*RS ORF2 would potentially explain the copy number results obtained for the *mys* dot blot (Fig. 2c)

Contrasting *mys*, *mys*RS, mORF1, and mORF2

Overall, the detailed search for *mys*-1 elements led to the discovery of three additional related LTR-retrotransposon families in *Peromyscus*, two of which harbor only one of the ORFs related to *mys*-1. We aligned individual elements from each of the four LTR-retroelement families and generated a molecular phylogeny (Fig. 15) to ascertain their evolutionary relationships. Depicted from the molecular phylogeny is (1) *Peromyscus mys* elements form a distinct clade; (2) *Peromyscus mys*RS elements form a clade with *Peromyscus mys* forming a paraphyletic group; (3) *Peromyscus mys*RS forms a distinct clade from *O. torridus mys*RS (88% bootstrap value); (4) the *O. torridus mys*RS elements appear to form two distinct subfamilies based on whether the *P. leucopus mys*-1 versus either *P. maniculatus mys*-1 or *mys*RS elements were used as query sequences; and (5) the mORF1 family is the most distantly related, which is consistent with containing a large amount of sequence not found in the other retroelement families. We have classified Ot (PI) 11 as a *mys* element, although the full-length elements phylogenetically fall within the *O. torridus mys*RS element clade that were obtained via a *P. leucopus mys*-1 screen. Additionally, each retrotransposon family displays concerted evolution consistent with recent activity. To further address this point, we screened orthologous loci for the presence or absence of some of the individual elements.

Screening orthologous loci for recent retrotransposon activity

To assess relatively recent activity of the various LTR-retroelement families, we analyzed orthologous loci in *Peromyscus* for several individual elements (Table 1). In some cases, library screens queried with sequences flanking individual elements did not yield matches to the corresponding locus from the available assembly. We analyzed two *P. eremicus mys*RS element-containing orthologous loci in the

related *P. californicus*; both species in the subgenus *Haplomylomys* (Fig. 1), as well as in the more distantly related *P. maniculatus*. These individual elements were limited to *P. eremicus* (Table 1). This provides evidence of *mys*RS integrations that are limited to *P. eremicus*. We also analyzed five *P. maniculatus* *mys*RS elements of which three were present and two were absent in the closely related *P. leucopus* species. We were able to identify the orthologous locus in *P. polionotus*, the sister species of *P. maniculatus*, for one of these two absence loci in *P. leucopus* and found the presence of the element, indicating integration post-dating the divergence of the *P. leucopus*/*P. maniculatus* clades.

To assess recent *mys* integrations we analyzed four orthologous loci of the *P. maniculatus* *mys* elements in other *Peromyscus* species (Table 1). We found two to be present in *P. leucopus* and one absent, indicating activity of the latter occurred after the divergence leading up to these species. The fourth *mys* ortholog could not be identified in the *P. leucopus* assembly, although we determined presence in *P. polionotus*, the sister species of *P. maniculatus*.

To assess recent mORF2 LTR-retroelement activity we analyzed four individual mORF2-containing loci from *P. maniculatus* and *P. eremicus*. We found one *P. maniculatus* mORF2 element present in the *P. leucopus* ortholog, and the other in *P. polionotus* (the flanking sequences in *P. leucopus* could not be identified). For the two *P. eremicus* mORF2 loci we analyzed, we found both to be absent in orthologous *P. maniculatus* loci, but also absent in the orthologous site in the more closely related *P. californicus*. Therefore, there has been activity of *mys*RS and mORF2 elements specific to the *P. eremicus* lineage.

Table 1. Analysis for the presence or absence of retroelements within orthologous *Peromyscus* loci.

Element	Source	Library	Present or Absent
1) <i>mysRS</i>			
Pe <i>mysRS</i> 56082383	<i>P. eremicus</i>	<i>P. maniculatus</i>	absent
Pe <i>mysRS</i> 56082383	<i>P. eremicus</i>	<i>P. californicus</i>	absent
Pe <i>mysRS</i> 4961104	<i>P. eremicus</i>	<i>P. maniculatus</i>	absent
Pe <i>mysRS</i> 4961104	<i>P. eremicus</i>	<i>P. californicus</i>	absent
Pm <i>mysRS</i> 12-54005644	<i>P. maniculatus</i>	<i>P. leucopus</i>	absent
Pm <i>mysRS</i> 12-54005644	<i>P. maniculatus</i>	<i>P. polionotus</i>	present
Pm <i>mysRS</i> 11-9715318	<i>P. maniculatus</i>	<i>P. leucopus</i>	present
Pm <i>mysRS</i> 3-2991789	<i>P. maniculatus</i>	<i>P. leucopus</i>	present
Pm <i>mysRS</i> 10-660859	<i>P. maniculatus</i>	<i>P. californicus</i>	absent
Pm <i>mysRS</i> 10-660859	<i>P. maniculatus</i>	<i>P. leucopus</i>	absent
Pm <i>mysRS</i> 20-36151548	<i>P. maniculatus</i>	<i>P. leucopus</i>	present
2) <i>mys</i>			
Pm <i>mys</i> 23-33312091	<i>P. maniculatus</i>	<i>P. leucopus</i>	absent
Pm <i>mys</i> 1-70574111	<i>P. maniculatus</i>	<i>P. leucopus</i>	present
Pm <i>mys</i> 1-99318323	<i>P. maniculatus</i>	<i>P. leucopus</i>	present
Pm <i>mys</i> 5-92303600	<i>P. maniculatus</i>	<i>P. polionotus</i> *	present
3) mORF2			
Pm 21-28563327	<i>P. maniculatus</i>	<i>P. polionotus</i> *	present
Pm 18-786447	<i>P. maniculatus</i>	<i>P. leucopus</i>	present
Pe 32375871	<i>P. eremicus</i>	<i>P. maniculatus</i>	absent
Pe 32375871	<i>P. eremicus</i>	<i>P. californicus</i>	absent
Pe 58262094	<i>P. eremicus</i>	<i>P. maniculatus</i>	absent
mORF1			
Pm 7-10377608	<i>P. maniculatus</i>	<i>P. leucopus</i>	present
Pm 6-20227267	<i>P. maniculatus</i>	<i>P. leucopus</i>	present**
Pm 6-20227267	<i>P. maniculatus</i>	<i>P. californicus</i>	absent

*Sequence from *P. leucopus* assembly not attainable. **Contains mORF1 plus additional insert.

We additionally assessed orthologous mORF1-containing loci, albeit more difficult to analyze based on the large size. Using flanking sequences of the near full-length Pm 7-10377608 element as well as the full-length Pm 6-20227267 element-containing loci, we identified these mORF1 elements within the orthologous loci of *P. leucopus* loci. However, of interest was the finding that the flanking sequences for Pm 6-20227267 were separated by 12,015 nucleotides in *P. leucopus*, the variation resulting from an insertion of about 3840 bp. We queried the NCBI genetic database with the 3840 bp insert sequence with the top match displaying 95% coverage and 97.25% identity to a *P. maniculatus* solute carrier family 2 member 12 (Slc2a12) transcript variant (accession no. XM_042262210.1). Further matches of the insertion corresponded to the same gene in additional cricetid rodents followed by lesser matches to other rodent families. No matches beyond a few nucleotides were identified via a Repbase search. Additionally, the Pm 6-20227267 element was lacking in *P. californicus* orthologous locus supporting mORF1 activity in the genus *Peromyscus*.

Discussion

The overarching objective of this investigation was to further explore the origins, expansion, and evolution of the novel *mys* family and potential subfamilies of LTR-retroelements of the understudied Cricetidae family of rodents (Wichman et al. 1985), as the *P. leucopus mys*-1 element has been the only isolated full-length member with two nearly intact ORFs. Additionally, evidence for recent activity of *mys* elements has been illustrated in the genus *Peromyscus* (Lee et al. 1996; Sawby and Wichman 1997), and seemingly to be highly active in the lineage leading to the *P. maniculatus*/*P. leucopus* clade based on the rapid generation of different subfamilies (Lee et al. 1996) with continued activity supported by the presence or absence of elements in orthologous loci of individual *P. leucopus* mice (Sawby and Wichman 1997). We therefore investigated *mys* elements in more detail, which yielded several surprising results. These include (1) that there are only a few full-length copies of *mys* elements, and these elements are limited to members of the Neotominae subfamily of cricetid rodents; (2) the identification of a family of elements related to, but distinct from *mys*, that we refer to as *mysRS*, has amplified to a much greater extent than *mys*, as well as ascertaining full-length *mysRS* elements; (3) the identification of a novel full-length LTR-retroelement of over 8 kbp in length, with only the ORF1 sequence having high sequence similarity to that of *mys* (and less to *mysRS*), but does not harbor any ORF2-related sequences, and therefore referred to as mORF1; (4) the identification of a novel full-length LTR-retroelement containing ORF2 sequences, but not ORF1, that we refer to as mORF2, which exhibits greater similarity to the *mysRS* ORF2 sequence than that of *mys*, but appears related to both elements; (5) determined at least three of the four LTR-retroelement families, *mys*, *mysRS*, and mORF1, appear to have all originated in a common ancestor to the Neotominae subfamily of cricetid rodents, with mORF2 seemingly limited to the *Peromyscus* genus and potentially derived from a *mysRS* element; and (6) the previously high estimates of *mys* were likely the result of cross-hybridization to partial and full-length mORF1 and mORF2 sequences, plus the ORF2 region of *mysRS* elements. The lack of cross-hybridization in the quantitative dot blots between the ORF1 segments of *mysRS* and *mys* elements, along with the rarity of full-length

mys elements in genomes of Neotominae rodents, reflect a higher copy number estimate of *mys*RS elements.

We have presented detailed analyses that advance our knowledge of genomics in an understudied group of rodents, leading to the finding of four LTR-retroelement families limited to the Neotominae subfamily of cricetid rodents, with one seemingly more limited to a single genus within the subfamily. The *mys* and *mys*RS elements are distinct in sequence albeit structurally similar. Although the evolutionary timing of their origins is evident, the source, which would differ between *mys* and *mys*RS based on distinct sequence differences particularly within the ORF1s and LTRs, is not apparent. Additionally, our data are consistent with existence of two distinct subfamilies of *mys*RS in *O. torridus* based primarily on distinct clades of LTRs, as well as presence or absence of a small ERV2-n1_Mio_I insertion within ORF2, which is absent from the *P. leucopus mys* queried sequences. It has previously been reported that *mys*-1 contains structural similarities to endogenous retroviruses such as the LTRs, a pyrimidine-rich stretch of sequences, a tRNA-binding site, and some amino acid similarity to conserved regions of reverse transcriptase proteins (Wichman et al. 1985). However, the screening of nucleotide and protein databases, as well as the repetitive element database, Repbase (Kohany et al. 2006), yields back to *mys* with no significant matches to other known LTR-retroelements. It is possible that the origin and the ability for propagation of individual *mys* and *mys*RS elements occurred by recruitment of effective LTRs via mechanisms previously proposed for ERVs and retroviruses such as gene conversion (Kijima and Innan 2010) and recombination (Hayward 2017). This would be consistent with the proposed origin of porcine endogenous retroviruses (PERVs) suggested to occur by recombination following the horizontal transmission of endogenous retroviral sequences from murid rodents (Denner, 2021).

The origin of the 8,000 nucleotide mORF1 LTR-retroelement family apparently occurred at roughly the same period as both the *mys* and *mys*RS families, and likely the result of horizontal transfer, of which in addition to PERVs, there are several known examples (Platt II et al. 2018). In addition to PERVs, the mORF1 GenBank screen seemingly supports the integration of a *Pseudomonas* retroviral sequence that incorporated the *mys* ORF1 sequence in the reverse orientation. Therefore, the ORF would be downstream of the *pol* gene when viewed in the reverse orientation. A precedence for this has been observed in plants (Vicent and Casacuberta 2020) and referred to as an antisense additional open reading frame (aORF). Since the mORF2 element exhibits a more recent ancestry than the other three elements, as it is seemingly limited to *Peromyscus* species, then it likely originated from a recombination event in which ORF1 was deleted. This potentially occurred in a *mys*RS element based on sequence similarity and the greater number of full-length *mys*RS elements than *mys*.

All four LTR-retroelement families have been recently active. This is supported by our results that include their origin in an ancestor of the Neotominae family and expansion within this limited group of rodents, along with concerted evolution, as well as the presence and absence of elements in orthologous loci of *Peromyscus*. The molecular phylogenies of full-length *mys* and *mys*RS elements as well as their components demonstrate distinct clades between *O. torridus* and *Peromyscus* elements, or subfamilies of elements, along with *P. eremicus mys*RS elements forming a distinct clade from *P. maniculatus*. The

molecular phylogeny of mORF1 yielded a distinct clade of *O. torridus* versus *P. maniculatus*. Additionally, with support of strong bootstrap values, distinct clades were observed for mORF2 elements among *P. maniculatus*, *P. eremicus*, and *P. aztecus*. The demonstration of concerted evolution is consistent with activity of a limited number of active retrotranspositionally-competent source genes (Deininger et al. 1992). Recent activity is also supported by the lack of accumulated random mutations over time, which would have been observed in molecular phylogenies that consisted of older elements. The greater intra-specific sequence similarities versus that of inter-specific elements is consistent with recent activity.

Our analysis of *Peromyscus* LTR-retroelement containing loci that are absent within orthologs of other *Peromyscus* species in sister taxa is also consistent with recent activity of the four LTR-retroelement families. Additionally, a presence-absence variant of a *mys* element (*mys*-6) was identified among *P. leucopus* individuals, with a correlation to northern versus southern populations (Sawby and Wichman, 1997). This adds to the pre-existing list of active retroelement families in Neotominae rodents, particularly within the *Peromyscus* genus. For example, non-autonomous SINE retrotransposons have also apparently been active in cricetid rodents. Notably, in a study by Veniaminova et al. (2007) the lone cricetid representative assessed, *Microtus socialis*, had the most abundant number of B1 elements (roughly 1.2 million) among representatives of 21 families of analyzed rodents. Additionally, among various subfamilies of cricetid rodents, members of Neotominae (e.g. *Reithrodontomys* and *Peromyscus*) and Arvicolinae (e.g. *Microtus*) appear to be the only groups to maintain both L1 and B1 activity (Rinehart et al. 2005; Yang et al. 2019). Although ID elements have shown sporadic levels of amplification among various rodents, both *P. californicus* and *P. maniculatus* have relatively large amounts of genomic copies (Kass et al. 1996). The B2 SINE family has dramatically amplified in genomes of the Muroidea superfamily of rodents (Kass et al. 1997), which includes Muridae and Cricetidae. It is apparent from these studies and this investigation that *Peromyscus* contains several active retrotransposon families. Additionally, we have identified recent SINE integrations in *Peromyscus* species (manuscript in preparation).

Genomes with several active families of autonomous elements have the potential to increase the activity of non-autonomous elements, as SINEs have been shown to be dependent on the L1 (non-LTR) element proteins for their propagation (Dewannieux and Heidmann 2005; Garcia-Perez et al. 2007). Additionally, the LTR-retrotransposon machinery of the MusD element has been demonstrated to be utilized by the ETn transposable element in *Mus* for its propagation (Ribet et al. 2004). This activity leads to insertional mutagens that can impact genome function and phenotype (reviewed in Feschotte and Gilbert 2012). Genetic impacts include yielding genomic rearrangements such as chromosomal translocations, cause the disruption of gene regulation by providing both new promoter sequences (within the LTRs) and new epigenetic markings, as well as by providing enhancer sequences, as well as by providing new polyadenylation signals or binding sites for transcription factors. There is potential then for these insertions to contribute to reproductive isolation and speciation (reviewed in Kass 2001d hne et al. 2008). With the correlation between the *Peromyscus* genus containing over 50 species and several subspecies, particularly of the wide-ranging *P. maniculatus* (Carelton 1989) with the presence of several active autonomous and non-autonomous retroelements, including the Neotominae- and *Peromyscus*-specific

LTR-retroelement families characterized in this study, we propose a plausible model of an association between retroelement activity and the generation of *Peromyscus* species. This is consistent with the correlation between the relatively large number of species and subspecies within the genus *Mus* (<https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=10088>). LTR-retrotransposons comprise a sizeable proportion of the genome of *M. musculus* (approximately 10%) contributing to about 10% of spontaneous mutations, as well as having an estimated 3,000 competent L1 retrotransposons (reviewed in Richardson et al. 2015). These factors lend support to our proposed model. In contrast, there is a lack of evidence for the current proliferation LTR-retrotransposons in the human genome (Beck et al. 2011), with only an estimated 80–100 competent L1 elements (Brouha et al. 2003). An additional correlation involves an association of numerous species of voles of the genus *Microtus* (<https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Tree&id=10053&lvl=3&lin=f&keep=1&srchmode=1&unlock>) with a highly active, likely non-autonomous unique retrotransposon, that also demonstrated no relevant sequence homology to sequences in the NCBI database and is limited to the *Microtus* genus (Neitzel et al. 2002). It is therefore conceivable that the genomic impact of the families of LTR-retroelements originating in the ancestor of the Neotominae subfamily has contributed to the large number of *Peromyscus* species and has the potential to continue in this role.

Declarations

Acknowledgments:

Support for this research includes an Eastern Michigan University Sabbatical Leave Award and Provost Research Support Award to D.H.K.; an EMU Biology Department Meta Hellwig and Don Brown Research Fellowship and EMU Honors Undergraduate Research Fellowship to A.S.; EMU Symposium Undergraduate Research Fellowship (SURF) to M.S.; EMU McNair Scholarship to S.B.; Yongsheng Bai for assistance in sequence database submissions and Kamal Abuarquab for technical support.

Conflict of Interest: The authors declare no conflict of interest.

Ethical Approval

N/A

Authors' Contributions

D.H.K. designed the experiment, generated data, analyzed the genomic data from the DNA libraries and databases, and wrote the main manuscript text as well as created the figures. S.B., A.S., M.S., D.S., and M.C. each contributed important data to the work and contributed to the writing of the manuscript. All authors reviewed and approved the manuscript.

Funding

Support for this research includes an Eastern Michigan University Sabbatical Leave Award and Provost Research Support Award to D.H.K.; an EMU Biology Department Meta Hellwig and Don Brown Research Fellowship and EMU Honors Undergraduate Research Fellowship to A.S.; EMU Symposium Undergraduate Research Fellowship (SURF) to M.S.; EMU McNair Scholarship to S.B.

Competing Interests

All authors have no relevant financial or non-financial interests to disclose.

Availability of Data

DNA sequences generated in this study were submitted to GenBank. Accession numbers for sequences of cloned *mys* intra-ORF1 PCR products from *P. maniculatus* and *N. alstoni* are ON778008-ON778016. Accession numbers for sequences of cloned intra-mysRS PCR products from *N. alstoni*, *P. leucopus*, *R. fulvescens*, and *O. leucogaster* are ON677500-ON766521. *P. eremicus* and *P. aztecus* partial DNA libraries have accession number PRJNA847558.

References

1. Beck CR, Garcia-Perez JL, Badge RM, Moran JV (2011) LINE-1 elements in structural variation and disease. *Annu Rev Genomics Hum Genet* 12:187–215. <https://doi.org/10.1146/annurev-genom-082509-141802>
2. Brouha B, Schustak J, Badge RM, et al (2003) Hot L1s account for the bulk of retrotransposition in the human population. *Proc Natl Acad Sci U S A* 100:5280–5285. <https://doi.org/10.1073/pnas.0831042100>
3. Cantrell MA, Carstens BC, Wichman HA (2009) X chromosome inactivation and Xist evolution in a rodent lacking LINE-1 activity. *PLoS One* 4:e6252. <https://doi.org/10.1371/journal.pone.0006252>
4. Cantrell MA, Ederer MM, Erickson IK, et al (2005) MysTR: an endogenous retrovirus family in mammals that is undergoing recent amplifications to unprecedented copy numbers. *J Virol* 79:14698–14707. <https://doi.org/10.1128/JVI.79.23.14698-14707.2005>
5. Cantrell MA, Filanoski BJ, Ingermann AR, et al (2001) An ancient retrovirus-like element contains hot spots for SINE insertion. *Genetics* 158:769–777. <https://doi.org/10.1093/genetics/158.2.769>
6. Carleton MD (1989) Systematics and Evolution. In: Kirkland Jr GL and Layne JN (eds) *Advances in the Study of Peromyscus (Rodentia)*. Texas Tech University Press, Lubbock, Texas, pp 7-142
7. Casavant NC, Sherman AN, Wichman HA (1996) Two persistent LINE-1 lineages in *Peromyscus* have unequal rates of evolution. *Genetics* 142:1289–1298. <https://doi.org/10.1093/genetics/142.4.1289>
8. Chow JC, Ciaudo C, Fazzari MJ, et al (2010) LINE-1 activity in facultative heterochromatin formation during X chromosome inactivation. *Cell* 141:956–969. <https://doi.org/10.1016/j.cell.2010.04.042>
9. Deininger PL, Batzer MA, Hutchison CA, Edgell MH (1992) Master genes in mammalian repetitive DNA amplification. *Trends Genet* 8:307–311. [https://doi.org/10.1016/0168-9525\(92\)90262-3](https://doi.org/10.1016/0168-9525(92)90262-3)

10. Denner J (2021) The origin of porcine endogenous retroviruses (PERVs). *Arch Virol* 166:1007–1013. <https://doi.org/10.1007/s00705-020-04925-8>
11. Dewannieux M, Heidmann T (2005) L1-mediated retrotransposition of murine B1 and B2 SINEs recapitulated in cultured cells. *J Mol Biol* 349:241–247. <https://doi.org/10.1016/j.jmb.2005.03.068>
12. Ding W, Lin L, Chen B, Dai J (2006) L1 elements, processed pseudogenes and retrogenes in mammalian genomes. *IUBMB Life* 58:677–685. <https://doi.org/10.1080/15216540601034856>
13. Elbarbary RA, Lucas BA, Maquat LE (2016) Retrotransposons as regulators of gene expression. *Science* 351:aac7247. <https://doi.org/10.1126/science.aac7247>
14. Feschotte C, Gilbert C (2012) Endogenous viruses: insights into viral evolution and impact on host biology. *Nat Rev Genet* 13:283–296. <https://doi.org/10.1038/nrg3199>
15. Garcia-Perez JL, Doucet AJ, Bucheton A, et al (2007) Distinct mechanisms for trans-mediated mobilization of cellular RNAs by the LINE-1 reverse transcriptase. *Genome Res* 17:602–611. <https://doi.org/10.1101/gr.5870107>
16. Gifford RJ, Blomberg J, Coffin JM, et al (2018) Nomenclature for endogenous retrovirus (ERV) loci. *Retrovirology* 15:59. <https://doi.org/10.1186/s12977-018-0442-1>
17. Hayward A (2017) Origin of the retroviruses: when, where, and how? *Curr Opin Virol* 25:23–27. <https://doi.org/10.1016/j.coviro.2017.06.006>
18. Holloway JR, Williams ZH, Freeman MM, et al (2019) Gorillas have been infected with the HERV-K (HML-2) endogenous retrovirus much more recently than humans and chimpanzees. *Proc Natl Acad Sci U S A* 116:1337–1346. <https://doi.org/10.1073/pnas.1814203116>
19. Janecek LL, Longmire JL, Wichman HA, Baker RJ (1993) Genome organization of repetitive elements in the rodent, *Peromyscus leucopus*. *Mammalian Genome* 4:374–381. <https://doi.org/10.1007/BF00360588>
20. Jansa SA, Giarla TC, Lim BK (2009) The Phylogenetic Position of the Rodent Genus *Typhlomys* and the Geographic Origin of *Muroidea*. *Journal of Mammalogy* 90:1083–1094. <https://doi.org/10.1644/08-MAMM-A-318.1>
21. Jurka J (2000) Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet* 16:418–420. [https://doi.org/10.1016/s0168-9525\(00\)02093-x](https://doi.org/10.1016/s0168-9525(00)02093-x)
22. Kass DH, Berger FG, Dawson WD (1992) The evolution of coexisting highly divergent LINE-1 subfamilies within the rodent genus *Peromyscus*. *J Mol Evol* 35:472–485. <https://doi.org/10.1007/BF00160208>
23. Kass DH, Jamison N (2007) Identification of an active ID-like group of SINEs in the mouse. *Genomics* 90:416–420. <https://doi.org/10.1016/j.ygeno.2007.05.007>
24. Kass DH, Kim J, Deininger PL (1996) Sporadic amplification of ID elements in rodents. *Journal of Molecular Evolution* 42:7–14. <https://doi.org/10.1007/BF00163205>
25. Kass DH, Kim J, Rao A, Deininger PL (1997) Evolution of B2 repeats: the muroid explosion. *Genetica* 99:1–13. <https://doi.org/10.1007/BF02259494>

26. Kass DH, Raynor ME, Williams TM (2000) Evolutionary history of B1 retroposons in the genus *Mus*. *J Mol Evol* 51:256–264. <https://doi.org/10.1007/s002390010087>
27. Kass H (2001) Impact of SINEs and LINEs on the Mammalian Genome. *CG* 2:199–219. <https://doi.org/10.2174/1389202013350968>
28. Kazazian HH, Moran JV (2017) Mobile DNA in Health and Disease. *N Engl J Med* 377:361–370. <https://doi.org/10.1056/NEJMra1510092>
29. Kijima TE, Innan H (2010) On the estimation of the insertion time of LTR retrotransposable elements. *Mol Biol Evol* 27:896–904. <https://doi.org/10.1093/molbev/msp295>
30. Kohany O, Gentles AJ, Hankus L, Jurka J (2006) Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. *BMC Bioinformatics* 7:474. <https://doi.org/10.1186/1471-2105-7-474>
31. Lander ES, Linton LM, Birren B, et al (2001) Initial sequencing and analysis of the human genome. *Nature* 409:860–921. <https://doi.org/10.1038/35057062>
32. Lee RN, Jaskula JC, van den Bussche RA, et al (1996) Retrotransposon *Mys* was active during evolution of the *Peromyscus leucopus-maniculatus* complex. *J Mol Evol* 42:44–51. <https://doi.org/10.1007/BF00163210>
33. Leite RN, Kolokotronis S-O, Almeida FC, et al (2014) In the Wake of Invasion: Tracing the Historical Biogeography of the South American Cricetid Radiation (Rodentia, Sigmodontinae). *PLoS ONE* 9:e100687. <https://doi.org/10.1371/journal.pone.0100687>
34. Moran JV, DeBerardinis RJ, Kazazian HH (1999) Exon shuffling by L1 retrotransposition. *Science* 283:1530–1534. <https://doi.org/10.1126/science.283.5407.1530>
35. Morrish TA, Gilbert N, Myers JS, et al (2002) DNA repair mediated by endonuclease-independent LINE-1 retrotransposition. *Nat Genet* 31:159–165. <https://doi.org/10.1038/ng898>
36. Mouse Genome Sequencing Consortium, Waterston RH, Lindblad-Toh K, et al (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* 420:520–562. <https://doi.org/10.1038/nature01262>
37. Neitzel H, Kalscheuer V, Singh AP, et al (2002) Copy and paste: the impact of a new non-L1 retroposon on the gonosomal heterochromatin of *Microtus agrestis*. *Cytogenet Genome Res* 96:179–185. <https://doi.org/10.1159/000063036>
38. Piriyaongsa J, Mariño-Ramírez L, Jordan IK (2007) Origin and evolution of human microRNAs from transposable elements. *Genetics* 176:1323–1337. <https://doi.org/10.1534/genetics.107.072553>
39. Platt RN, Vandeweghe MW, Ray DA (2018) Mammalian transposable elements and their impacts on genome evolution. *Chromosome Research* 26:25–43. <https://doi.org/10.1007/s10577-017-9570-z>
40. Ribet D, Dewannieux M, Heidmann T (2004) An active murine transposon family pair: retrotransposition of “master” *MusD* copies and *ETn* trans-mobilization. *Genome Res* 14:2261–2267. <https://doi.org/10.1101/gr.2924904>

41. Richardson SR, Doucet AJ, Kopera HC, et al (2015) The Influence of LINE-1 and SINE Retrotransposons on Mammalian Genomes. *Microbiol Spectr* 3:MDNA3-0061–2014. <https://doi.org/10.1128/microbiolspec.MDNA3-0061-2014>
42. Rinehart TA, Grahn RA, Wichman HA (2005) SINE extinction preceded LINE extinction in sigmodontine rodents: implications for retrotranspositional dynamics and mechanisms. *Cytogenet Genome Res* 110:416–425. <https://doi.org/10.1159/000084974>
43. Roy AM, Gong C, Kass DH, Deininger PL (1998) Recent B2 element insertions in the mouse genome. *DNA Seq* 8:343–348. <https://doi.org/10.3109/10425179809034078>
44. Sawby R, Wichman HA (1997) Analysis of orthologous retrovirus-like elements in the white-footed mouse, *Peromyscus leucopus*. *J Mol Evol* 44:74–80. <https://doi.org/10.1007/pl00006123>
45. Smith MF, Patton JL (1999) Phylogenetic Relationships and the Radiation of Sigmodontine Rodents in South America: Evidence from Cytochrome b. *Journal of Mammalian Evolution* 6:89–128. <https://doi.org/10.1023/A:1020668004578>
46. Steppan SJ, Adkins RM, Anderson J (2004) Phylogeny and Divergence-Date Estimates of Rapid Radiations in Muroid Rodents Based on Multiple Nuclear Genes. *Systematic Biology* 53:533–553. <https://doi.org/10.1080/10635150490468701>
47. Tamura K, Stecher G, Kumar S (2021) MEGA11: Molecular Evolutionary Genetics Analysis Version 11. *Molecular Biology and Evolution* 38:3022–3027. <https://doi.org/10.1093/molbev/msab120>
48. Veniaminova NA, Vassetzky NS, Kramerov DA (2007) B1 SINEs in different rodent families. *Genomics* 89:678–686. <https://doi.org/10.1016/j.ygeno.2007.02.007>
49. Vicient CM, Casacuberta JM (2020) Additional ORFs in Plant LTR-Retrotransposons. *Frontiers in Plant Science* 11:555. <https://doi.org/10.3389/fpls.2020.00555>
50. Wallace MR, Andersen LB, Saulino AM, et al (1991) A de novo Alu insertion results in neurofibromatosis type 1. *Nature* 353:864–866. <https://doi.org/10.1038/353864a0>
51. Wichman HA, Potter SS, Pine DS (1985) Mys, a family of mammalian transposable elements isolated by phylogenetic screening. *Nature* 317:77–81. <https://doi.org/10.1038/317077a0>
52. Williams SL, Ramírez-Pulido J, Baker RJ (1985) *Peromyscus alstoni*. *Mammalian Species* 1–4. <https://doi.org/10.2307/3503917>
53. Woods-Samuels P, Wong C, Mathias SL, et al (1989) Characterization of a nondeleterious L1 insertion in an intron of the human factor VIII gene and further evidence of open reading frames in functional L1 elements. *Genomics* 4:290–296. [https://doi.org/10.1016/0888-7543\(89\)90332-7](https://doi.org/10.1016/0888-7543(89)90332-7)
54. Yang L, Scott L, Wichman HA (2019) Tracing the history of LINE and SINE extinction in sigmodontine rodents. *Mob DNA* 10:22. <https://doi.org/10.1186/s13100-019-0164-5>

Figures

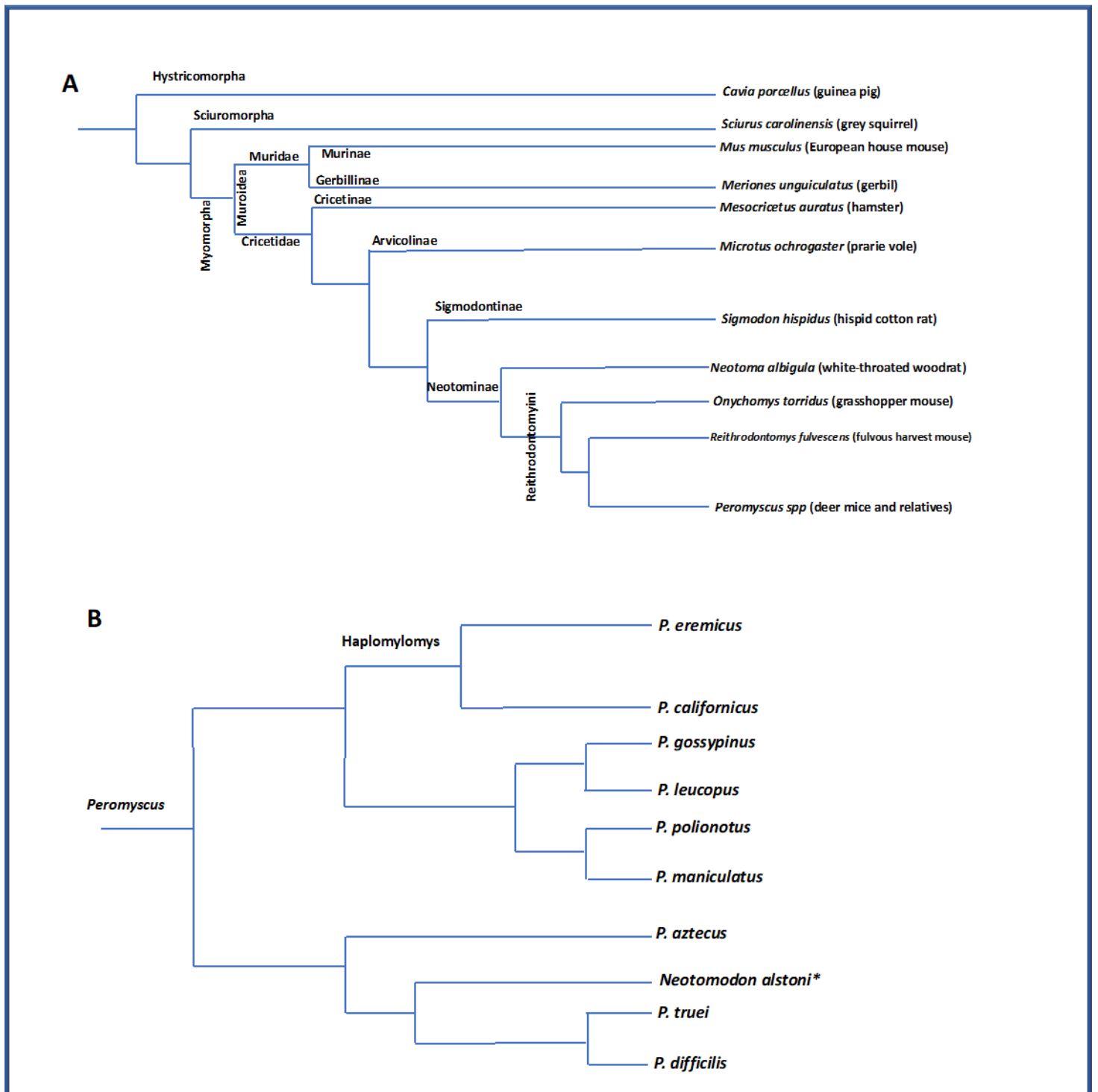


Figure 1

General phylogeny of rodents incorporated into the study based on data from Carleton (1980); Smith and Patton (1999); Steppan et al. (2004); Jansa et al. (2009); and Leite et al. (2014). (A) Representatives of the three suborders of rodents with emphasis of the Cricetidae family and subfamilies. Reithrodontomyini represents the taxonomic level of a tribe within the subfamily Neotominae. (B) Phylogenetic relationship of members of the genus *Peromyscus*, with *Haplomylomys* representing a subgenus. **Neotomodon alstoni* has also been characterized as *Peromyscus alstoni* (Williams et al. 1985).

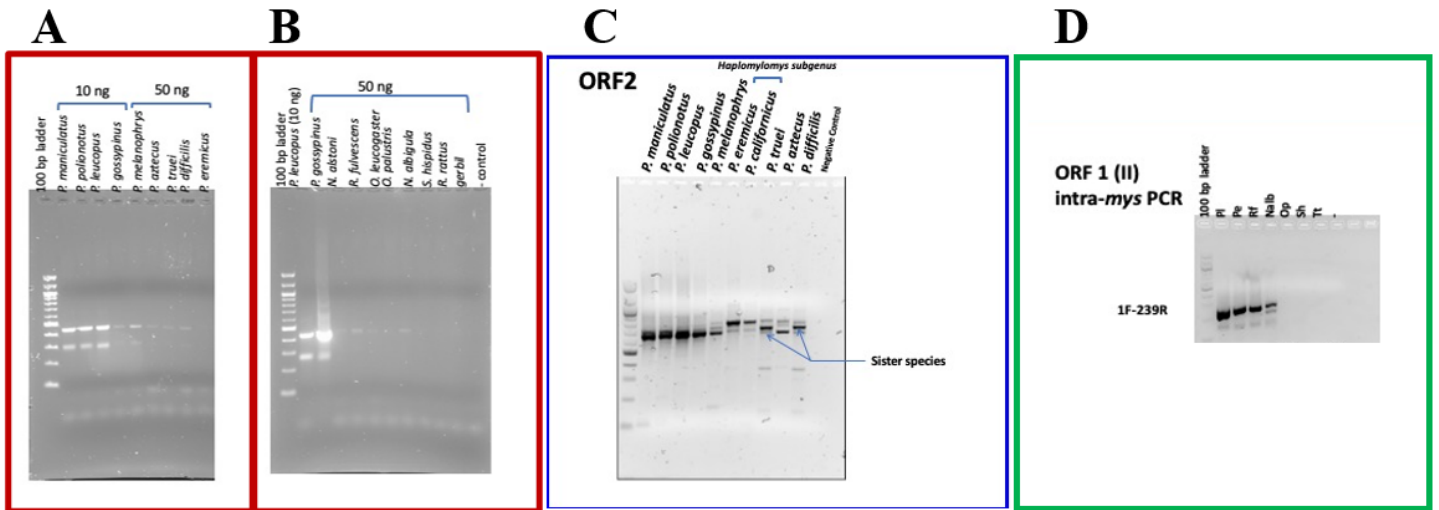


Figure 2

Agarose gel analysis of PCR amplified (a) intra-ORF1 segments from DNA of rodents in the genus *Peromyscus*; (b) intra-ORF1 segments from cricetid and murid rodents of the superfamily Muroidea. Amount of template DNA is indicated, and an alternative *P. gossypinus* DNA sample was used in B; (c) intra-ORF1 PCR utilizing primers based on highly conserved sequences (d). intra-ORF2 PCR with primers 1580 and 2210 from DNA of various species of *Peromyscus*; *Pl* = *P. leucopus*, *Pe* = *P. eremicus*, *Rf* = *Reithrodontomys fulvescens*, *Nalb* = *Neotoma albigula*, *Op* = *Oryzomys palustris*, *Sh* = *Sigmodon hispidus*, *Tt* = *Tylomys tumbalensis*, - negative control, *Pm* = *P. maniculatus*, *Pa* = *P. aztecus*, *Nals* = *Neotomodon alstoni*

```

10      20      30      40      50      60      70      80      90      100
Mys Pm 1 TGC AAGAGTAGACC --- AAGACAGCTGGAAGCCTGGACAGTCACCTAATGTTTTAGCATTGTTGGTGCATTCAAATTTGGCTACAGGCCCTAGAGTC
Mys Pm 2 TGC AAGAGTAGACC --- AAGACAACCTGGCAGCCTGGACAGTCACCTGATGTTTTCAGCATTGTTGGTGCATTCAAATTTGGCTACAGGCCCTAGAGTA
Mys Pm 3 TGC AAAAGTAGACC --- AAGACAGCTGGCAGAAATGGACAGTCACCTAATGTTTTTCAGCATTGTCGGTGCATTCAAATTTGGCTACAGGCCCTAGAGTA
Mys Pm 4 TGC AAGAGTAGACC GACCAAGACAGCTGGCAGCCTGGACAGTCACCTAATGTTTTTCAGCATTGTTGGTGCATTCAAATTTGGCTACAGGCCCTAGAGTA
Mys Pm 5 TGC AAGAGTAGACC --- AAGACAGCTGGCAGCCTGGACAGTCACCTAATGTTTTTCAGCATTGTTGGTGCATTCAAATTTGGCTACAGGCCCTAGAGTA
Mys Neo 1 TGC AGG-GTAGACA --- GAGATAGCTGACAGCCTGGACAGTCATCCAAAGTTCCTTTGTAAAGTTGGGACATCTGTTTTGAGCTCACAGACCTAGAGTC
Mys Neo 3 TGC AAGAGTAGACC --- CAGACAGCTGGCAGCCTGGACAGTCACCTAATGTTTTTCAGCATTGTTGGTGCATTCAAATTTGGCTACAGGCCCTAGAGT-
Mys Neo 4 TGC AAGAGTAGACT --- GAGACAGCTGACAGCCTGGACAGTCACCTAATGTTTTTCAGCATTGTTGGTGCATTCAAATTTGGCTACAGGCCCTAGAGT-
Mys Neo 5 TGC AGG-GTAGACA --- GAGACAGCTGACAACTGGACAGTCATTCAAAGTTCCTTTGTAAAGTTGGGACCTGTTTTCAGCTAGCAGGCCCTAGAGTC

110     120     130     140     150     160     170     180     190     200
Mys Pm 1 TCTGACAGACCATTTT TAGAA --- GCAGGAATTCGAGAGACCATTACCCCTGT --- CTTGGCAGATTACAGTGGTCGTTTTCCCTTGTGTC-CCGC
Mys Pm 2 TCTGACAGACCATTTT CAGAA --- GCAGGATTTCTGAAAGACCATTACCCCTGT --- CTTGGCAGAGTACAGTGGTCGCTTTCCCTTGTGTC-CCGC
Mys Pm 3 TCTGACAGACCATTTT CAGAA --- GCAGGATTTCTGAAAGACCATTACCCCTGT --- CTTGGCAGAGTACAGTGGTCGCTTTCCCTTGTGTC-CCGT
Mys Pm 4 TATGACAGACCATTTT CATAA --- GCAGGAATTCGTAAGACCATTCTTACC-TGT --- CTTGGCAGAGTACAGTGGTCGCTTTCCCTTGTGTC-CCGT
Mys Pm 5 TCTGACAGACCATTTT CAGAA --- GCAGGATTTCTGAAAGACCATTACCCCTGT --- CTTGGCAGAGTACAGTGGTCGCTTTCCCTTGTGTC-CCGC
Mys Neo 1 TCTCAGTCATTTTTTTCTGTGCTGTAGAATGTCTGGCAGTTTCCTCTGGAAGCAGGAACCTGAAGBACCATTTTGTCAAGCAAAAGTTCAGTGTGCAC
Mys Neo 3 TCTGACAGACCATTTT CAGAA --- GCAGGAATTCGTAAGACCATTCTTACCCTGT --- CTTGGCAGAGTACAGTGGTCGCTTTCCCTTGTGTC-CCGC
Mys Neo 4 TCTGACAGACCATTTT CAGAA --- GCAGGAATTCGTAAGATACCATTCTTACCCTGT --- CTTGGCAGAGTACAGTGGTCGCTTTCTTGTGTG-CCAC
Mys Neo 5 TCTCAGTCATTTTTTTCTGTGCTGTAGAATGTCTGGCAGTTTCCTCTGGAAGCAGGAACCTGAAGBACCATTTTGTCAAGCAA-GTTCAGTGTGCAC

210     220     230     240     250     260     270     280     290     300
Mys Pm 1 TTGTCCAGAAAGGACAGCATTATATTTGTA CTGTCA -GCCATCAAGGC -AAGGGCAGTTCCTTTGCCCAGTAGGCCATTTTGTGCCAAG--AAGACAAAC
Mys Pm 2 TTGTCCAGAAAGGACAGCATTGCATTTGTA CTGTCA -GCCATCAAGGC -AAGGGCAGTTCCTTTGCCCAGTAGGCCATTTTGTGCCAAA--AAGACAAAC
Mys Pm 3 TTGTCCAGAAAGGACAGGATGATATTCGTACTGTCA -GCCATCAAGGC -AAGGGCAGTTCCTTTGCCCAGTAGGTCATTTTGTGCCAAG--AAGACAAAC
Mys Pm 4 GTGACCCAGAAAGGTATAGCATTATATTCGTACTGTCA -GCCGTCAAGGC -AAGGGCAGTTCCTTTGCCCAGTAGGCCATTTTGTGCCAAG--AAGACAAAC
Mys Pm 5 TTGTCCAGAAAGGACAGCATTGCATTTGTA CTGTCA -GCCGTCAAGGC -AAGGGCAGTTCCTTTGCCCAGTAGGCCATTTTGTGCCAAA--AAGACAAAC
Mys Neo 1 CT-TCCATATGGGTCCGTACAT--GTTCACTGTCAAGCAGTCCAGGC -AAGAACAGTTCCTTTGCCCAATAGCCTATTTT-TGTTAAGGTGAAGATAAAC
Mys Neo 3 TTGTCCAGAAAGGACACCATTT---GTTCACTGTCA -GCAGTAGAGGC -AAGGGCAGTTCCTTTGCCAGTAGGCCATCTTGTGCCAAG--AAGACAAAC
Mys Neo 4 TTGTCCAGAAAGGACAGCATTGGATTCTGTACTGTCA -GCAGTTGAGGCCAAGGGCAGTTCCTTTGCCCAGTAGGCCATCTTGTGCCAAG--AAGACAAAC
Mys Neo 5 CT-TCCATATGGGTCCGTACAT--GTTCACTGTCAAGCAGTCCAGGC -AAGAACAGTTCCTTTGCCCAATAGCCTACTTT-TGCCAAGGTGAAGATAAAC

310     320     330     340     350     360     370     380     390     400
Mys Pm 1 TTCCCAATGGAAATGTCTTAGAAGCCCAACATTCTCTCGGGATCAAAT-TGGTGCAGCCAGGAGCAATTTGTGTCTCAC-ATCAA-----CAGAATTC
Mys Pm 2 TTCCAAATGGAAATGTCTTAGAAGCCCAACATTCTCTCGGGATCAA-T-TGGTGCAGCCAGGAGCAATTTGTGCCTCAC-GTCAA-----CAGAATTC
Mys Pm 3 TTCCAAATGGAAATGTCTTAGAAGCCCAACATTCTCTCCAGATCAAAT-TGGTGCAGCCAGGAGCAATTTGTATCTCAT-ATCAA-----CAGAATTC
Mys Pm 4 TTCCAAATGGAAATGTCTTAGAAGCCCAACATTCTCTCGGGATCAAAT-TGGTGCAGCCAGGAGCAATTTGTGTCTCAT-GTC-----C-----
Mys Pm 5 TTCCAAATGGAAATGTCTTAGAAGCCCAACATTCTCTCGGGATCAA-T-TGGTGCAGTCCAGGAGCAATTTGTGTCTCAT-GTCAA-----C-GAATTTT
Mys Neo 1 -TCCATATGGAG-TGTCCTTCGATGCCCATGCTCCTCT-GAAGTAAA-TCTGGTGTTCGCCAGGAGCAGACATGCTCAGCTGTCAGAACGCTCTAAATTTT
Mys Neo 3 ATCCAACTAGAAATGTCTTAGAAGCCCAACATTCTCTCGGGATCAAAT-TGGTGCAGCCAGGAGCAATTTGTGTCTCAC-ATCAA-----CAGAATTT
Mys Neo 4 ATTGAAC TAGAAATGTCTTAGAAGCCCAAGCATTCTCTCGGGATCAAAT--GGTGTGCCAGGAGCAATTTGTGTCTCAC-GTCAA-----TAGAATTC
Mys Neo 5 -TCCATATGGAG-TGTCCTTCGATGCCCATGCTCCTCT-GAAGTAAAATCCAGTGT-GCCAGGAGCAGACGTGTCTCATGTGCCAGAACGCTCTAAATTTT

410     420     430     440     450     460     470
Mys Pm 1 AAGTTATTT-AAATGCCATATTCTCTAGGTCTATGAAGTGTTTGAAGATTATCTATCTGAAATATCTG
Mys Pm 2 AAGTTATTT-AAATGCCATATTCTCTAGGTCTATGAAGTGTTTGAAGATTACCTATCTATCTGAAATATATC
Mys Pm 3 AAGTTATTT-AAATGCCATATTCTCTAGGTCTATGAAGTGTTTGAAGATTACCTATCTATCTGAAATATATC
Mys Pm 4 -----
Mys Pm 5 AAGTTATTT-AAATGCCATATTCTCTAGGTCTATGAAGTGTTTGAAGATTACCTATCTATCTGAAATATATC
Mys Neo 1 AAAACATTTTAAATGCCATATTCTCTAGGTCTTTGAAGTATTTGAAGATTACCTATCTATCTGAAATATATC
Mys Neo 3 AAGTTATTT-AAATGCCATATTCTCTAGGTCTATGAAGTGTTTGAAGATTACCTTTTATCTGAAATATATC
Mys Neo 4 AAGTTATTT-AAATGCCATATTCTCTAGGTCTATGAAGTGTTTGAAGATTACCTTTTATCTGAAATATATC
Mys Neo 5 AAAACATTTTAAATGCCATATTCTCTAGGTCTTTGAAGTATTTGAAGATTACCTATCTATCTGAAATATATC

```

Figure 3

Alignment of sequences from ORF1-generated clones from DNA of *P. maniculatus* (Pm) and *N. alstoni* (Neo). Shaded regions represent conserved sequences. Clones bracketed in red represent similar but distinct sequences to *mys*ORF1

Intra-*mysRS* PCR

P. leucopus
N. alstoni
R. fulvescens
O. leucogaster
O. palustris
T. tumbalensis
N. albigula
M. auratis
M. unguiculatus
- control

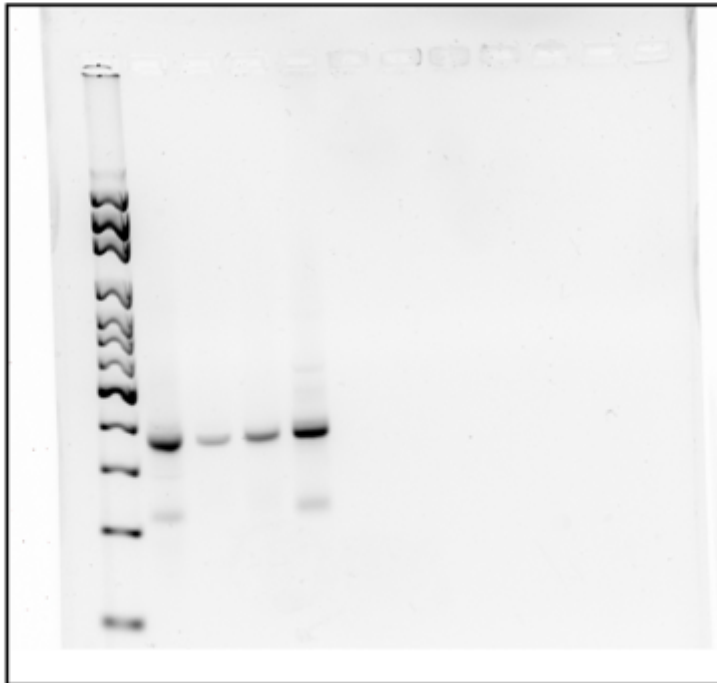


Figure 4

Amplification of intra-*mysRS* sequences among cricetid rodents (primers F60-R430)

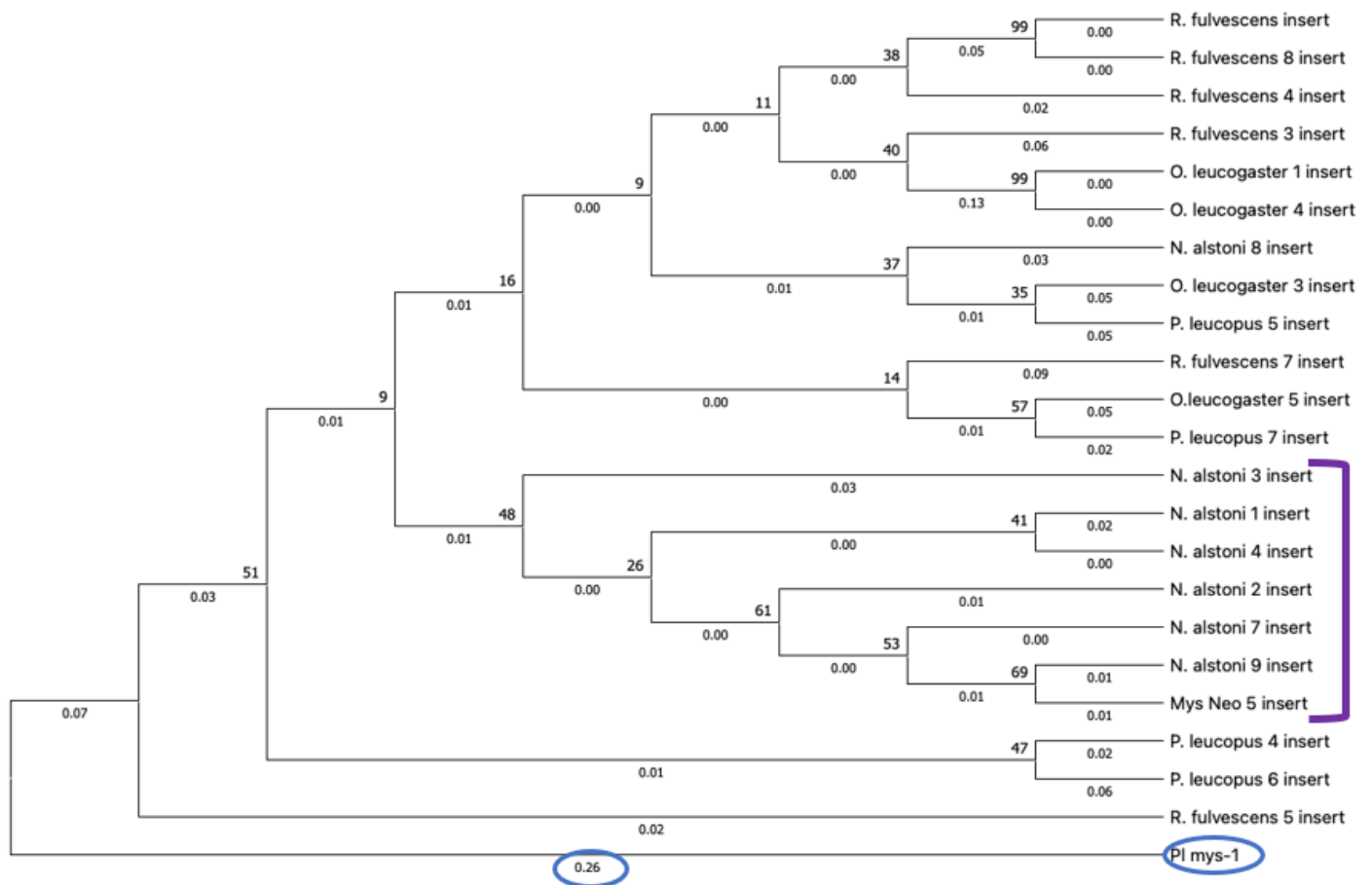


Figure 5

Molecular phylogeny of sequences from intra-*mysRS* ORF1 PCR cloned fragments aligned to the corresponding ORF1 region of *P. leucopus*. Alignment generated via maximum likelihood using the Hasegawa-Kishino-Yano (HKY+G) model. Bootstrap values based on 500 replicates and estimated pair-wise distances applied by Neighbor-Join and BioNJ algorithms. LogL = -1356.58. Clade of *N. alstoni* sequences bracketed and *mys-1* depicted as an outgroup

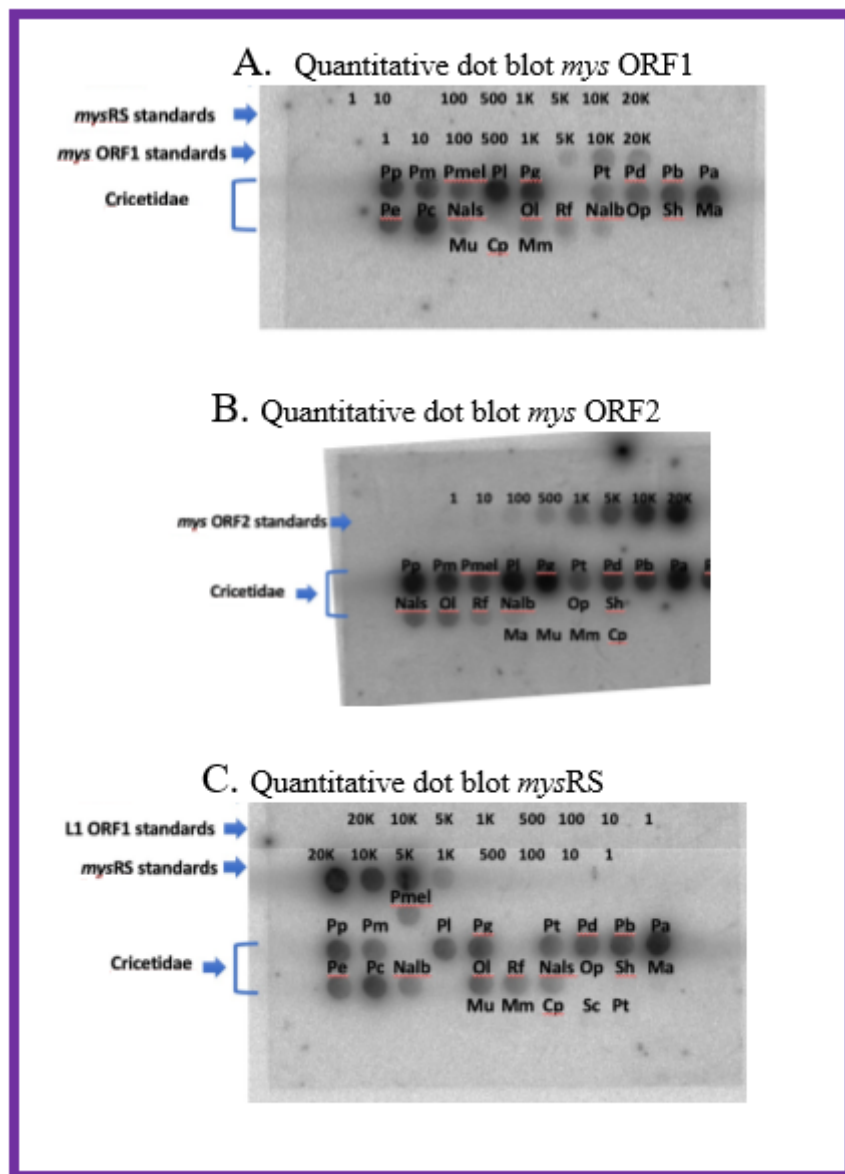


Figure 6

Quantitative dot blots of (a) *mys*-1 based on hybridization to a *P. gossypinus* ORF1 probe. Pp = *Peromyscus polionotus*; Pm = *P. maniculatus*; Pmel = *P. melanotis*; Pl = *P. leucopus*; Pg = *P. gossypinus*; Pt = *P. truei*; Pd = *P. difficilis*; Pb = *P. boylii*; Pa = *P. aztecus*; Pe = *P. eremicus*; Pc = *P. californicus*; Nals = *Neotomodon alstoni*; Ol = *Onychomys leucogaster*; Rf = *Reithrodontomys fulvescens*; Nalb = *Neotoma albigula*; Op = *Oryzomys palustris*; Sh = *Sigmodon hispidus*; Ma = *Microcricetus auratus*; Mu = *Meriones unguiculatus*; Cp = *Cavia porcellus*; Mm = *Mus musculus*; (b) *mys*-1 based on hybridization to a *P. polionotus* ORF2 probe. Pp = *Peromyscus polionotus*; Pm = *P. maniculatus*; Pmel = *P. melanotis*; Pl = *P. leucopus*; Pg = *P. gossypinus*; Pt = *P. truei*; Pd = *P. difficilis*; Pb = *P. boylii*; Pa = *P. aztecus*; Pe = *P. eremicus*; Pc = *P. californicus*; Nals = *Neotomodon alstoni*; Ol = *Onychomys leucogaster*; Rf = *Reithrodontomys fulvescens*; Nalb = *Neotoma albigula*; Op = *Oryzomys palustris*; Sh = *Sigmodon hispidus*; Ma = *Microcricetus auratus*; Mu = *Meriones unguiculatus*; Cp = *Cavia porcellus*; Mm = *Mus musculus*; (c) the *mys*RS element based on hybridization to a *P. maniculatus* *mys*RS ORF1 probe. Pp = *Peromyscus*

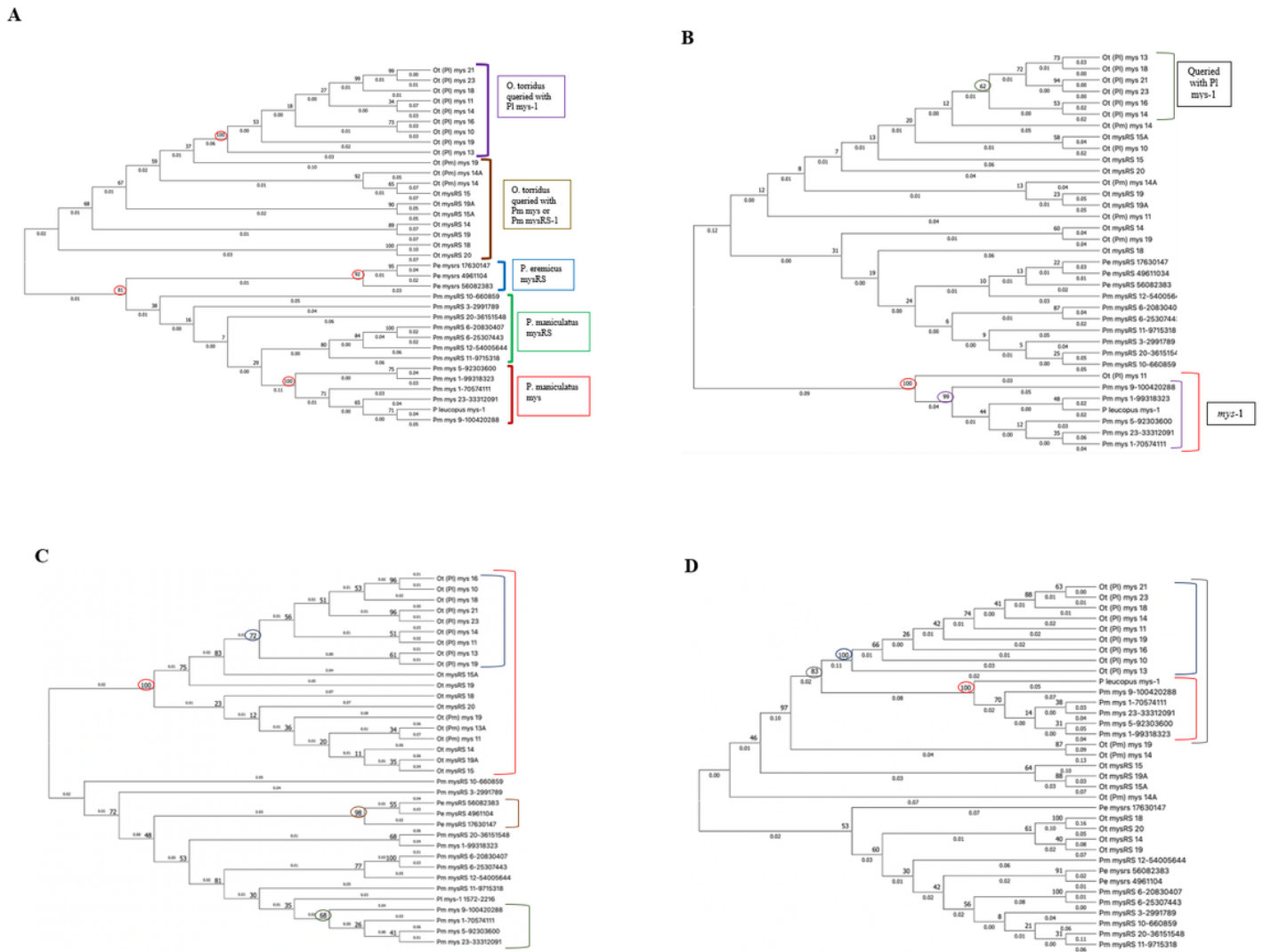


Figure 9

Molecular phylogenies of full or near full-length *mys* and *mysRS* elements and the various components of the elements. **(a)** Molecular phylogeny of full or near full length elements isolated from database screens using presumptive *mys* and *mysRS* elements. The maximum likelihood consensus tree (500 replicates) was generated using the HKY+G parameter model of substitution with estimated pairwise distances applied by NJ and BioNJ algorithms. LogL = -22303.13. Sequence designations for *O. torridus* indicate the queried sequence (*P. leucopus mys*, *P. maniculatus mys*, or *P. maniculatus mysRS*) along with the chromosome number, as each chromosome assembly of *O. torridus* has a unique GenBank accession number. If there is more than one sequence from the same chromosome, then a letter designation is provided. *P. eremicus* designations correspond to the isolate generated by GeneWiz, and the *P. maniculatus* designation corresponds to the chromosome and approximate position; **(b)** Molecular phylogeny of the ORF1 regions of full-length *mys* and *mysRS* elements. The maximum likelihood consensus tree (500 replicates) was generated from sequences aligned with Muscle and using the Tamura 3 (T92+G) parameter model of substitution with estimated pairwise distances applied by NJ and BioNJ algorithms. LogL = -3323.56; **(c)** Molecular phylogeny of the ORF2 regions of full-length *mys* and

mysRS elements (Fig. S3). The maximum likelihood consensus tree (500 replicates) was generated from sequences aligned with Muscle and using the Tamura 3 (T92+G) parameter model of substitution with estimated pairwise distances applied by NJ and BioNJ algorithms. LogL =-5049.21; **(d)** Molecular phylogeny of the 5'LTR regions of full-length *mys* and *mysRS* elements (Fig. S3). The maximum likelihood consensus tree (500 replicates) was generated from sequences aligned with Muscle and using the Tamura 3 (T92+G) parameter model of substitution with estimated pairwise distances applied by NJ and BioNJ algorithms. LogL =-3638.48. Notable monophyletic groups are indicated.

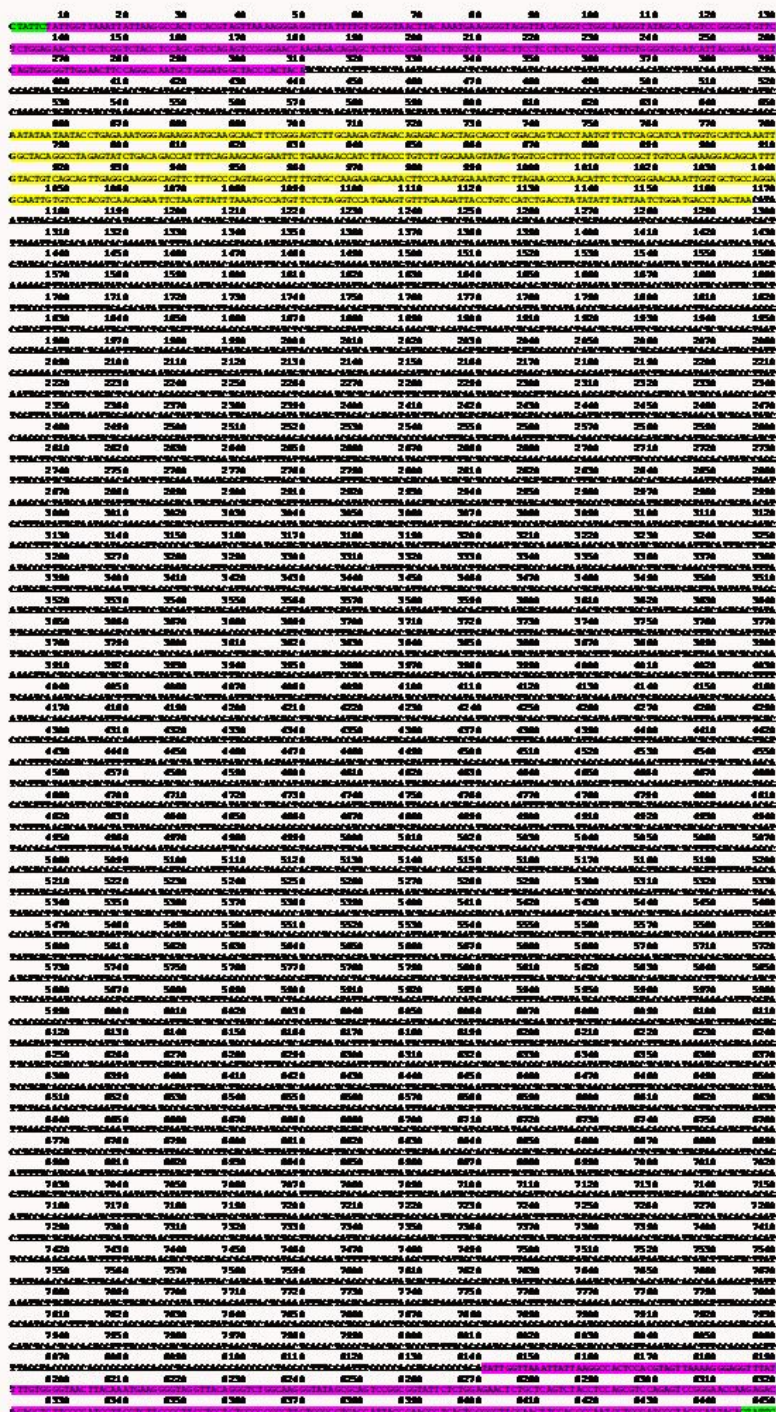


Figure 10

Full-length mORF1 from *O. torridus* (chromosome 23). Green highlighted regions correspond to the FDRs, pink highlights to the LTRs, and the yellow highlighted region to ORF1

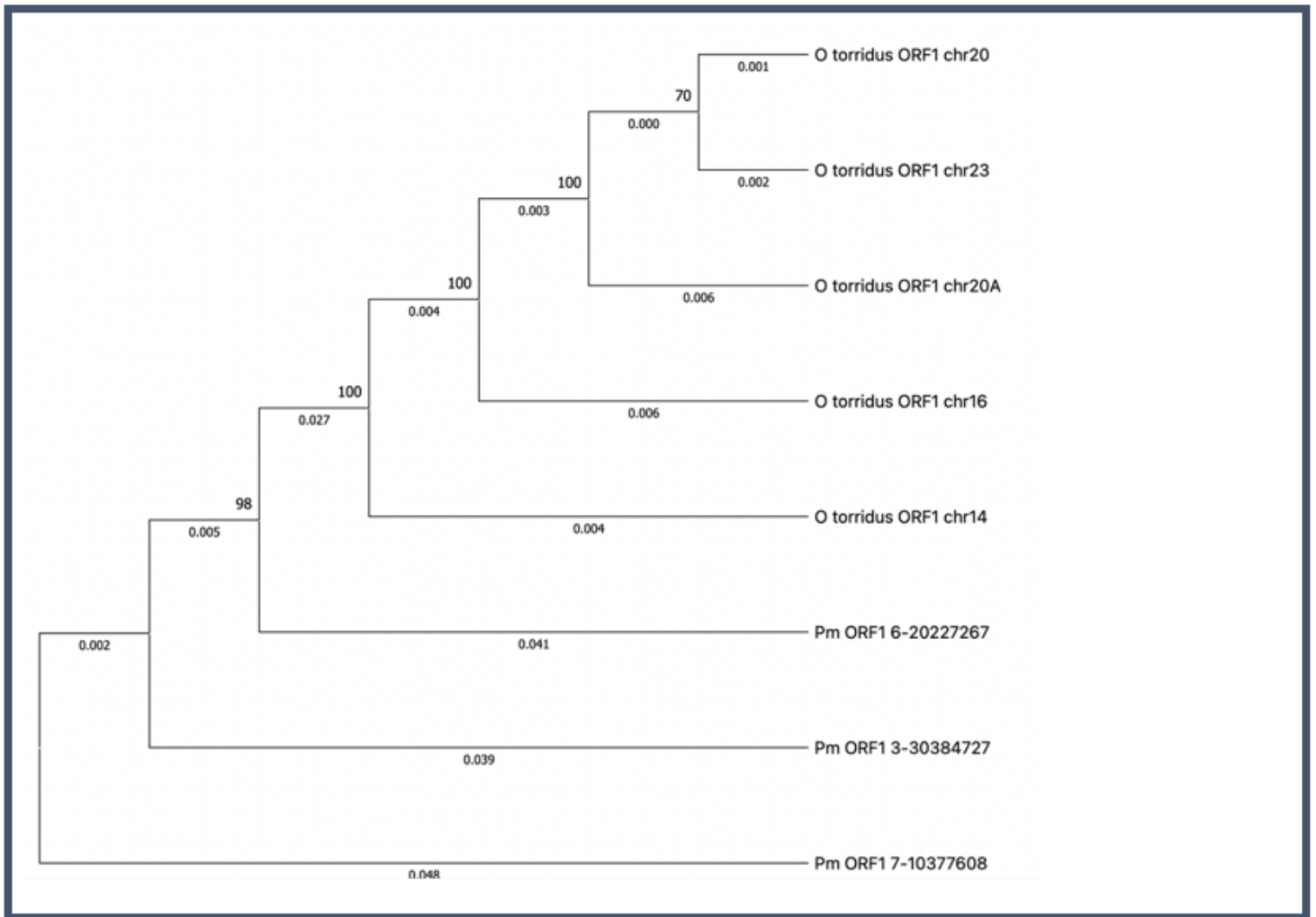


Figure 11

Molecular phylogeny of the full and near-full length mORF1 elements (from figure S5). The maximum likelihood consensus tree (500 replicates) was generated from sequences aligned with Muscle and using the HKY+G parameter model of substitution with estimated pairwise distances applied by NJ and BioNJ algorithms.

LogL =-1377.90

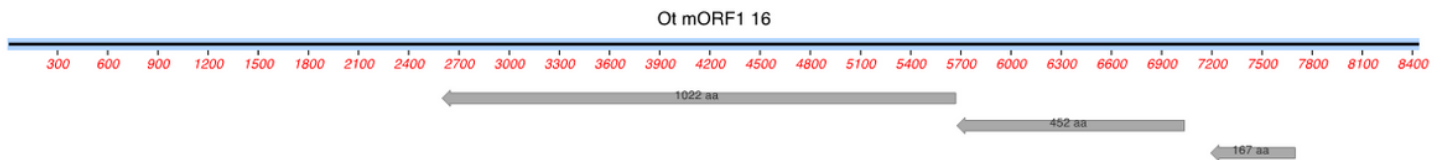


Figure 12

ORF analysis of the *O. torridus* mORF1 16 element, with a minimum of 100 amino acids in both plus and minus strands.

10	20	30	40	50	60	70	80	90	100	110	120
AGAAAT	TGTAGCTAGAGTTT	CTGTCTTGCC	CACAGTCAGG	ACAAATCTTT	GTAAACCGCC	AGTCCCACAG	CCGCTCAGA	ACCAACCAAG	TAAACACAG	AGACTTAT	ATTGCTTACAAA
130	140	150	160	170	180	190	200	210	220	230	240
CTGTATGGCC	ATGACAGGCTT	CTTGCTAACT	ATTCTTATAG	CTTAAATTAAT	CCATTCTATA	AAATCTAACC	GTGGCTTACC	GGCATTTCAC	ATTCTGCTT	GTGCTGGC	AGTGGCTGGC
250	260	270	280	290	300	310	320	330	340	350	360
AGTGACTCCT	GCCTTTCTGT	CTTTCTTTCT	CTCTCTGTT	AGTCCC	GCCTATACT	TCCTGCCT	AGCCACTGG	CCAAATCAG	TGTATTTAT	TATTGACCA	ATCAGAGCA
370	380	390	400	410	420	430	440	450	460	470	480
ATACAGAAC	ATCCCACAGCA	CTCCCTCCCC	CTTTTTC	CAAAAGGA	AGTTTAA	ACCTTAACA	AAAGTAA	AATTACAT	AATAATTG	GGGAATTG	GGCGTAGCT
490	500	510	520	530	540	550	560	570	580	590	600
TGCTGGAGG	GGGGACGCT	GTATCTTAT	GGGGACACA	AGAAAAAT	TTTAGGAT	CATGGAAT	AGTCCAT	GAGGCTG	TATTGTCT	GAGCCAG	CTGCCTTCA
610	620	630	640	650	660	670	680	690	700	710	720
TCTGGGCC	ATGGCGTC	ATCAGAGCC	CTTCAGGT	GGTCTTGG	CTGGTCAA	ACCTGAT	GTATCTTA	ATCTGGA	ACAAATCC	ATAGCCTC	TGGCTTTCT
730	740	750	760	770	780	790	800	810	820	830	840
TTTCCAA	AGCAACAT	ATCCTTAT	ATCCAAAT	TTTGAAG	TCAAGGT	ACCTTTAA	AAATACAT	TTTGGCA	TAACTCA	ACAGCTTT	TGCAATCAA
850	860	870	880	890	900	910	920	930	940	950	960
AGACAAT	CCAGATTCT	CTGTGTA	ATATTCAT	CTTTATG	TGGCTTAT	TTTTTAT	ATTAATTT	TAATCTG	CTCTTTAA	AGACTTTT	TAAAACTAT
970	980	990	1000	1010	1020	1030	1040	1050	1060	1070	1080
ACCTTTT	TGCTCTTT	CAAGCTAC	GTATATTT	TACACAC	ATTGTAA	ACTATTAC	ATCTGA	ATCTGTCT	TATTGTG	AATCTATT	GCTTTAA
1090	1100	1110	1120	1130	1140	1150	1160	1170	1180	1190	1200
CTATGGCT	GCTGGCTC	TGCCCACCT	CAGCTTCC	CAACATGG	CGGTGGT	ACATTTACC	GGCATCTC	TGGGAGCT	ATCATG	GGGTCTAT	GCTTTTAT
1210	1220	1230	1240	1250	1260	1270	1280	1290	1300	1310	1320
CTTTTTT	GTGTTTGT	ACTAGCAA	AGGCTAA	ATCTACCA	TGCAGCTT	AAATGTG	CCACTTGC	AGAGGCTC	ATTC	CCGCATA	CTGCAGGT
1330	1340	1350	1360	1370	1380	1390	1400	1410	1420	1430	1440
TCAAACG	GCAGCTG	CTGCTCAT	TTGAGAG	AGACAATT	AGGAGG	CTGTTTT	AGTCCG	TTTTAG	ATCTTTT	CTCAGG	TTTAGG
1450	1460	1470	1480	1490	1500	1510	1520	1530	1540	1550	1560
TAACTAG	AGTTTTCT	GCTTTG	CCCACAG	TCAAGCA	AAATCTT	GTCAAC	CGCCAG	TCCCAC	AGCCGCT	CAGAAC	CAACCA
1570	1580	1590	1600	1610	1620	1630	1640	1650	1660	1670	1680
CCATGGC	AGGCTTCT	TGCTAACT	ATTCTTAT	AGCTTAA	ATTAAT	CCATTTCT	ATAAACT	TAACCG	TGGCTTAC	CGGCATCT	TACATG
1690	1700	1710	1720	1730	1740	1750	1760	1770	1780	1790	1800
CTTCTGC	CTTCTG	TTCTTTCT	CTCTCTG	TAGTCCC	GCCTATA	CTTCTGC	CTAGCC	ACTGGC	CAATCA	ATGTTTAT	TTATTG
1810	1820										
GAACAT	CCCACAG	CAAGAAAT									

Figure 13

Full-length *Peromyscus* mORF2 element Pm 21-28563327. Green highlights = FDRs, blue highlights = LTRs, yellow highlights = region corresponding to *mys-1* ORF2

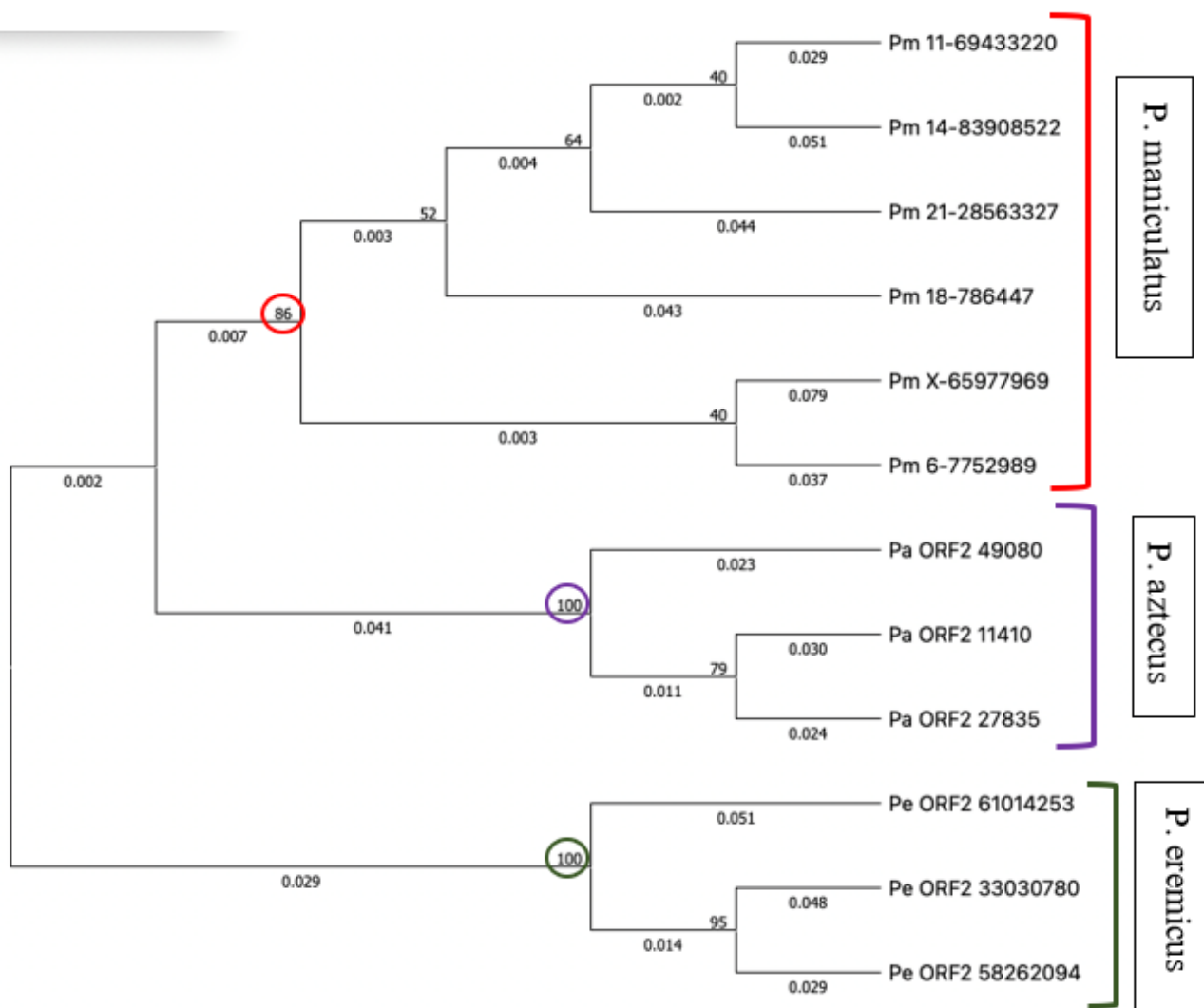


Figure 14

Molecular phylogeny of *Peromyscus* ORF2 sequences generated using the HKY-G model of maximum likelihood and applying pairwise distances using the Neighbor-Join and BioNH algorithms based on a maximum composite likelihood approach. LogL = -644.32

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [myspapersupplementalfigures93022.docx](#)