

THE PENNSYLVANIA STATE UNIVERSITY
SCHREYER HONORS COLLEGE

DEPARTMENT OF BIOLOGY

THE DIVERSITY OF ENDOGENOUS RETROVIRUSES IN WYOMING MULE DEER

STEPHANIE N. WILLIAMS
SPRING 2019

A thesis
submitted in partial fulfillment
of the requirements
for a baccalaureate degree
in Biology
with honors in Biology

Reviewed and approved* by the following:

Mary L. Poss
Professor of Biology and of Veterinary and Biomedical Sciences
Thesis Supervisor

Daniel J. Cosgrove
Professor and Holder of the Eberly Chair in Biology
Honors Adviser

* Signatures are on file in the Schreyer Honors College

Abstract

Endogenous retroviruses (ERVs) are RNA viruses that integrate, by chance, into the germ cells of a host, permitting vertical transmission.²⁹ The number of ERV integration sites within the genome can vary between individuals within a species due to unique, germ-line infections or expansion caused by retrotransposition, a process during which DNA copies of an initial ERV integrate into the host genome at different locations via an RNA transcript intermediate.²⁹ Over time, ERVs can be lost from the population, actively silenced, or remain in the genome and become fixed.^{17, 22} ERVs within mule deer (CrERVs) are insertionally polymorphic and transcriptionally active.^{9, 31} We currently have data describing the distribution and location of CrERVs within the genomes of mule deer in Montana (MT). The data suggests that, on average, mule deer within Wyoming (WY) have more integrations than MT mule deer. This study strove to explore possible explanations for the observed difference in the number of CrERV integration sites within Wyoming (WY) and Montana (MT) mule deer. During these investigations, transcript data and a small survey of CrERV unique to WY mule deer provided no evidence to support the presence of novel CrERV in WY mule deer. However, evidence to suggest expansion via retrotransposition of one CrERV family was found in the WY mule deer. Because ERVs are capable of insertional mutagenesis, such information is critical in the investigation of whether or not CrERVs could contribute to the presence and distribution of diseases that affect mule deer.^{17, 20, 22} One such disease is Chronic Wasting Disease, which is endemic in wild mule deer populations in Wyoming (WY) and Northern Colorado (CO), but absent from populations to the north and west of this area, like Montana (MT), despite the high mobility and migratory habits of mule deer.^{2, 18}

Table of Contents

LIST OF FIGURES.....	iii
Chapter I: Introduction.....	pg. 1
The Retrovirus.....	pg. 1
An Overview of the Life Cycle of a Retrovirus.....	pg. 2
The Life Cycle of a Retrovirus: Structure Retroviral RNA.....	pg. 2
The Life Cycle of a Retrovirus: Reverse Transcription.....	pg. 3
The Life Cycle of a Retrovirus: Integration of Retroviral DNA into Host Genome.....	pg. 4
Endogenous Retroviruses.....	pg. 5
The Impact of Endogenous Retroviruses on a Host.....	pg. 5
The Evolution of Endogenous Retroviruses.....	pg. 7
<i>Odocoileus hemionus</i> and Cervid ERVs.....	pg. 8
Aim I.....	pg. 10
Aim II.....	pg. 11
Chapter II: Methods.....	pg. 12
Mule Deer Sample Tissue Collection, CrERV RNA Extraction, and First-Strand cDNA Synthesis.....	pg. 12
Amplification of CrERV cDNA.....	pg. 12
Isolation, Cloning, and Sequencing of CrERV cDNA.....	pg. 14
Phylogenetic Analysis of CrERV DNA Sequences.....	pg. 14
Splice Pattern Determination.....	pg. 15
Identifying CrERV with Novel Insertions in WY Mule Deer.....	pg. 15
Chapter III: Results.....	pg. 16
Aim I.....	pg. 16
Analyzing CrERV Spliced Envelope Sequences in Wyoming Mule Deer.....	pg. 16
Comparing CrERV Spliced Envelope Sequences of Wyoming and Montana Mule Deer.....	pg. 17
Splice Pattern Variation.....	pg. 23
Diversity of Transcriptionally Active CrERV within Individual Animals.....	pg. 24
Aim II.....	pg. 25
Chapter IV: Discussion.....	pg. 26
Literature Cited.....	pg. 29
Academic Vita.....	pg. 33

LIST OF FIGURES

Figure 1: Structure of an Infectious Retrovirus.....	pg. 1
Figure 2: Reverse Transcription.....	pg. 3
Figure 3: Creation of Target Site Duplication.....	pg. 5
Figure 4: Phylogenetic Tree of 14 Known CrERV.....	pg. 9
Figure 5	
5. a) Primer Name and Sequence.....	pg. 14
5. b) Primer Diagram.....	pg. 14
Figure 6: Maximum Likelihood Tree for Unique Transcripts in Wyoming Mule Deer.....	pg. 19
Figure 7: Maximum Likelihood Tree for Unique Transcripts in Montana and Wyoming Mule Deer.....	pg. 20
Figure 8: Minimum Spanning Network for MT and WY CrERV Clustered with In1 in Figure 7.....	pg. 23
Figure 9: Minimum Spanning Network for MT and WY CrERV Clustered with In12 in Figure 7.....	pg. 24
Figure 10: Chart of the Number of WY and MT Mule Deer with Transcriptionally Active In1 and In12 CrERV.....	pg. 25
Figure 11: The Two Splice Patterns and Alternative Splice Acceptor Observed within CrERV Sequences.....	pg. 26
Figure 12: Diversity of Unique Viruses within WY Mule Deer.....	pg. 28
Figure 13: Diversity of Unique Viruses within MT mule deer.....	pg. 29
Figure 14: Phylogenetic Tree of all CrERV Major Lineages and Two Unique CrERV Sequenced from WY Mule Deer.....	pg. 30

Chapter I: Introduction

The Retrovirus

A retrovirus is a ribonucleic acid (RNA) based virus with a unique structure and lifecycle from other viruses. A retroviral particle, or virion, is comprised of an envelope, an outer membrane derived from the host membrane, and a capsid, an inner membrane derived of viral protein. Coating the envelope are viral encoded glycoproteins, which bind the virion to a host cell. Within the capsid, there are two strands of identical ribonucleic acid (RNA) and three viral enzymes: integrase, protease, and reverse transcriptase (RT).^{6, 23} The two strands of RNA are dimerized at the 5' end and the RT enzyme is in association with the RNA dimer. The reverse transcriptase is anchored to the RNA dimer and primed by a host transfer RNA (tRNA) molecule.^{6, 14, 23}

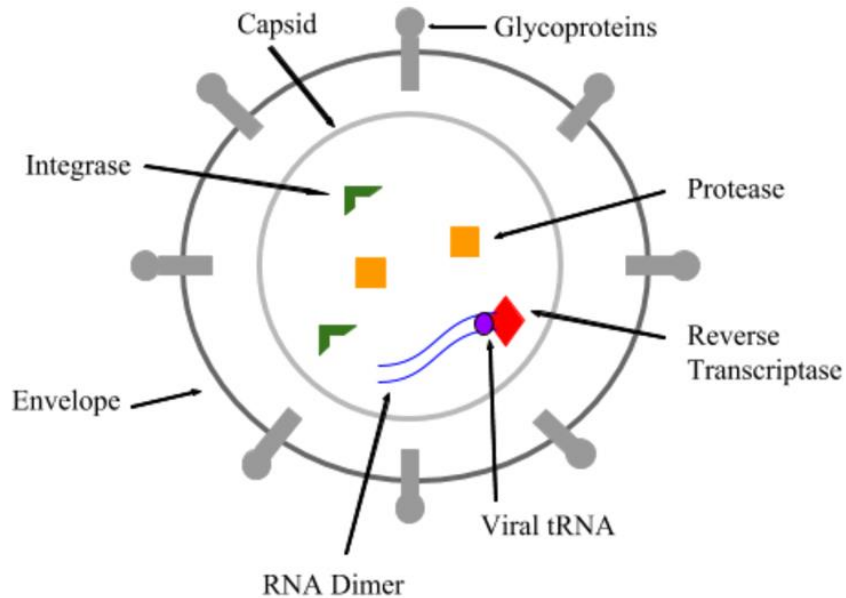


Figure 1. Structure of an Infectious Retrovirus. The capsid of the virion is surrounded by the envelope, which is coated by various glycoproteins. Contained within the capsid are the integrase, protease, and reverse transcriptase enzymes, the latter of which is in association with the dimer of two homologous strands of RNA.

An Overview of the Life Cycle of a Retrovirus

When a virion encounters a potential host cell, some of the glycoproteins on the viral envelope will bind and anchor the virion to receptor glycoproteins on the host cell. The specific glycoproteins involved in the process of virion absorption differ for each retrovirus. After the virion has bound to the host cell, the viral envelope fuses with the host plasma membrane and the capsid enters the host cell. Once the capsid is released into the host cytoplasm, it dissociates and reverse transcription of the viral RNA begins via the RT enzyme. The result of reverse transcription is double stranded viral deoxyribonucleic acid (DNA). The process of reverse transcription also results in flanking, identical long terminal repeats (LTRs) on both the 3' and 5' end of the viral genome. The double stranded viral DNA is translocated into the nucleus of its host, where it is integrated into the host genome via the integrase enzyme.²⁴ The viral RNA genome is replicated utilizing host mechanisms and machinery. Viral proteins, called *gag* proteins, drive the assembly of new viral particles, which acquire the host lipid membrane and bud from the host cell surface with the potential to infect new host cells.⁶

The Life Cycle of a Retrovirus: Structure Retroviral RNA

Each retroviral RNA genome encodes at least four genes: *gag*, *pro*, *pol*, and *env*. The *gag* region is responsible for the formation of the viral capsid and the assembly of new viral particles, as previously mentioned. The *pro* region encodes the protease enzyme and the *pol* region encodes the reverse transcriptase and integrase enzymes. The *env* region encodes the viral envelope glycoproteins, which are necessary for the entry into host cells. Each RNA genome also contains a flanking repeat (R) region and a unique region on both the 3' (U3) and 5' (U5) end of the RNA molecule. The R regions can range from 15 to 250 base pairs long while the unique

regions can vary in length from 70 to 250 base pairs. In total, retroviral genomes typically range from 7,000 base pairs to 12,000 base pairs in length (see **Figure 2** below).^{6, 27}

The Life Cycle of a Retrovirus: Reverse Transcription

Once the viral envelope has fused with the host membrane and the viral capsid has entered into the host cell, the contents of the viral capsid are released into the cytoplasm. The association of the RNA dimer and RT remains intact in the cytoplasm and uses the primer tRNA to initiate reverse transcription. Reverse transcription is the process of converting viral RNA into double stranded viral DNA. During reverse transcription, all genes are preserved and two flanking, identical long terminal repeats (LTR) are generated (**Figure 2**).^{6, 17} Two transcripts of the viral DNA can be created: a full-length transcript of the viral genome from which gag and enzymes are produced and a spliced transcript that only encodes the *env* gene.^{9, 30} The latter, also called the spliced *env* transcript, is a continuous transcript from the transcription start site through to the splice donor (SD) and then from the splice acceptor (SA) through to the transcription end site. The spliced *env* transcript makes up a larger portion of the total viral RNA transcripts within the host cell.⁶

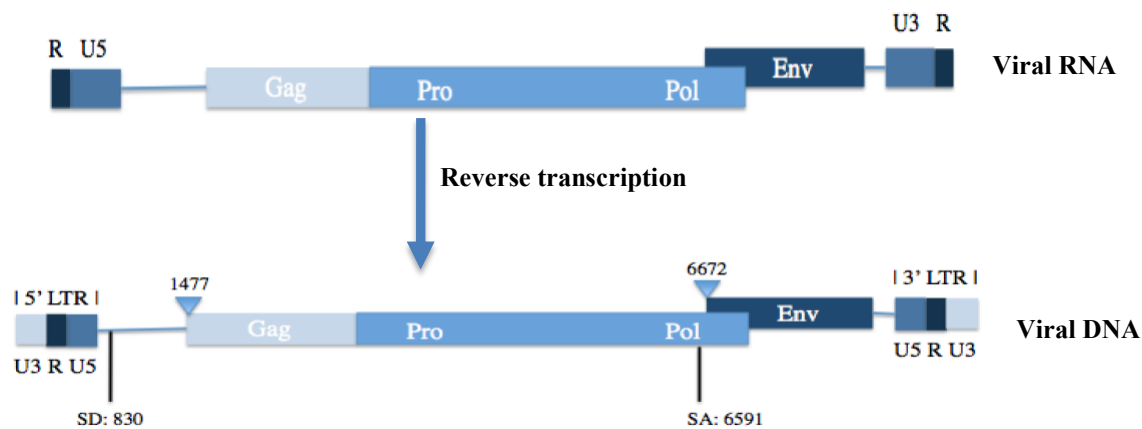


Figure 2. Reverse Transcription. During reverse transcription, viral RNA is converted into double stranded viral DNA via the viral reverse transcriptase enzyme. The process of reverse transcription preserves all viral genes and generates identical flanking long terminal repeats (LTRs). SD = splice donor. SA = splice acceptor. Starting base pair positions of the *gag* and *env* genes for Cervid endogenous retrovirus are indicated above the inverted triangles for size reference.^{9, 30}

Reverse transcription is prone to errors and genetic recombination for various reasons. Because the two homologous RNA strands are dimerized and, consequently, in close proximity, the RT enzyme is able to move between the two strands, using each as a template for reverse transcription. The ability to move between RNA strands during reverse transcription is a large source of genetic recombination in retroviral genomes.¹⁴ Similarly, a phenomenon called slipped-strand mispairing can occur during reverse transcription in which the RT enzyme falls off of the RNA template strand and begins transcribing the strand at a different location. Slipped-strand mispairing is most common in regions with homopolymer repeats.^{12, 15} Additionally, the RT enzyme does not have proof reading capabilities, causing an increased number of single nucleotide errors in the viral DNA.^{6, 26}

The Life Cycle of a Retrovirus: Integration of Retroviral DNA into Host Genome

After the viral RNA has been reverse transcribed into DNA, it is able to integrate into the host genome. Within the host, retroviruses tend to integrate into regions of high transcriptional activity due to the hindering steric features of the host chromatin in non-transcribing regions.⁴

Integrase, one of the enzymes contained within the capsid, cuts the host genome at a specific site in a nonlinear fashion, creating an overhang of host DNA.^{16, 24} The overhang can be four to seven base pairs long, depending on the integrase enzyme. A two base-pair overhang is created at the outermost end of each LTR of the viral DNA by the integrase enzyme as well. The overhang of the host and viral DNA must be complementary to each other in order for integration to be successful. The complementary overhangs of the viral DNA and host are aligned and host DNA repair mechanisms fill in the remaining complementary base pairs. Because the regions filled in by host machinery were bound together before the insertion of the viral DNA, the integration of viral DNA causes target site duplications, or identical sequences, on each end of the viral DNA.¹⁶

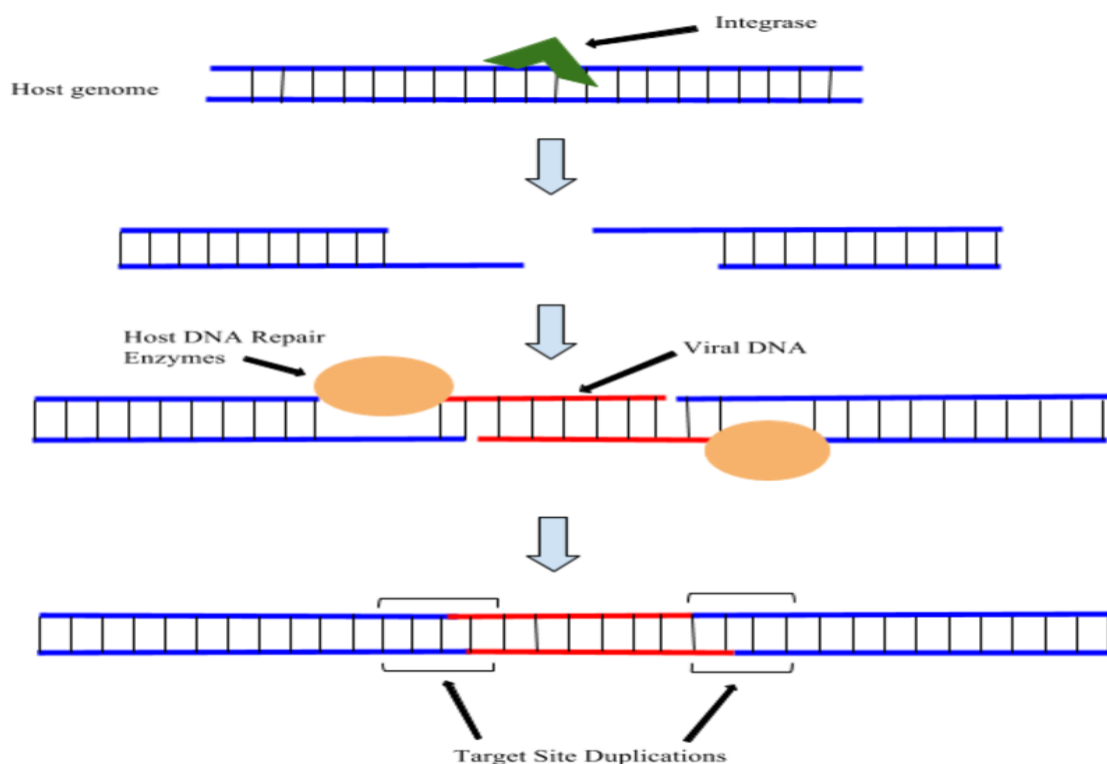


Figure 3. Creation of Target Site Duplication. Viral DNA is transported from the cytoplasm of the host to the nucleus, where it is integrated into the host genome via the viral enzyme, integrase. Integrase cuts the host genome and the viral genome to create overhangs on both the 5' and 3' end that are complementary between the host and viral DNA. The viral DNA is inserted into the host genome and host DNA repair mechanisms add complementary nucleotides to the remaining parts of the overhang, leaving target site duplications on each side of the viral DNA.

Endogenous Retroviruses

Endogenous Retroviruses (ERVs) arise from retroviruses that integrate, by chance, into the germ cells of a host.²⁹ Once integrated, ERVs can be vertically passed from parent to offspring.¹⁰ Consequently, an ERV can become fixed within a species over time. Fixation of a virus means that it is present in all individuals of the species at the same location within the genome. ERVs that are not fixed within a species are considered to be insertionally polymorphic.¹⁷

ERVs are common to all vertebrates, including humans, in which ERVs account for approximately 8% of the genome. Because chance integration into a germ line cell is not common, ERVs typically come to make up such a significant proportion of the genome through a combination of individual infections and intracellular retrotransposition.¹⁷ Retrotransposition is the process during which DNA copies of an initial ERV integrate into the host genome at different locations via an RNA transcript intermediate. Retrotransposed ERV sequences can typically be distinguished from one another by the flanking host sequence, as each ERV will occupy a unique site in the host genome.²⁹ Though ERVs can be transcriptionally active or inactive, only transcriptionally active ERVs are capable of retrotransposition.^{8, 17, 29}

The Impact of Endogenous Retroviruses on a Host

The insertion of retroviruses into the host genome can cause insertional mutagenesis, which is an alteration of the normal expression of host genes due to insertion of retroviral DNA near or in a host gene promoter. Insertional mutagenesis is possible because the ERV LTR is capable of acting as an alternative or, occasionally, the primary promoter for nearby host genes. In some cases, an ERV LTR can come to completely regulate the expression of the gene, a process called exaptation. Hosts can also exapt function from retroviral proteins. One of the most

well-known examples of exaptation of an ERV protein is the human endogenous retrovirus (HERV) family, HERV-W. A provirus in this family transcribes the human protein, syncytin, which assists in the formation of the placental membrane that separates mother from fetus.^{17, 20, 22}

Because of insertional mutagenesis, ERVs have been linked with disease. In humans, three HERV families, HERV-H, HERV-W, and HERV-K, have been associated with the neurodegenerative disease multiple sclerosis (MS).^{17, 21} Increased levels of multiple sclerosis-associated retrovirus (MSRV), a member of the HERV-W family, have been found in MS patients' blood, B-cells, cerebrospinal fluid, and MS lesions in the brain. Various studies have found that HERV-W is involved in inflammation and death of cells in such lesions. An increased level of HERV-H has been found in the serum and MS lesions in the brains of MS patients.^{21, 22} Elevated expression of HERV-K *pol* and insertionally polymorphic HERV-K113 in MS lesions in the brains of MS patients has also been observed.^{3, 21} Various other ERVs demonstrate polymorphisms in the LTRs that are common in patients with MS. A link between these polymorphisms and variations in MS epidemiology is hypothesized.³

Insertional mutagenesis is also believed to be associated with various cancers. For example, HERV-K is believed to produce proteins that bind to promyelocytic leukaemia zinc finger (PLZF) protein, which is critical in spermatogenesis. The binding of HERV-K proteins to PLZF is believed to activate protooncogenes and lead to tumors like seminomas.²² The integration of ERVs near protooncogenes is also believed to lead to certain types of lymphomas in humans.¹⁷ Additionally, retroviral RNA transcripts have been hypothesized to facilitate protein aggregation in numerous prion diseases.¹³

The Evolution of Endogenous Retroviruses

Once integrated, endogenous retroviruses are subject to the same evolutionary forces as the host genes, all of which can alter the expression and effects of the ERV.¹⁷ Typically, ERVs will accumulate numerous mutations over time. These mutations often render the ERVs transcriptionally inactive, or silenced, meaning that they are unable to create transcripts or retrotranspose.^{10, 29} In humans, only 7% of ERVs (HERVs) (approximately) are transcriptionally active.³ ERVs can also be actively silenced through host mechanisms like methylation and histone deacetylation.^{7, 8, 17}

ERVs that are detrimental to the host, meaning that they negatively affect reproductive ability, are more likely to be lost from the population or actively silenced. ERVs that are neutral or beneficial to their host are also likely to be actively silenced, but are more likely to remain in the genome and become fixed.^{17, 22} Insertion of ERVs near such sites that are critical to transcription and translation of proteins have the highest potential to be detrimental to the function of host genes. Accordingly, there are fewer fixed ERVs in proximity to host gene promoters and splice donor/acceptor sites than expected by chance.¹⁰

ERV replication and transcription can be halted through the formation of a solitary (solo) LTR. A solo LTR is formed when the 3' and 5' LTRs of an ERV undergo homologous recombination, excising all viral genes and leaving behind a single ERV LTR. Solo LTRs can still affect expression of proximal genes.²⁹ The formation of solo LTRs occurs most frequently in viruses that have recently integrated into a host, as these viruses are less likely to have accumulated mutations that would render the LTRs unique from each other.¹⁷

Odocoileus hemionus and Cervid ERVs

Odocoileus hemionus, commonly known as mule deer, are part of the cervidae (deer) family and currently range from central Mexico to northern Canada, including regions in northwest United States. There are typically low levels of genetic diversity between populations of mule deer, likely due to their long dispersal capabilities and large population sizes.^{18, 25} Since the speciation from sister species, *Odocoileus virginianus* (white-tailed deer), approximately 1.1 million years ago, mule deer have experienced numerous germ-line integrations of retroviruses. Because mule deer are a relatively young species, many of these cervid endogenous retroviruses (CrERVs) have endogenized recently and most are insertionally polymorphic (i.e. not fixed). Some CrERVs are also transcriptionally active.^{9, 31} Thus far, four lineages of CrERVs have been identified (**Figure 4**).¹⁸

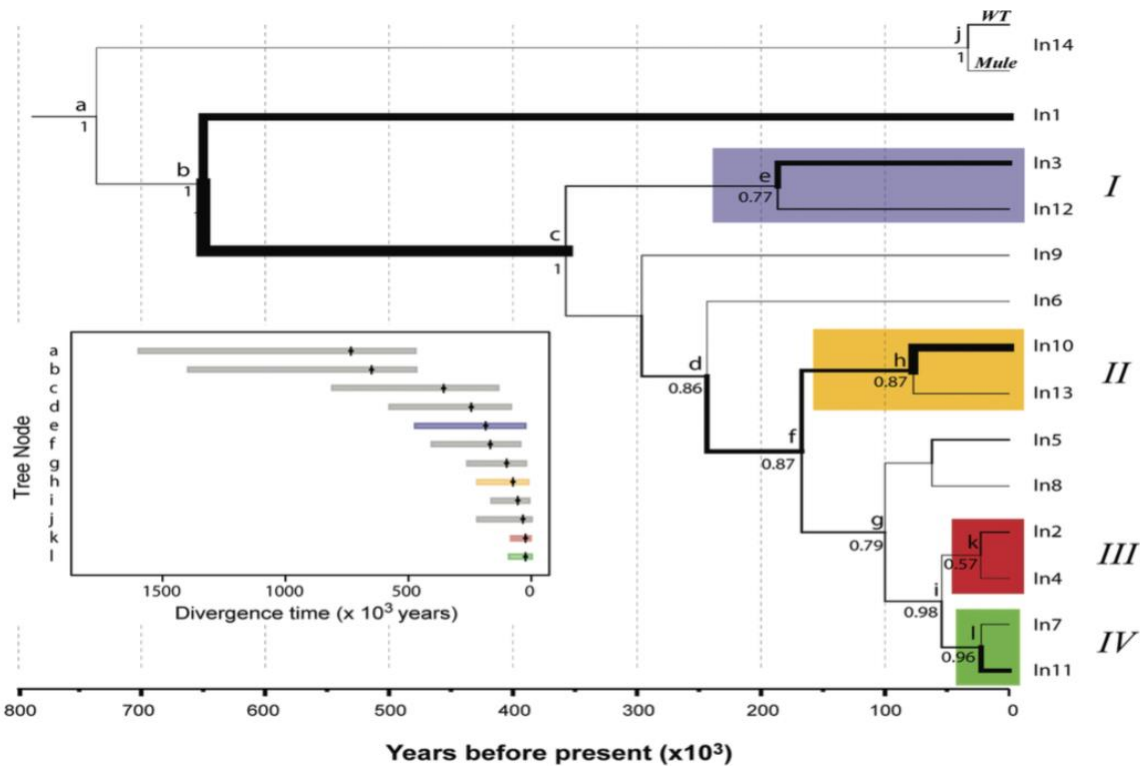


Figure 4. Phylogenetic tree of the 14 known CrERV with years before present (10^3) on the horizontal axis. The four lineages of CrERVs (I, II, III, and IV) are highlighted in purple, yellow, red, and green respectively.¹⁸

The transcriptionally active and insertionally polymorphic state of CrERVs raises the question of whether or not CrERVs could contribute to the presence and distribution of diseases that affect mule deer. One such disease is Chronic Wasting Disease (CWD), a prion disease that affects the cervidae family and is only found in North America and South Korea. Prion diseases, or transmissible spongiform encephalopathies, affect an organism by causing a host protein to misfold and aggregate within the central nervous system of the host. The accumulation of misfolded proteins leads to the neurodegeneration and inevitable death of the host.^{2, 28}

In the United States, CWD is endemic in wild mule deer populations in Wyoming (WY) and Northern Colorado (CO), but absent from populations to the north and west of this area, like Montana (MT), despite the high mobility and migratory habits of mule deer.^{2, 18} In order to determine if CrERVs could be involved in the susceptibility of mule deer to CWD, the profile of transcriptionally active CrERVs within two populations of mule deer will be compared: those from WY, a CWD endemic region, and those from MT, a CWD non-endemic region. Additionally, the integration site profile of CrERVs will be compared among the two populations and CrERV unique to either of the populations will be characterized.

The Poss Laboratory currently has data describing the distribution and location of CrERVs within the genomes of mule deer in MT. This data suggests that very few CrERVs found in MT mule deer are transcriptionally active. Characterization of transcriptionally active CrERV in MT deer was previously conducted. The Poss Laboratory also has data suggesting that mule deer within Wyoming (WY) have, on average, more CrERV integrations than MT mule deer. On average, MT mule deer have between 180 and 280 unique CrERV integrations, while WY mule deer contain approximately 333 CrERV integrations.¹ However, the transcriptional profiling of WY animals has not been completed.

This study attempts to identify a potential explanation for the quantitative difference in CrERV insertions between the two populations. The first potential explanation, investigated by Aim I, is that there is a difference in the transcriptional activity among the known CrERV families within WY and MT mule deer. Analyzing the differences in transcriptional activity could not only reveal a potential difference in retrotransposition, but could also reveal the presence of a novel family or subfamily of transcriptionally active CrERV in one or both populations. The second explanation, investigated by Aim II, is that WY mule deer have experienced infection and subsequent integration of novel CrERV lineages that may or may not be transcriptionally active. However, if novel CrERV are transcriptionally active, it should also be indicated in the exploration of Aim I.

Aim I

The first aim of this study is to identify all transcriptionally active CrERVs in mule deer from Wyoming (WY) and compare the phylogenetic diversity to that of the transcriptionally active CrERV previously found in Montana (MT) mule deer. The null hypothesis is that there are no differences in the transcriptionally active CrERVs in WY mule deer and in MT mule deer. An alternative hypothesis is that different CrERVs – specifically novel CrERV families – are transcriptionally active in WY mule deer and in MT mule deer. In Aim I, the transcriptional activity of CrERVs will be determined by generating complementary DNA (cDNA) from CrERV RNA extracted from lymph node samples of the mule deer. The cDNA will be used as a template for amplification of the CrERV spliced envelope transcript and the products will be sequenced. Phylogenetic analyses will be used to determine the lineage of all observed transcriptionally active CrERVs and the profile of transcriptionally active CrERV in WY and MT mule deer populations will be compared.

Aim II

The second aim of this study is to identify the profile of novel CrERV integration sites within the WY mule deer and compare to those previously identified in MT mule deer. Although the CrERV integration sites differ among the two populations of mule deer (WY and MT), it is not known if the lineages of CrERV are unique among the populations. CrERV integration sites are critical to analyze because the CrERV integration alone can affect the host genome (insertional mutagenesis). Consequently, only CrERV with integration sites near genes were investigated in Aim II. The null hypothesis is that there are no differences in the lineages of CrERV present in WY mule deer. If the null hypothesis were to hold true, it would suggest that the WY mule deer have experienced more retrotransposition of the same CrERV lineages found within the MT mule deer and our analysis would identify the CrERV lineage responsible for this activity. An alternative hypothesis is that WY animals will have unique CrERV integrations from MT animals. The integration sites of known CrERVs have been determined based on the flanking host sequences.¹ CrERV unique to each population will be determined from existing data on CrERV integration profiles and the sequence of unique CrERVs will be obtained by PCR amplification.

Chapter II: Methods

Mule Deer Sample Tissue Collection, CrERV RNA Extraction, and First-Strand cDNA

Synthesis

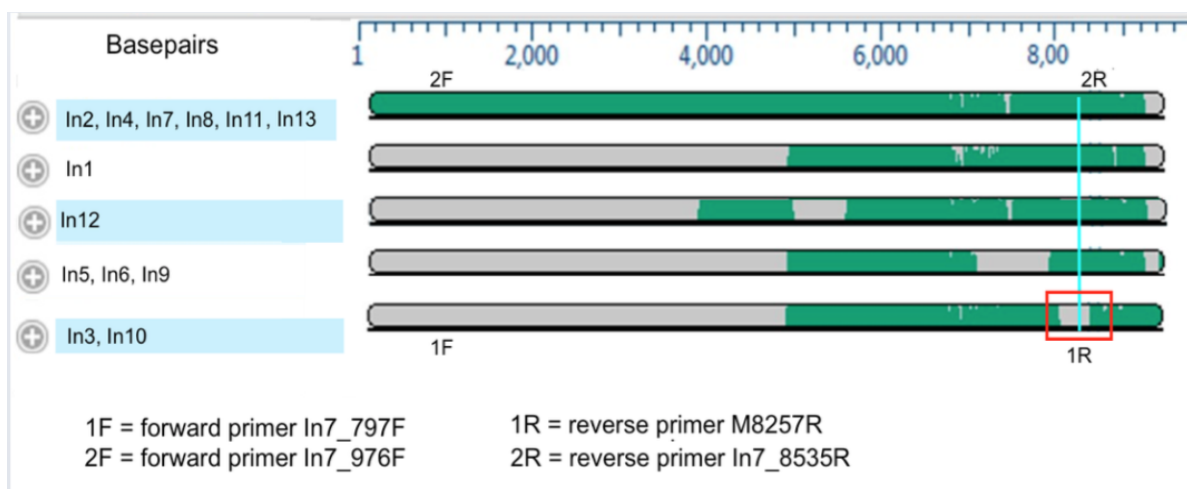
Tissue samples from the retropharyngeal lymph node were taken from 10 MT and 11 WY mule deer taken through hunter check stations in MT and WY respectively, as detailed by Kamath, *et al.* All RNA from the tissue samples was extracted using a TriZol extraction method. Complementary DNA (cDNA) was made from the CrERV RNA present within the extracted RNA using the AffinityScript Multiple Temperatures Reverse Transcription Protocol. Each PCR reaction was 20 μ L: 1 μ L of reverse primer In7_8555R, a primer specific to all CrERV, 2 μ L AffinityScript 2 Buffer, 0.8 μ L 100mM dNTP, 0.5 μ L RNase Block, 1 μ L AffinityScript Multiple Temperature Reverse Transcriptase, and the extracted RNA and nuclease-free water summing to 4.7 μ L (depending on how many microliters of RNA were extracted from the tissue sample). The mixture was incubated for 1 hour at 42°C followed by 15 minutes of incubation at 70°C.

Amplification of CrERV cDNA

The first-strand CrERV cDNA was used as a template for amplification by PCR using two primer sets: M8257R/In7_796F and In7_8535R/In7_797F (**Figure 5, a**). The former primer set excluded the In3 CrERV family (**Figure 5, b**). The In3 CrERV family was excluded with one primer set because this family produced significantly more transcripts, and consequently more DNA, than the other CrERV families of similar length, making it more difficult to detect other CrERV families during extraction. Consequently, using the two primer sets better sampled the diversity of transcriptionally active CrERV.

Primer Name	Primer Sequence
M8257R	AGAGCTGCGCAGACCCACCTT
In7_8535R	GCCTTAGCACCATAATCTGGATAGTATT
In7_796F	GATCCCGCGGCAGTTGACCA
In7_797F	ATCCCGCGGCAGTTGACC

a)



b)

Figure 5. a) Primer name and sequence. **b)** Primer diagram for CrERV showing the five main env structural variations found in CrERV and the families with those patterns. Blue vertical line highlights the position of primer 1R within the CrERV genome. Red box highlights absence of region including binding site of primer M8257R in In3 and In10 viruses.

For each animal, two 25µL PCR reactions were run for each primer set, meaning that a total of four PCR reactions were completed per animal. In addition to 1.25µL of the forward and reverse primer, 14.25µL of nuclease-free water, 5µL of 5x Buffer, 2µL of 2.5x dNTP, 0.25µL of Phusion HotStart enzyme (ThermoFisher Scientific), and 0.5µL of the template cDNA were included in each PCR reaction. The PCR conditions were 30 seconds at 98°C for initial denaturation, 10 seconds at 98°C for cyclic denaturation, 30 seconds at 60.8°C for primer set M8257R/In7_796F or 58.5°C for primer set In78535R/In7_797F for cyclic annealing, and 1 minute at 72°C for cyclic extension. Cyclic steps were run 34 times per PCR amplification.

Both primer sets amplified the spliced envelope region of the CrERV genome. The spliced envelope transcript was amplified because it has the highest genetic diversity within the CrERV genome and consequently yielded the highest diversity of CrERV sequences (unpublished observation). Additionally, by amplifying the spliced envelope transcript, the potential of sequencing proviral genome is eliminated.

Isolation, Cloning, and Sequencing of CrERV cDNA

The amplified spliced envelope region of the CrERV DNA was run in an agarose gel (1%) electrophoresis in order to separate the DNA by size. The DNA was isolated using the Qiagen QIAquick Gel Extraction Kit. The isolated CrERV DNA was cloned using the NEB PCR Cloning Kit (New England BioLabs) and the clones were sequenced using Sanger Sequencing.

Phylogenetic Analysis of CrERV DNA Sequences

The generated CrERV DNA sequences were analyzed using CLC Genomics Workbench 9.⁵ In order to identify unique CrERV sequences, all sequences were aligned and manually compared. Sequences were considered unique if they had three or more single nucleotide polymorphisms (SNPs). All SNPs found were verified by the trace data. Unique CrERV sequences from WY mule deer were used to generate a maximum likelihood phylogenetic tree to identify the lineage of unique WY CrERV sequences and to determine if novel lineages were present in the WY mule deer (**Figure 6**). This tree included previously deduced CrERV family (In1, In2, In3, etc.) reference sequences in order to classify the family of each CrERV sequence.¹⁸

In order to better analyze differences between MT and WY mule deer and the diversity within each CrERV family, a maximum likelihood tree with MT and WY mule deer sequences and the reference CrERV family sequences was created (**Figure 7**).⁵ Minimum Spanning

Networks were created for both the In1 and In12 CrERV reference sequences and the transcripts that clustered with each in order to better analyze the diversity of transcripts found within both of the CrERV families (**Figures 8 and 9**).¹⁹

Splice Pattern Determination

Splice pattern variation was determined by mapping the CrERV transcript sequences to the full-length In7 (accession number JN592050) CrERV family reference sequence in an exact alignment with no gap penalty.⁸

Identifying CrERV with Novel Insertions in WY Mule Deer

In order to determine the identity of the CrERV with novel insertions in CWD positive WY mule deer, two CrERV with novel insertions were sequenced by PCR amplifications and sequenced. The first amplification was a hemi-nested PCR that served as a screen to determine whether a virus was present at the targeted site. The first PCR amplification was completed using 0.5µL of both the viral specific primer and host primer PreLTR2F, 5µL 5x Buffer, 0.5µL 10 mM dNTPs, 1.0µL of (animal) DNA template, 0.25µL of Phusion HotStart Enzyme (ThermoFisher Scientific), and 17.25µL of water. The PCR conditions were 30 seconds at 98°C for initial denaturation, 10 seconds at 98°C for cyclic denaturation, 30 seconds at 63°C for annealing, and 45 seconds at 72°C for extension. The second denaturation through to the extension was repeated 34 times per amplification. The PCR products were then run on an agarose gel electrophoresis gel and, if a band of the correct size was identified during the screen, the animal DNA used in the first PCR amplification was used in a second PCR amplification.

The second PCR amplification was used to amplify the full-length virus, using 1.0µL of both the forward and reverse virus-specific primers. This PCR amplification also included 12.5µL of GoTaq long 2x Master Mix (Promega), 1.0µL of DNA template, and 9.5µL of water.

The PCR conditions were three minutes at 95°C for initial denaturation, 30 seconds at 94°C for cyclic denaturation, 30 seconds at 62°C for annealing, 8 minutes at 72°C for cyclic extension, and 10 minutes at 72°C for final extension. The second denaturation through the cyclic extension were repeated 36 times per reaction.

The PCR products were then run on an agarose gel electrophoresis. Products with full length virus bands were gel isolated (Monarch DNA Extraction Kit, New England BioLabs). Products of the gel isolation were used as a template in a third PCR using a viral primer and a virus-specific primer. The contents of the GoTaq Long PCR remained the same, except for the volume of DNA and water added, each of which varied depending on the concentration of the gel isolation product. The PCR conditions were three minutes at 95°C for initial denaturation, 30 seconds at 94°C for cyclic denaturation, 30 seconds at 58°C for annealing, five minutes at 72°C for cyclic extension, and 10 minutes at 72°C for final extension. The second denaturation through to the cyclic extension was repeated 36 times per reaction. Following PCR amplification, products were sequenced via Sanger Sequencing. A phylogenetic tree including the two CrERV sequences determined and other CrERV sequences that represented all known, major CrERV phylogenetic groups was made using PhyML with a GTR+I+G model (Yang, *et al.*, manuscript in preparation).

Chapter III: Results

Aim I

Analyzing CrERV Spliced Envelope Sequences in Wyoming Mule Deer

In determining the identity of all transcriptionally active CrERV within the WY mule deer population, Sanger sequencing the CrERV clones provided multiple CrERV spliced envelope transcripts from each WY mule deer. The unique (i.e. not identical to other transcripts from the same animal) transcripts and the reference envelope sequences of CrERV families 1-13 were used to generate a maximum likelihood tree (**Figure 6**).^{5, 18} All unique CrERV spliced envelope transcripts from the WY mule deer clustered with one of the reference sequences, suggesting that transcriptionally active, novel CrERV families are not present within the WY mule deer evaluated. Of the 13 reference sequences, the In1, In3, In5, and In12 families demonstrated transcriptional activity within the WY mule deer; however, a majority of the spliced envelope transcripts from WY mule deer clustered with the In1, the In3, and the In12 families. Only transcripts from animal WY1608 clustered with the In5 family. The In1 and In12 families demonstrated the highest diversity of transcripts.

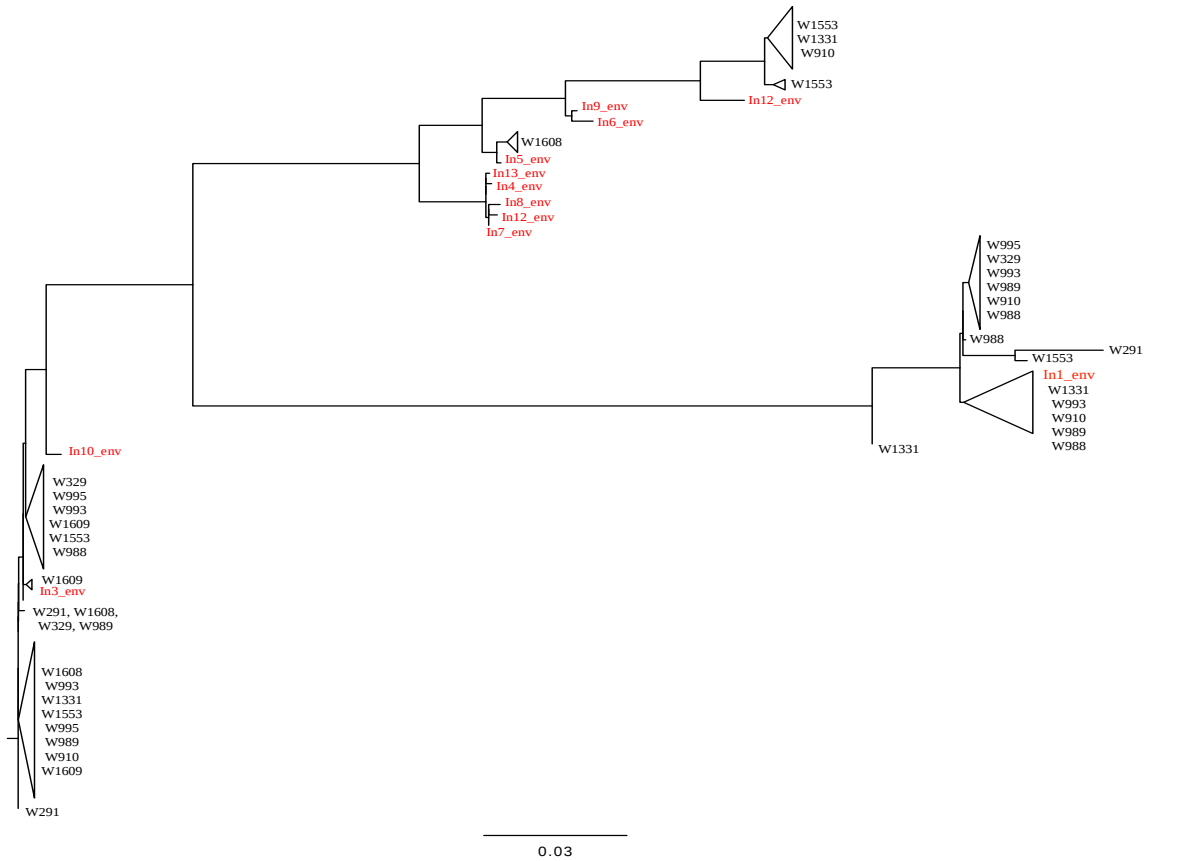


Figure 6. Maximum likelihood tree for all unique transcripts in Wyoming mule deer (W = Wyoming). Unique transcripts are represented solely by the name of the WY animal in which they were found. Triangles indicate collapsed branches with little phylogenetic diversity. Red font indicates a reference CrERV.¹⁹ Genetic diversity is determined by the horizontal length of branches, as noted by the scale provided.

Comparing CrERV Spliced Envelope Sequences of Wyoming and Montana Mule Deer

In order to analyze differences in the lineages of transcriptionally active CrERV between MT and WY mule deer, a maximum likelihood phylogenetic tree was generated using the unique CrERV sequences (transcripts) from each WY and MT animal and the reference CrERV family sequences (**Figure 7**).^{5, 18}

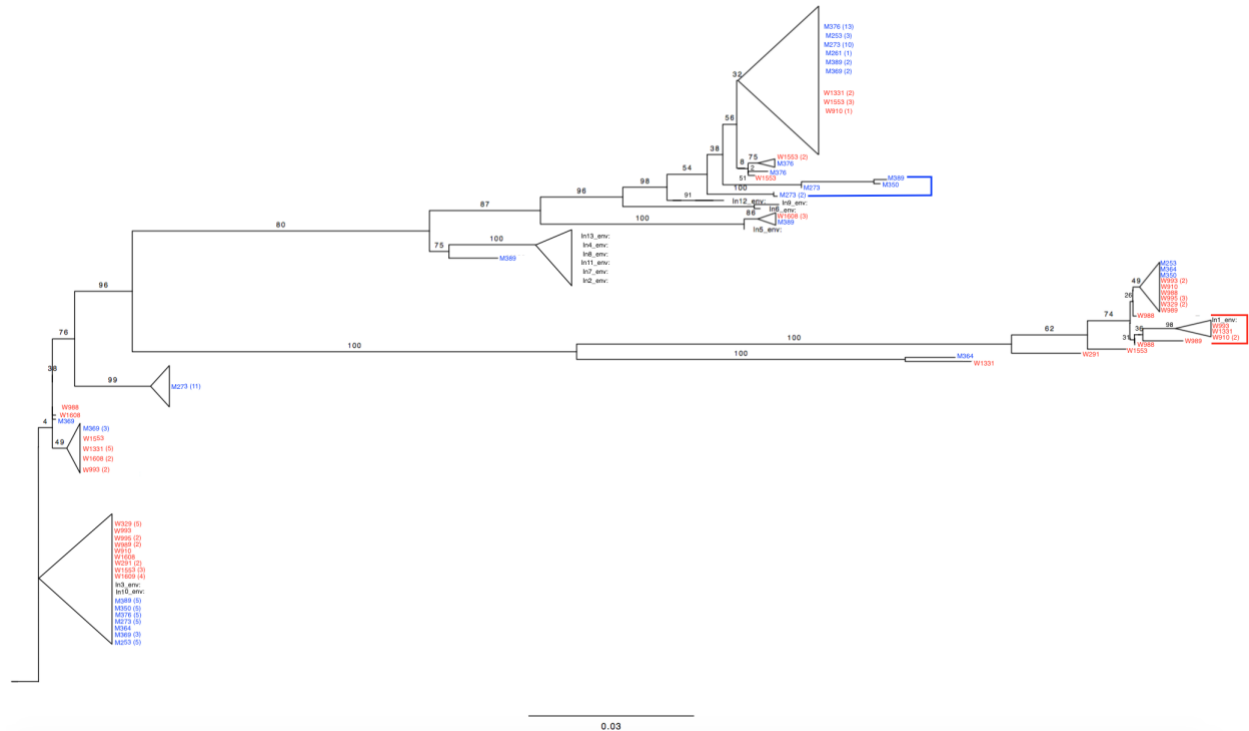


Figure 7. Maximum likelihood tree for all unique sequences in Montana and Wyoming mule deer. Blue font = Montana animal(s). Red font = Wyoming animal(s). Unique sequences are represented solely by the name of the animal in which they were found. The number of unique transcripts in each lineage from each animal is shown in parentheses next to the animal name. Triangles indicate collapsed branches with little phylogenetic diversity. Blue brackets show MT exclusiveness and red brackets show WY exclusiveness. Genetic diversity is determined by the horizontal length of branches, as noted by the scale provided.

The In12 family is dominated by sequences from MT mule deer and a cluster of sequences within the In12 family is exclusive to MT mule deer (bootstrap = 100). The In1_env is dominated by WY mule deer and there is a WY exclusive cluster within the In1 family (bootstrap = 98). Other sequences from WY mule deer surround the WY exclusive cluster with lower bootstrap support. Within the In3 family, one cluster of sequences is exclusively found in one MT mule deer (M273) (bootstrap = 99). Despite the formation of distinct clusters around the In3 reference sequence, there are few differences between the sequences within the clusters and the populations representing them, indicating little diversity among the In3-like sequences for both WY and MT mule deer. Only one CrERV sequence from a MT animal (M389) clustered

with the In2, In4, In8, In7, In11, and In13, CrERV families. The In5-like sequences are found in one WY mule deer (W1609) and one MT mule deer (M389).

In order to better visualize differences between the unique sequences of each MT and WY mule deer and the diversity within each CrERV family, minimum spanning networks (MSN) were created for the In1 and In12 CrERV lineages, as these families showed the most diversity (**Figures 8 and 9**).^{18, 19} Both networks included the CrERV family reference sequence and the CrERV sequences from WY and MT mule deer that associated with each reference sequence in the WY/MT maximum likelihood tree (**Figure 7**).

For the purposes of analysis, sequences are unique if they are not identical to other sequences from the same animal or different animals. Additionally, sequences were defined as being in different clusters in the MSNs if there were 7 or fewer SNPs between them (as indicated by the number in the parenthesis over each branch).

The MSN for the In1 shows two distinct clusters, Cluster A and Cluster B (**Figure 8**). Cluster B is WY exclusive and contains the In1 reference sequence, while cluster A contains viruses present in both WY and MT animals, but is dominated by WY mule deer. The two clusters are separated by viruses (CrERV sequences) W910sp_23 and W988sp_4, which differ by eight SNPs.

The six unique viruses in Cluster B were extracted from four WY animals: W1331 (contains two unique viruses in cluster B), W910 (contains two unique viruses in cluster B), W993, and W291. One of the viruses, WY1331sp_10, is identical to the In1_env reference sequence. Two of the viruses in the cluster B contain deletions, as indicated by the dashed line. Because cluster B includes viral sequences with large deletions, it differs slightly from the WY exclusive In1 branch observed in the ML tree (**Figure 7**); the four viruses in Cluster B without

deletions make up the WY exclusive branch (bootstrap = 98) observed in **Figure 7**. It is possible that all unique viruses in cluster B are the In1 reference sequence with multiple single nucleotide polymorphisms (SNP) and deletions. This type of diversity is expected because the In1 family has been in the mule deer genome for over 700,000 years and ERVs are subject to the same evolutionary forces as their host.

Cluster A contains eight unique viruses, which were found in seven different WY mule deer and three different MT mule deer. Of the eight viruses, four were found exclusively in WY mule deer, three are found exclusively in MT mule deer, and one was found in both WY and MT mule deer. Two of the viruses found exclusively in WY animals were found in W988, while the two other viruses exclusively from WY mule deer were found in W989 and W1553. The virus found in WY and MT animals was found in six WY deer and one MT mule deer (M253). Two of the viruses found exclusively in MT were found in M364. The other virus found exclusively in MT mule deer was found in M350. Two out of the 8 viruses in cluster A contain deletions, one from a WY mule deer and the other from a MT mule deer (W1553 and M364). Cluster A showed much less diversity than cluster B, suggesting that cluster A may be composed of a recently retrotransposed virus.

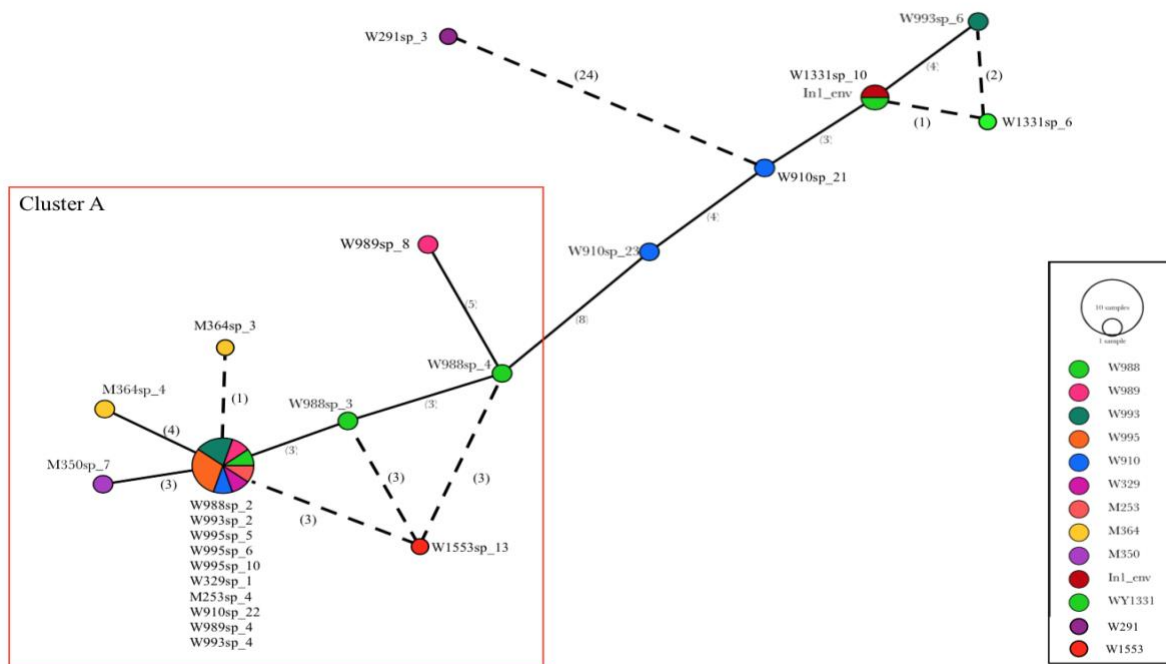


Figure 8. Minimum Spanning Network of CrERV sequences from MT and WY mule deer that were clustered with In1_env in **Figure 7**. Dotted lines indicate sequences that contain a deletion with phylogenetic information. Cluster A is indicated with a red box and label. The number of differences in base pairs between sequences is shown in parentheses. Transcripts within the same circle are identical or differ by 2 or less SNPs.

The MSN for the In12 family shows three distinct clusters – of which two were MT dominant (Cluster B and C) and one was MT exclusive (Cluster A) – and one outlier virus (from an MT animal) (**Figure 9**). Cluster A contains two unique viruses from two different MT animals, M389sp_9 and M350sp_6, which differed from each other by six SNPs. The latter virus differed from the most closely related virus of Cluster B by 36 SNPs. With respect to **Figure 7**, Cluster A represents part of the MT exclusive In12 branch indicated by the MLT. The remaining part of the MT exclusive branch in the MLT is the outlier, M273sp_73. This outlier differed from the most closely related virus of Cluster B by 27 SNPs.

Cluster B, the largest cluster, contains 3 unique viruses. Of the 12 animals that contained sequences found in Cluster B, only four were WY mule deer (WY1608, WY1609, W910, and

W1331); thus, Cluster B is MT dominant. One of the unique viruses is exclusively from a WY mule deer (W1331), another is exclusively from a MT mule deer (M389), and the third is found in both WY and MT mule deer. Cluster B varied from the most closely related unique virus of the Cluster C by 8 SNPs.

Cluster C has five unique viruses from four different animals, three of which were MT mule deer (M167, M261, and M364); therefore, cluster C MT dominant. Of the five unique viruses, four were found exclusively in MT mule deer. This cluster was closest related to the reference In12 sequence (via transcript M261sp_1, 41 SNPs) of all In12 clusters. Cluster B and C were represented as one cluster in the ML tree of unique viruses from MT and WY mule deer (Figure 7).

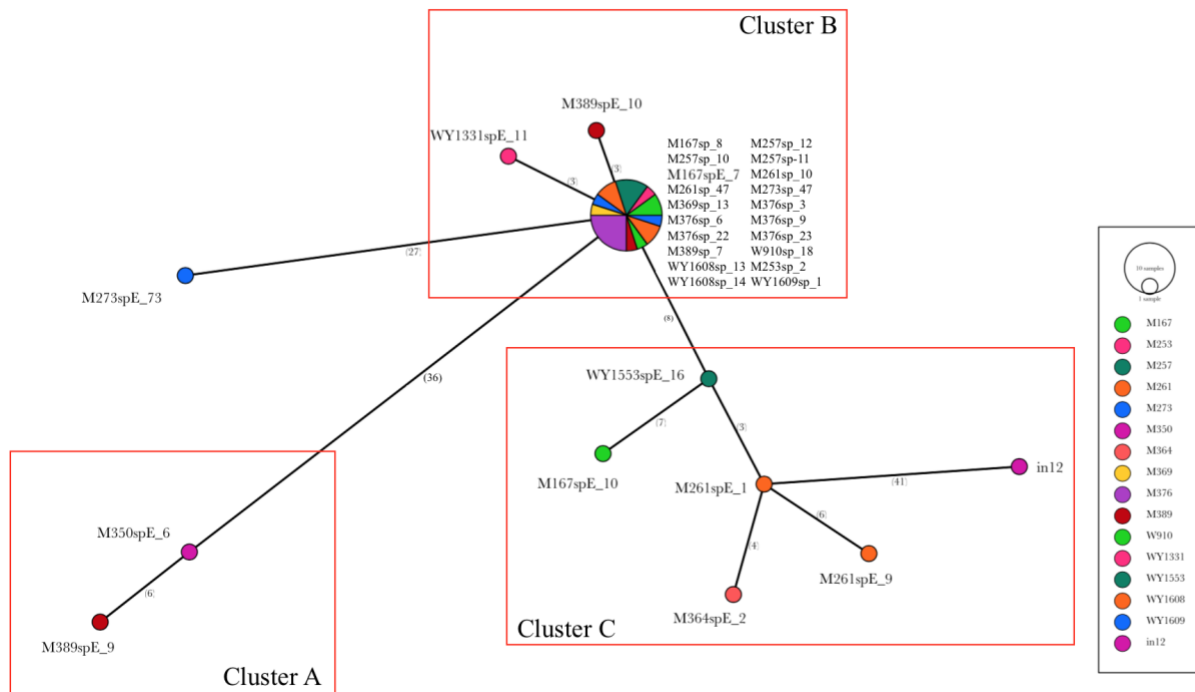


Figure 9. Minimum Spanning Network of CrERV sequences from MT and WY mule deer that were clustered with In12_env in Figure 7. Clusters A, B, and C are indicated with red boxes and labeled accordingly. The number of differences in base pairs between sequences is shown in parentheses. Transcripts within the same circle are identical or vary by 2 or less SNPs.

The data regarding the comparison of the transcriptionally active WY and MT viruses from **Figures 7, 8, and 9** is summarized in **Figure 10**. Overall, more MT mule deer were found to contain In12-like viruses than WY mule deer, while more WY mule deer appeared to contain In1-like viruses than MT mule deer. The In1 family contained a WY exclusive cluster (B) and a WY dominant cluster (A), while the In12 family contained two MT dominant clusters (B and C) and a MT exclusive cluster (A).

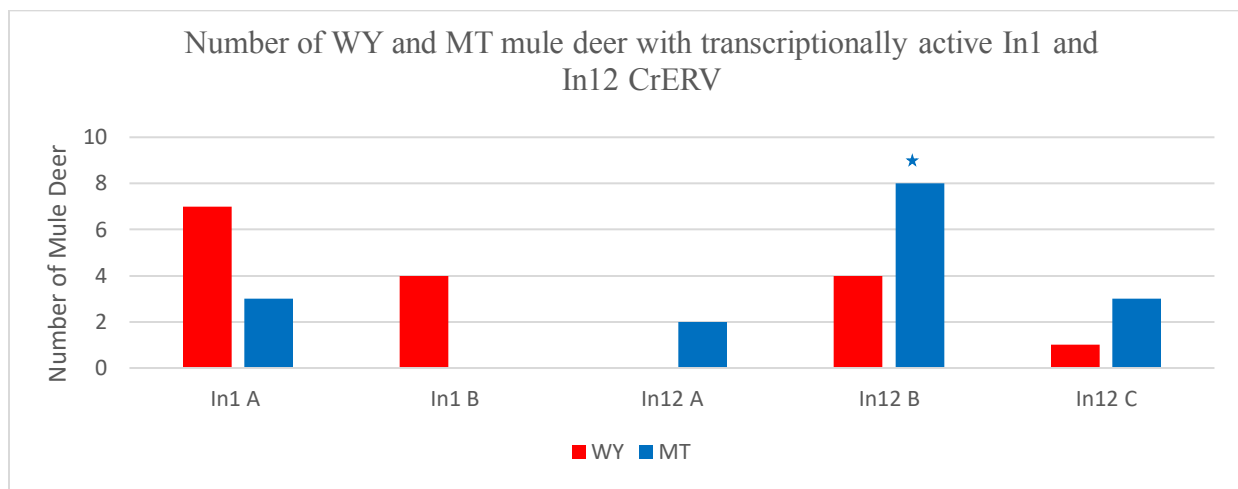


Figure 10. The number of WY (red) and MT (blue) mule deer within the In1 and In12 clusters. The star indicates the presence of an outlier (M273).

Splice Pattern Variations

Among all transcripts, two unique splice pattern variations were found (**Figure 11**). One of the patterns is the canonical splice pattern (**Figure 11**, “1”) and the second splice pattern contained additional splice acceptors and donors (**Figure 11**, “2”). The second splice pattern variation was found in viruses within the In3 and In12 CrERV families. An alternative splice acceptor (**Figure 11**, “Alt.”) for the canonical splice donor was also observed in some In3 viruses. The alternative splice acceptor was due to a deletion in the genome sequence, relative to the full length In7 virus, that ablated the use of the canonical splice acceptor.

There appear to be no difference between the MT and WY mule deer in expression of the second splice pattern or use of the alternative splice acceptor. A total of 5 transcripts from 4 animals (W1553, M369, M253, and W1609) contained the second splice pattern. The majority of these transcripts were In3-like, while one transcript from W1553 (W1553sp_21) was In12-like. Eight In-3 like transcripts from 5 animals (W989, M350, W1609, M273, and M369) used the alternative splice pattern.

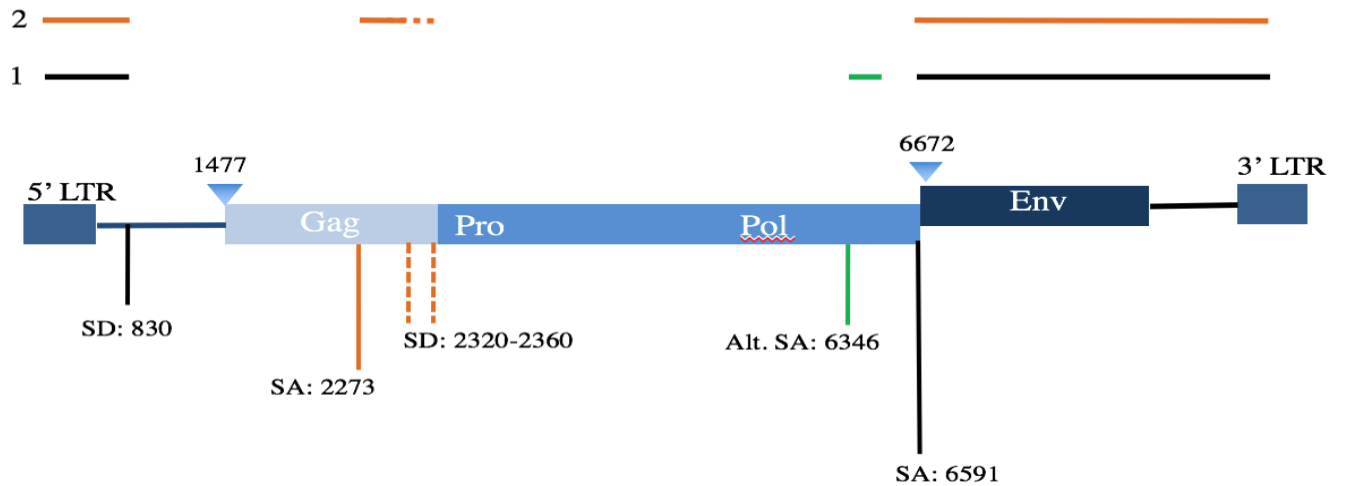


Figure 11. The two splice patterns observed within CrERV sequences. Solid vertical lines indicate base pair positions of splice donors (SD) and splice acceptors (SA). Black vertical lines indicate the canonical SD and SA. Black horizontal lines show the transcript produced using the canonical splice pattern. Orange vertical lines indicate the SDs and SAs of the second splice pattern. Orange horizontal lines show the transcript produced using the second splice pattern. The green vertical line indicates the alternative SA for the canonical SD. The green horizontal line shows the additional part of the canonical transcript produced with alternative SA. Note: transcripts using the alternative SA are followed by a deletion that would ablate the use of the canonical SA.

Diversity of Transcriptionally Active CrERV within Individual Animals

The diversity of CrERV within individual mule deer is shown in **Figures 12 and 13**, which were generated from the data shown in **Figures 7, 8, and 9**. As shown, there are mule deer with high and low diversity of transcriptionally active CrERV lineages within both populations; however, MT mule deer appear to have more diversity in the number of transcriptionally active,

unique (i.e. not identical to other viral transcripts within the same animal) viruses and within the number of CrERV families with which those unique viruses are associated. WY mule deer contained a range of two to five transcriptionally active, unique viruses from two to four different CrERV families or clusters (**Figure 12**). MT mule deer contained a range of one to six unique, transcriptionally active viruses, from a one to five CrERV families and clusters (**Figure 13**). Accordingly, the animals with the highest and lowest diversity of CrERV families were from MT (M389 and M257 respectively).

No WY animals contained more than one transcriptionally active In12 virus and, again, only one WY animal (W1553) contained a In12 Cluster C virus. All WY mule deer, except two (W1608 and W1609), were found to contain an In1-like unique viruses, the former of which (W1608) was the only WY mule deer found to contain a transcriptionally active In5-like virus. Therefore, most WY mule deer express members of the In1 family, however, expression of diverse members of the family occur within individual WY mule deer. Only two WY mule deer (W993 and W910) contained transcriptionally active viruses from both In1 clusters. Unlike WY mule deer, only 3 MT mule deer contained transcriptionally active viruses from the In1 family, all of which were in cluster A. Additionally, an In3-like virus was found in every WY mule deer sampled, while unique In3-like viruses were not found in three MT mule deer (M261, M257, and M167).

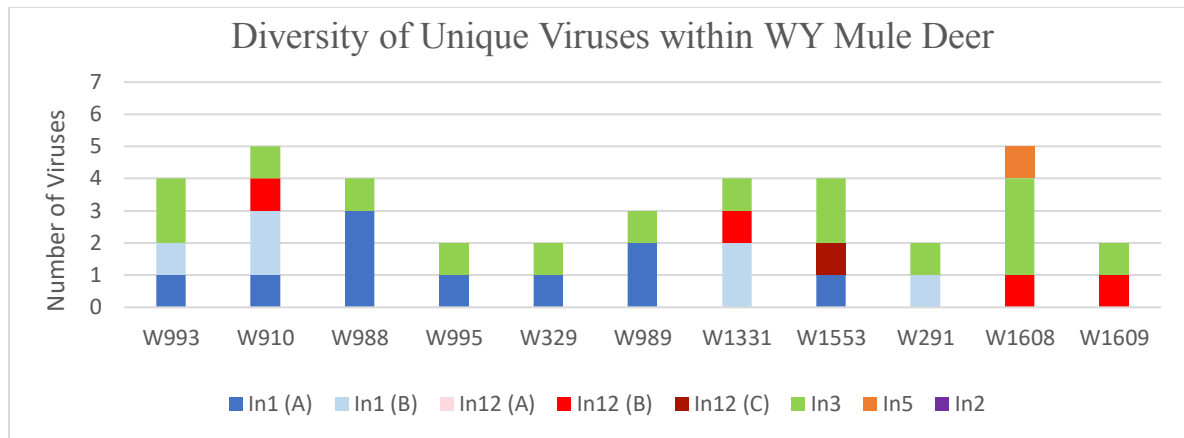


Figure 12. Diversity of Unique Viruses within WY mule deer.

All MT mule deer, except two (M350 and M364), contained one or more transcriptionally active, unique viruses from the In12 Cluster B. Therefore, expression of diverse members of the In12 family occurs within individual MT mule deer and not necessarily among the MT population. Animal M350 was the only mule deer to contain an In12-like virus from Cluster A and no other In12-like viruses. Again, only 3 MT mule deer (M364, M253, and M350) contained transcriptionally active In1 viruses, all of which were from cluster A. Animal M389 was the only MT mule deer to contain a transcriptionally active In5-like virus and the only mule deer sampled (from WY or MT) to contain a transcriptionally active In2-like virus. M389 was also the only animal to contain transcriptionally active viruses from both the In12 A and B cluster. Two of the MT mule deer, M261 and M167, were found to contain viruses in both the In12 Cluster B and Cluster C. All MT mule deer besides the three fore-mentioned mule deer (M389, M261, and M167) were found to contain transcriptionally active viruses from only one In12 cluster.

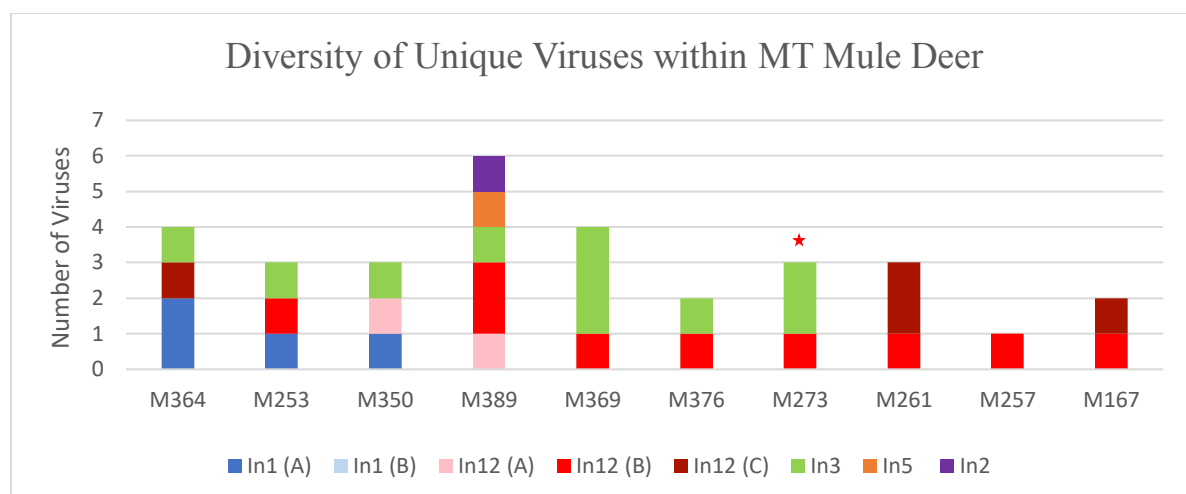


Figure 13. Diversity of Unique Viruses within WY mule deer. Star indicates the presence of an outlier (M273sp_73) in the In12 CrERV family (ref. **Figure 9**).

Aim II

Identifying CrERV with Novel Insertions in WY Mule Deer

In order to determine whether any of the CrERV found in WY mule deer were of novel CrERV lineages with respect to previously identified CrERV, several proviral CrERV sequences found only in WY were sequenced. The phylogenetic tree included two CrERV sequences (W291_S7236 and W1331_S901) generated from the extracted DNA of two different WY mule deer (W291 and W1331, respectively) and previously sequenced CrERV representative of all known, major CrERV phylogenetic groups (Yang, *et al.*, manuscript in preparation). The phylogenetic tree indicated that the two sequences were not novel CrERV lineages. W291 S7236 clustered among In6-like and In12-like CrERV, suggesting that it could be a recombinant of the two families. W1331 S901 clustered with In7-like CrERV. Both families are relatively young lineages and could provide evidence of expansion via retrotransposition in WY animals.

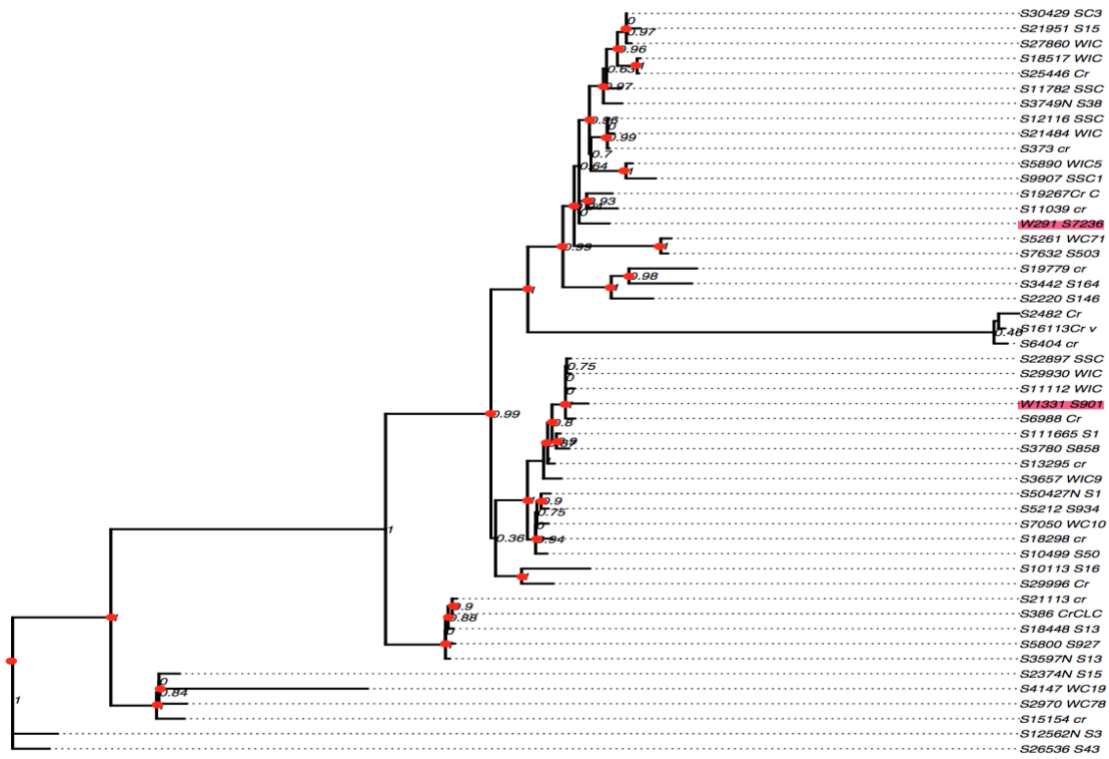


Figure 14. Phylogenetic tree of all previously sequenced CrERV representative of all known, major CrERV phylogenetic groups and two unique CrERV sequenced from WY mule deer (pink highlight) (Yang, *et al.*, manuscript in preparation). Red dots indicate supported nodes.

Chapter IV: Discussion

This study strove to explore possible explanations for the observed difference in the number of CrERV integration sites within Wyoming (WY) and Montana (MT) mule deer. On average, MT mule deer have 250 unique CrERV integrations, while WY mule deer contain approximately 333 CrERV integrations.¹ The first aim of this study used transcript data to determine if increased transcriptional activity of either a novel or known CrERV could account for a difference in CrERV integrations. The transcripts identified in WY were the same as those previously reported in MT mule deer. An aberrant splice pattern could indicate a difference in genome sequence in a known CrERV lineage, which could affect retrotransposition. However, WY and MT mule deer showed no difference in splice patterns based on the transcript data; both the canonical splice pattern and an alternative splice pattern were found in both MT and WY mule deer. Therefore, it is unlikely that the increased number of CrERV integrations in WY mule deer is due to the infection of a novel CrERV family.

Aim II investigated the phylogeny of CrERV lineages found only in WY mule deer by analyzing proviral sequences. Unlike aim I, this approach could detect a novel virus that was not transcriptionally active. Two CrERV integrations site that were only found in WY mule deer were sequenced; both were from known phylogenetic groups described in MT mule deer. Consequently, there was no support based on this limited sample to indicate a novel CrERV infection event. However, because these families were relatively young, they could provide evidence of retrotransposition or an increased number of germline infections. We previously determined that a CrERV in the In7 lineage, one of the two phylogenetic groups identified in Aim II, recombined with a virus in the In1 family to create an In3 virus, which subsequently retrotransposed giving rise to the closely related In3 family (Yang, *et al.*, manuscript in

preparation). In order to investigate potential recombination, retrotransposition, and multiple unique germ-line infections, full-length viruses would need to be sequenced to determine their phylogenetic relatedness. Recombination would be indicated by phylogenetic discordance with two CrERV families. Retrotransposition would present as closely related monophyletic CrERV while multiple infections would contain higher diversity in the branch structure of a monophyletic group. More CrERV found only in WY animals should be sequenced and their phylogeny analyzed to thoroughly explore this possibility.

Although this study found no evidence that WY mule deer contain CrERV from novel CrERV families or from CrERV families absent from MT mule deer, the results indicate that there are differences in the diversity of transcriptionally active CrERV between the WY and MT mule deer populations. This diversity was most notable in the In1 CrERV families. The maximum likelihood tree (MLT) (**Figure 7**) shows that there was higher sequence diversity in In1 env transcripts in WY compared to MT mule deer. A minimum spanning network (MSN) for In1 clearly shows two distinct clusters, with cluster B only found in WY mule deer. Because cluster A showed much less diversity than cluster B, it may be composed of a recently retrotransposed CrERV. Such retrotransposition could explain the increased number of CrERV integrations in the WY mule deer, as cluster A is WY dominant. Additionally, the lack of transcripts from cluster B in MT mule deer suggests a WY-exclusive retrotransposition of an In1 virus.

During these investigations, transcript data (Aim I) and a small survey of CrERV unique to WY mule deer (Aim II) provided no evidence to support the presence of novel CrERV in WY mule deer. However, evidence to suggest expansion via retrotransposition of the In1 lineage was found in the WY mule deer.

Literature Cited

- ¹Bao, L., Elleder, D., Malhotra, R., DeGiorgio, M., Maravegias, T., Horvath, L., Carrel, L., Gillin, C., Hron, T., Fabryova, H., Hunter, D., and Poss, M. Computational and Statistical Analyses of Insertional Polymorphic Endogenous Retroviruses in a Non-model Organism. *Computation*. (2014). **2**: 221-245.
- ²Bourne, D.C. Chronic wasting disease of cervids. *Small Ruminant Research*. (2015). **128**: 72-78.
- ³Brutting, C., Emmer, A., Kornhuber, M., and Staege, M. S. A survey of endogenous retrovirus (ERV) sequences in the vicinity of multiple sclerosis (MS)-associated single nucleotide polymorphisms (SNPs). *Molecular Biology Reports*. (2016). **43 (8)**: 827-836.
- ⁴Bushman, F., Lewinski, M. Ciuffi, A., Barr, S., Leipzig, J., Hannenhalli, S., and Hoffman, C. Genome-wide Analysis of Retroviral DNA Integration. *Reviews Microbiology*. (2005). **3 (11)**: 838-858.
- ⁵CLC Genomics Workbench 9.0 (<https://www.qiagenbioinformatics.com/>)
- ⁶Coffin, JM, Hughes, SH, Varmus, HE, editors. *Retroviruses*. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press; 1997. The Place of Retroviruses in Biology. Available from: <https://www.ncbi.nlm.nih.gov/books/NBK19382/>.
- ⁷Cohen, C., Lock, W.M., and Mager, D.L. Endogenous retroviral LTRs as promoters for human genes: A critical assessment. *Gene*. **448**: 105-114.
- ⁸Dewannieux, M. and Heidmann, T. Endogenous retroviruses: acquisition, amplification, and taming genome invaders. *Current Opinion in Virology*. 2013. **3**: 646-656.

- ⁹Elleder, D., Kim, O., Padhi, A., Bankert, J.G., Simeonov, I., Schuster, S.C., Wittekindt, N.E., Motameny, S. and Poss, M. Polymorphic Integrations of an Endogenous Gammaretrovirus in the Mule Deer Genome. *Journal of Virology*. (2012). **86 (5)**: 2787-2796.
- ¹⁰Feschotte, C. and Gilbert, C. Endogenous viruses: insights into viral evolution and impact on host biology. *Nature Reviews Genetics*. 2012. **13**: 283-296.
- ¹¹Frommer M., McDonald, L.E., Millar, D.S., Collis, C.M., Watt, F., Grigg, G.W., Molloy, P.L., Paul, C.L. A Genomic Sequencing Protocol that Yields a Positive Display of 5-methylcytosine Residues in Individual DNA Strands. *Proceedings of the National Academy of Sciences of the United States of America*. (1992). **89**:1827–1831.
- ¹²Fungtammasan, A., Tomaszewicz, M., Campos-Sánchez, R., Eckert, K.A., DeGiorgio, M., Makova. K.D. Reverse Transcription Errors and RNA–DNA Differences at Short Tandem Repeats. *Mol. Biol. Evol.* 2016; **33 (10)**: 2744-2758.
- ¹³Gomes, M.P.B., Cordeiro, Y., and Silva, J.L. The peculiar interaction between mammalian prion protein and RNA. *Prion*. (2008). **2(2)**: 64-66.
- ¹⁴Greuthe, Jane. The Retroviral RNA dimer linkage: different structures may reflect different roles. *Retrovirology*. 2004. **1 (22)**.
- ¹⁵Hamburgh, M.E., Curr, K.A., Monaghan, M., Rao, V.R., Tripathi, S., Preston, B.D., Sarafianos, S., Arnold, E., Darden, T., and Prasad, V.R. Structural Determinants of Slippage-mediated Mutations by Human Immunodeficiency Virus Type 1 Reverse Transcriptase. *Journal of Biological Chemistry*. 2006. **281 (11)**: 7421-7428.
- ¹⁶Hindmarsh, P. and Leis, J. Retroviral DNA Integration. *Microbiology and Molecular Biology Reviews*. (1999). **63 (4)**. 836-843.

- ¹⁷Jern, P. and Coffin, M. Effects of Retroviruses on Host Genome Function. *Annual Reviews Genetics*. 2008. **47**: 709-732.
- ¹⁸Kamath, P. L., Elleder, D., Bao, L., Cross, P. C., Powell, J. H., and Poss, M. The Population History of Endogenous Retroviruses in Mule Deer (*Odocoileus hemionus*). *Journal of Heredity*. (2014). **105(2)**: 172-187.
- ¹⁹Leigh, J. W., Bryant, D. and Nakagawa, S. (2015). Popart: Full-feature software for haplotype network construction. *Methods Ecol. Evol.*, **6**.
- ²⁰Maksakova, I.A., Romanish, M.T., Gagnier, L., Dunn, C.A., van de Lagemaat, L.N., Mager, D.L. Retroviral Elements and Their Hosts: Insertional Mutagenesis in the Mouse Germ Line. *PLoS Genetics*. (2006). **2(1)**.
- ²¹Morandi, E., Tanasescu, R. Tarlinton, R.E., Constantinescu, C.S., Zhang, W., Tench, C., and Gran, B. The association between human endogenous retroviruses and multiple sclerosis: A systematic review and meta-analysis. *PLoS ONE*. (2017). **12 (2)**.
- ²²Moyes, D., Griffiths, D.J., and Venables, P.J. Insertional polymorphisms: a new lease of life for endogenous retroviruses in human disease. *Trends in Genetics*. (2007). **23 (7)**.
- ²³Muriaux, D., Mirro, J., Harvin, D., Rein, A. RNA is a structural element in retrovirus particles. *Proceedings of the National Academy of Sciences*. 2001. **98 (9)**: 5246-5251.
- ²⁴Nisole, S. and Saib, A. Early steps of retrovirus replicative cycle. *Retrovirology*. 2004. **1 (9)**.
- ²⁵Powell, J.H., Kalinowski, S.T., Higgs, M.D., Ebinger, M.R., Vu, N.V., and Cross, P.C. Microsatellites indicate minimal barriers to mule deer *Odocoileus hemionus* dispersal across Montana, USA. *Wildlife Biology*. (2013). **19**: 102-110.
- ²⁶Roberts, J.D., Bebenek, K., Kunkel, T.A. The accuracy of reverse transcriptase from HIV-1. *Science*. 1988. **212 (4882)**: 1171-1173.

- ²⁷Seibert, S.A., Howell, C.Y., Hughes, M.K., Hughes, A.L. Natural Selection on the gag, pol, and env, Genes of Human Immunodeficiency Virus 1 (HIV-1). *Molecular Biology and Evolution*. 1995. **12** (5): 803-813.
- ²⁸Sigurdson, C.J. A prion disease of cervids: Chronic wasting disease. *Veterinary Research*. (2008). **39** (4).
- ²⁹Stoye JP. Studies of endogenous retroviruses reveal a continuing evolutionary saga. *Nature Reviews Microbiology*. 2012. **10** (6): 395-406.
- ³⁰Uren, A.G., Kool, J., Berns, A., and van Lohuizen, M. Retroviral insertional mutagenesis: past, present, and future. *Oncogene*. (2005). **24**: 7656-7672.
- ³¹Wittekindt, N. E., Padhi, A., Schuster, S. C., Qi, J., Zhao, F., Tomsho, L. P. Kasson, L. R., Packard, M., Cross, P., and Poss, M. Nodeomics: Pathogen Detection in Vertebrate Lymph Nodes Using Meta-Transcriptomics. *PLOS ONE*. (2010). **5** (10): e13432.

Academic Vita of Stephanie Williams

Objective	Career as a physician, specializing in pediatric medicine and care
Education	Pennsylvania State University, Schreyer Honors College (2015 - present) • Major: Biology with a focus in genetics and development • Minor: Psychology Hershey High School (2011 - 2015)
Medical Experience	Atlantis Project: internship in El Bierzo Hospital in Ponferrada, Spain (05/12/16 - 06/19/16) • Assisted and observed physicians in numerous specialties Internship at the Pennsylvania State University Hershey Medical Center (HMC) through the PA Youth Apprenticeship Program (2013 - 2015) • Two hours a day, five days a week • Observed in 36 specialties and outside facilities (2013 - 2014) • Observed in the Neonatal Intensive Care Unit, the Pediatric Intensive Care Unit, and the Pediatric Intermediate Care Unit for one semester and in the Emergency Department for another (2014 - 2015) • Attended weekly, educational lectures about various medical professions, wrote weekly reports, compiled frequent case studies, and presented a final case study for staff of HMC and Hershey High School
Research Experience	Conduct primary research in Dr. Mary L. Poss' Laboratory (2016 - present) • Study the diversity of transcriptionally active endogenous retroviruses in mule deer • Awarded a \$1,000 Undergraduate Research Support Grant by the Office of Science Engagement, Eberly College of Science (2017)
Awards and Scholarships	Pennsylvania State University • The Evan Pugh Scholar Senior Award (2018) • President Sparks Award (2016 and 2017) • The President's Freshman Awards (2016) Hershey High School • Hershey High School Academic Excellence Award (2015) • Rotary Club Scholarship used to attend the Rotary Youth Leadership Award Conference (2014)

- Honors Societies**
- National Society of Leadership and Success** (2017 – present)
 - Alpha Epsilon Delta: National Health Pre-Professional Honor Society** (2015 - 2017)
 - National Honor Society** (2013 – 2015)
- International Service**
- Atlantis Project** (05/12/16 – 06/19/16)
 - Taught English to the Spanish doctors working in El Bierzo Hospital in Ponferrada, Spain through conversation (8 hours a week)
 - People-to-People Youth Cultural Ambassador** (6/27/13 – 7/15/13)
 - Performed community service aimed at bridging the gap between American and Chinese students
 - Taught English to high school students and assisted in a school for the intellectually disabled
- Service and Community Outreach**
- Centre County Women’s Resource Center** (2018 – present)
 - Counselor/Advocate
 - Answer domestic violence, sexual assault, and stalking hotline
 - Provide appropriate resources and emotional support and aid in creating safety and escape plans
 - Advocate for victims during forensic sexual assault examinations
 - Assist victims obtain various protection orders
 - Peers Helping Reaffirm, Educate, and Empower (PHREE)** (2017-present)
 - Executive board, Event Planning Chair: design and organize educational programs in cooperation with the PSU’s Gender Equity Center
 - On Site Coordinator for the Red Cross Blood Drives** (2017 - present)
 - PSU Dance MaraTHON, “Operations” Committee member** (Oct. 2016 – Feb. 2017, Oct. 2017 – Feb. 2018)
 - Hershey High School Mini-THON: Overall Chair** (2014 – 2015)
 - Organized Mini-THON and fundraising events
 - Key Club: Vice President** (2014 – 2015)
 - Organized and performed community service and outreach
- Employment**
- The Corner Room** (2018 – present)
 - Wait Staff
 - Transition Partner** for the Intensive English Communication Program (2017-present)
 - Assist international students in their transition to the United States and Penn State
 - Exam grader** for PSU’s general chemistry II course (2016 – 2017)
 - Employed by Dr. Raymond Schaak
 - Derry Psychiatric Services, Harrisburg, PA** (6/12 – 8/14)
 - Secretarial duties including maintaining patient records, checking in patients, and managing patient scheduling
 - Get the Picture Corporation** (6/12 – 8/13)
 - Sales Associate