

THE PENNSYLVANIA STATE UNIVERSITY
SCHREYER HONORS COLLEGE

DEPARTMENT OF BIOLOGY

THE IMPACT OF AN ENDOGENOUS RETROVIRUS ON MULE DEER ISY1 GENE
EXPRESSION

KALEB BOGALE
SPRING 2017

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 Poss
Professor of Biology and Professor of Veterinary and Biomedical Sciences
Thesis Advisor

Bernhard Lüscher
Professor of Biology and Professor of Biochemistry and Molecular Biology
Honors Adviser

* Signatures are on file in the Schreyer Honors College.

ABSTRACT

Endogenous retroviruses (ERVs) are retroviruses that infect germline cells and become permanent fixtures at a specific site in the host genome. In a continually evolving ERV-host relationship, ERVs may maintain the ability to completely restructure nearby host gene expression with their long terminal repeat (LTR). I am studying how an insertional polymorphic cervid ERV (CrERV), S2220, in *Odocoileus hemionus* (mule deer) impacts the nearby host pre-mRNA splicing ISY1 gene expression. I hypothesized CrERV S2220 restructured the ISY1 gene regulatory network and altered the transcript profile by acting as an alternative, bidirectional promoter of the ISY1 gene and CrERV S2220.

First, the prevalence and genotype of CrERV S2220 in 28 mule deer from Montana and Wyoming was investigated. We found CrERV S2220 was fixed and the most abundant state, comprising 25/28 total mule deer samples, in both Montana and Wyoming mule deer populations. Second, the host genome flanking the CrERV S2220 integration site was compared between mule deer from Montana or Wyoming and mule deer with or without CrERV S2220. The host genome near the integration site was conserved in mule deer.

Third, the predicted mule deer ISY1 gene model was generated through a comparative genomic analysis with close relatives, *Bos taurus* (domestic cow) and *Ovis aris* (sheep). Since no cervid genomes are available, we were the first to compare the gene structure of the cervid ISY1 gene (or any other cervid gene) and found the gene model is highly conserved in cervid genomes. Fourth, a qualitative transcript profile for the mule deer ISY1 gene was reported. Notably, we found the CrERV S2220 LTR is an alternative, bidirectional promoter of the ISY1 gene, as hypothesized. Two splicing variants in the mule deer ISY1 gene were found that initiated from the CrERV S2220 LTR. Future studies will identify if the CrERV S2220 LTR initiates a novel ISY1 gene transcript not found in the pre-integration site.

TABLE OF CONTENTS

LIST OF FIGURES	iii
LIST OF TABLES	iv
ACKNOWLEDGEMENTS	v
Chapter 1 Introduction	1
Infectious Retrovirus Structure	1
Retrovirus Lifecycle: Reverse Transcription of Retrovirus Genome	2
Retrovirus Lifecycle: Retroviral Genome Insertion into Host Genome	3
Endogenous Retroviruses.....	4
Endogenous Retroviruses Impact Host Gene Expression.....	4
Evolutionary Dynamics of Endogenous Retroviruses	6
<i>Odocoileus hemionus</i> (mule deer) and cervid ERVs	7
Pre-mRNA Splicing ISY1 Gene Structure and Transcript Profile	8
Research Question: Impact of CrERV S2220 on Mule Deer ISY1 Gene Expression.....	10
Chapter 2 Methods	11
Mule Deer Sample Collection & Genomic DNA Extraction ²⁰	11
CrERV S2220 Presence/Absence Screen	11
CrERV S2220 Population Structure	12
Mule Deer ISY1 Gene and Transcript Structure.....	13
Chapter 3 Results	14
CrERV S2220 Presence/Absence Screen	14
Predicted ISY1 Gene Model in Mule Deer	18
Chapter 4 Discussion	22
Appendix A Supplemental Information	25
BIBLIOGRAPHY	26

LIST OF FIGURES

Figure 1. Infectious Retrovirus Structure.	1
Figure 2. Formation of Long Terminal Repeats During Reverse Transcription. ⁷	2
Figure 3: Retroviral Genome Integration into Host Genome.	3
Figure 4: Mechanisms for ERV-Mediated Impact on Host Gene Expression ¹³	5
Figure 5: Insertional Activation of Oncogene Evi-1 by Endogenous Murine Leukemia Virus ⁶	6
Figure 6: Endogenous Retrovirus Integration Nearby pre-mRNA Splicing ISY1 Gene.....	8
Figure 7: Primary Transcript of ISY1 Gene in Sheep ²³	9
Figure 8: ISY1 Gene Transcript Profile in <i>Bos Taurus</i> from http://genome.ucsc.edu ²⁴	9
Figure 9: Primers S2220aF & S2220aR Flank CrERV S2220 Integration Site	12
Figure 10: Agarose Gel (1%) of CrERV S2220 Presence/Absence Screen	15
Figure 11: Distribution of CrERV S2220 in Montana and Wyoming Mule Deer.....	16
Figure 12: Predicted* Mule Deer Exon Profile of ISY1 Gene	19
Figure 13: RNA Transcription Sequencing of the Mule Deer ISY1 Gene	20
Figure 14: Observed Mule Deer ISY1 Gene Splice Variants	21

LIST OF TABLES

Table 1: Hardy-Weinberg Table for CrERV S2220 Occupied /Pre-Integration Site	17
Table 2: Hardy-Weinberg Table for CrERV S2220 Provirus/Solo LTR.....	18
Table 3. Primers Utilized for Genetic Sequencing Near CrERV S2220 Integration Site	25

ACKNOWLEDGEMENTS

A special thank you to my parents, Tadesse Bogale and Anketse Bogale, my brother, Daniel Bogale, my cousin, Elias Bogale, and my family. We have only grown closer since I left home for Penn State, and I whole-heartedly believe my successes are only a reflection of your tireless love and support. To Yisak Bogale and Meklit Bogale, you've both matured so much over these four years! I'm so thankful for your growth and encourage you both as you find your passions, whether its sports, science, or video games 😊

I am deeply appreciative of my research advisor, Dr. Mary Poss, and all of your personal mentorship over the past three years. You were more than just my research advisor, you've helped me mature as a person and scientist with your continued encouragement. Thank you for believing in me and challenging me to pursue The National Institutes of Health postbac program, even when I doubted myself.

I am very thankful for the day-to-day training and guidance throughout my time in the Poss lab, especially from Theodora Maravegias and Morgan Ferrell. Huge thanks for the rest of the Poss lab, including Lei Yang, Stephanie Williams, and former lab member Lan Nguyen.

I am grateful for all the opportunities through The Millennium Scholars Program, particularly from the director Star Sharp. You went above and beyond your expectations, making yourself available whenever called upon by myself or another MSP scholar. You personally set up the groundwork for the MSP program, and I am honored to be among the first cohort of millennium scholars.

Chapter 1

Introduction

Infectious Retrovirus Structure

Retroviruses, which infect all mammals and many other vertebrates, have a unique lifecycle beginning with their circulation as infectious retroviruses. The retrovirus structure contains an inner viral capsid surrounded by an outer membrane with envelope proteins (Figure 1). Within the viral capsid are two homologous strands of ribonucleic acid (RNA) in an association with a host tRNA that anchors the reverse transcriptase, a retroviral enzyme, to the retroviral genome.^{1,2} There are additional enzymes crucial for the retrovirus life cycle contained within the viral capsid, including integrase and proteases. The envelope proteins in the retrovirus outer membrane mediate the fusion and subsequent entry into a subset of host cells that are complementary. During fusion, the viral lipid bilayer combines with the host membrane, releasing both RNA strands, enzymes of the viral capsid (reverse transcriptase, integrase, and protease), and structural proteins into the host cell cytoplasm.^{1,3}

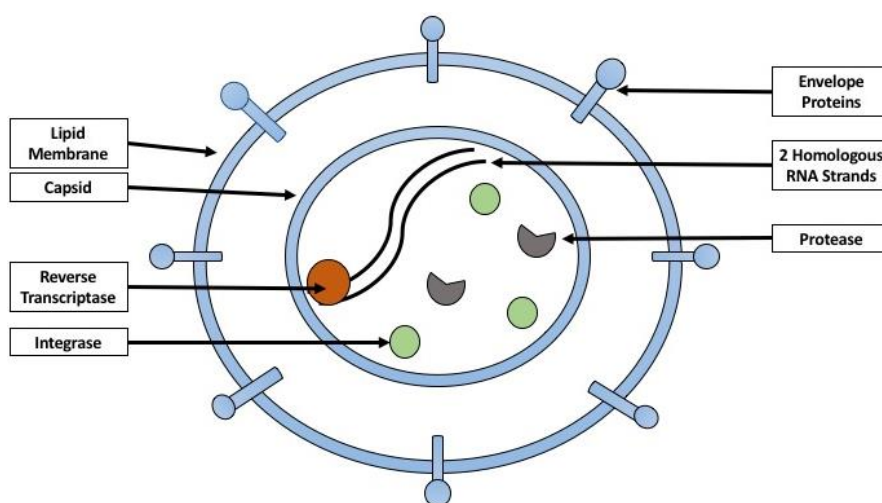


Figure 1. Infectious Retrovirus Structure.

Inside the viral capsid, both RNA strands are in an association with reverse transcriptase. The viral capsid is surrounded by an outer membrane that contains envelope proteins necessary for the fusion and subsequent entry into host cells. Upon entry into a host cell, enzymes of the viral capsid and structural proteins enter the host cell cytoplasm.

Retrovirus Lifecycle: Reverse Transcription of Retrovirus Genome

The two homologous RNA strands maintain their association with reverse transcriptase inside the host cytoplasm. In the host cytoplasm, reverse transcriptase subsequently utilizes host tRNA to catalyze the conversion of the RNA genome into double-stranded deoxyribonucleic acid (dsDNA) in a process called reverse transcription. Reverse transcriptase may move between the two homologous RNA strands during reverse transcription. This template switching during DNA synthesis is the principal source of genetic recombination in infectious retroviruses. In addition, reverse transcriptase has poor proof reading activity, causing an increased frequency of errors during DNA synthesis.^{1, 3, 4}

Reverse transcription also generates two identical long terminal repeats (LTRs) that flank the viral DNA not present in viral RNA (Figure 2). The 5' and 3' LTR, labelled in Figure 2, act as the viral promoter, enhancer, and transcription start and end sites. The 5' LTR normally acts as the viral promoter and transcription start site by binding host transcription factors, while the 3' LTR typically indicates the transcription termination site.^{1, 3, 5} The 3' LTR maintains viral promoter and enhancer ability that is hypothesized to be obstructed by the activity of the 5' LTR.⁶

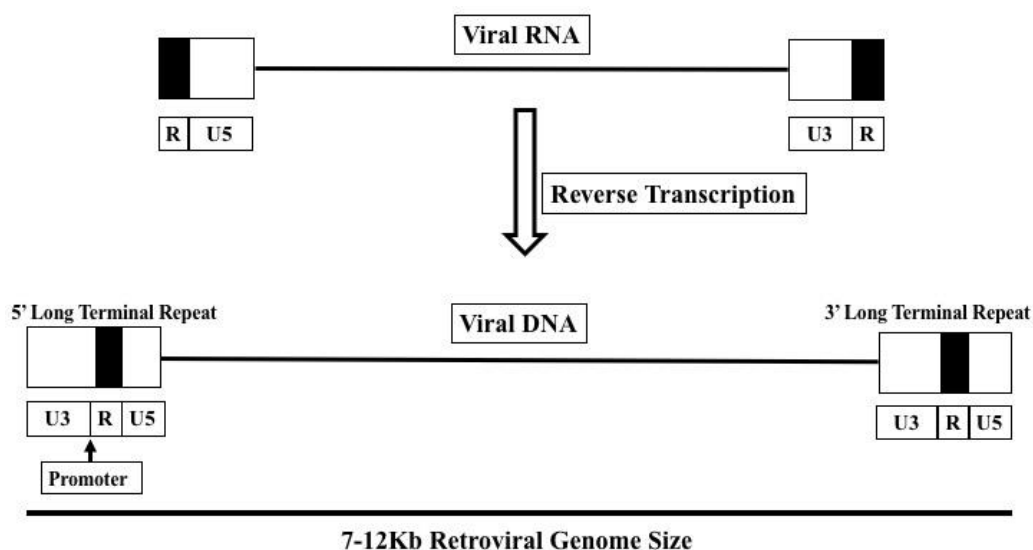


Figure 2. Formation of Long Terminal Repeats During Reverse Transcription.⁷

During reverse transcription, a retroviral enzyme (reverse transcriptase) catalyzes the conversion of the RNA genome (typically within 7-12Kb size) into double-stranded DNA. This also generates two identical long terminal repeats that flank the viral DNA, which aren't present in the viral RNA genome.

Retrovirus Lifecycle: Retroviral Genome Insertion into Host Genome

The retroviral genome is transported into the nucleus and integrated into the host genome with a retroviral enzyme, integrase. When integrase forms an association with the host genome, both terminal LTRs are attached, as shown on Figure 3. Integrase cuts the host genome leaving 5' overhangs to form the insertion site for the retroviral genome. The retrovirus is called a provirus after integration into the host genome. Host DNA repair enzymes fill the overhang leaving identical sequences flanking the virus integration, called target site duplications. The integrated retroviral genome and target site duplications are both structural variants in the host genome (Figure 3).^{1, 3, 8}

Retroviruses have integration preferences near host chromosomal regions that contain actively expressed genes and transcription start sites. They exhibit decreased integration near regions with condensed chromatin structure and long intergenic regions.⁹ Similar to cellular genes, once integrated retroviruses utilize their LTR promoter to recruit host RNA Polymerase II to catalyze its RNA transcription. The generation of RNA transcripts, some of which are translated into proteins, is critical to form new retroviruses. In a highly regulated process, the retrovirus life cycle is completed when new retroviruses assemble and acquire the host lipid bilayer upon exiting the cell.^{1, 3}

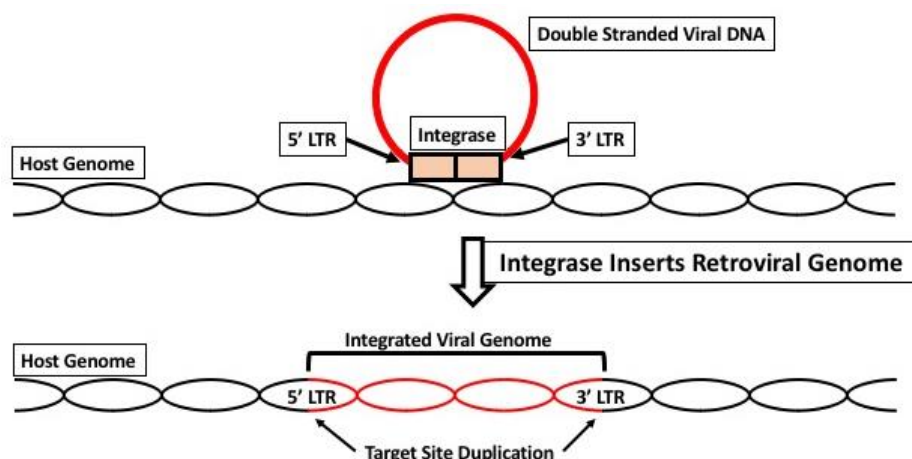


Figure 3: Retroviral Genome Integration into Host Genome.

The retroviral genome is transported into the host nucleus and integrated into the host genome with a retroviral enzyme, integrase. Integrase cuts the host genome leaving 5' overhangs to form the insertion site for the retroviral genome. Host DNA repair enzymes fill the overhang leaving identical sequences flanking the virus integration, called target site duplications.

Endogenous Retroviruses

Although retroviruses usually infect somatic cells, when a retrovirus infects a germline cell it is passed to the host progeny as an endogenous retrovirus (ERV). ERVs no longer retain infectious activity if the accumulation of epigenetic modifications and mutations make envelope encoded activity dysfunctional. The frequency of an ERV may increase in the host population over multiple generations if the ERV carrier maintains the ability to produce viable progeny.^{10, 11}

ERVs essentially function the same as other host genes. Both ERVs and host genes are present in every host cell and occupy a permanent fixture in a specific host genome location. They both contain a transcriptional promoter and have similar rates of genetic mutation. Typically, ERVs are silenced by epigenetic modifications and do not produce all components of functional retroviruses, although some retain transcriptional activity and produce RNA that may be translated into proteins.^{11, 12} Transcriptionally active ERVs are mobile genetic elements through an RNA intermediate, maintaining their original fixture in the genome while integrating a duplicate at another site of the genome. ERV transposons tend to integrate closer to chromosomal regions with expressed/active genes and transcription start sites.^{10, 12}

Endogenous Retroviruses Impact Host Gene Expression

ERVs can either contribute to the activation or deactivation (insertional mutagenesis) of nearby host genes. ERV LTRs can be primary promoters, alternative promoters, secondary promoters, or enhancers of the host gene (Figure 4). The LTRs may enhance the native promoter activity of host cellular genes over large distances via chromatin loop interaction, sometimes exhibiting tissues-specificity.^{13, 14} When ERV LTRs are secondary or alternative promoters, they may provide a secondary transcription start site for host genes.^{6, 13} The LTRs may also be bidirectional promoters, meaning they can initiate transcription in both directions; sometimes initiating both an ERV and host gene.⁵

ERVs can also render host genes dysfunctional after integrating within a host gene or its promoter region (Figure 4). The ERV integration may disrupt the promoter and prevent the binding of host transcription factors, which prevents any RNA transcript production. The ERV integration may also alter the transcript reading frame or signal early termination of transcripts, which form transcripts encoding unstable mutant peptides or truncated transcripts respectively.^{6, 13}

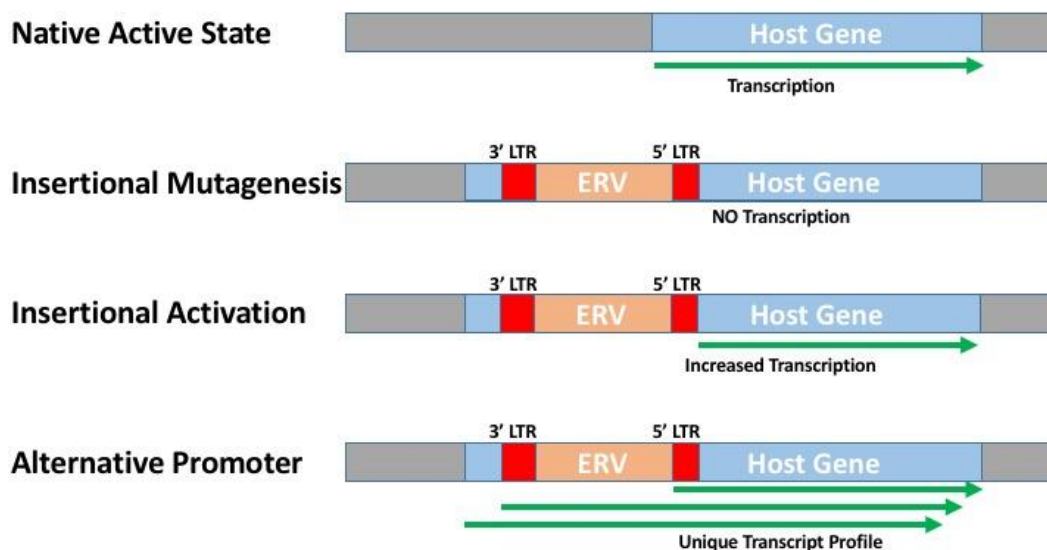


Figure 4: Mechanisms for ERV-Mediated Impact on Host Gene Expression¹³

Endogenous retroviruses have the potential to impact the expression of nearby cellular genes. When the ERV insertion prevents the transcription of a host gene, it is termed insertional mutagenesis. The long terminal repeat of an ERV may act as an enhancer or alternative promoter of cellular genes. ERVs acting as an alternative promoter may provide a secondary transcription start site for host genes.

One instance of ERVs contributing to disease is when ERVs dysregulate host oncogenes, causing breakdown of regulation with proper control signals or expression in the wrong cell types. For example, the 3' LTR activation of endogenous forms of Murine Leukemia Virus in mice triggers insertional activation of the host oncogene Evi-1 gene (Figure 5). Notably, elevated Evi-1 activation is not present in all individuals with this ERV, only those that have their 5' LTR inactivated subsequently releasing repression of 3' LTR promoter activity.⁶ In addition to functioning as the Evi-1 gene promoter, the ERV initiates a novel transcript not found in the native state (Figure 5). Interestingly, those transcripts initiated from the LTR also contain intronic regions under normal cellular conditions. Elevated expression of

oncogene Evi-1 is found in several hematopoietic malignancies and negatively effects terminal differentiation of hematopoietic cells via repression of transcription.^{15, 16, 17}

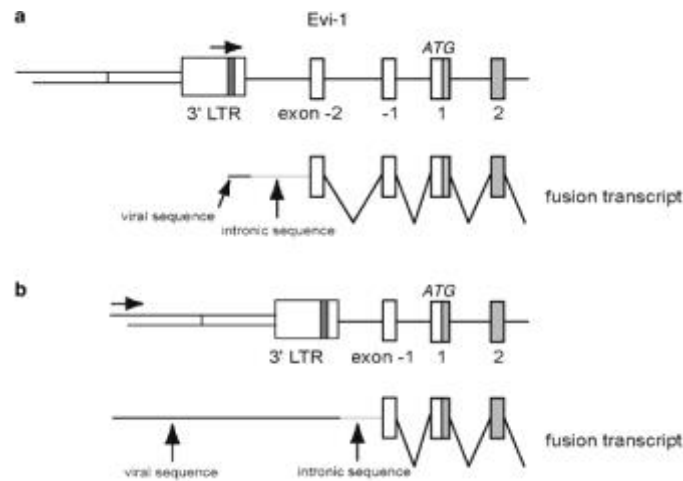


Figure 5: Insertional Activation of Oncogene Evi-1 by Endogenous Murine Leukemia Virus⁶

Endogenous forms of the murine leukemia virus may activate the Evi-1 proto-oncogene. Insertions are in the sense orientation either upstream (a) or within the first exons (b) of the Evi-1 gene.⁶ This results in a novel fusion transcript that translates to a normal, wild-type protein. Exons are drawn as boxes (coding sequences in grey) and transcriptional and translational start site of the Evi-1 gene is indicated (horizontal arrow and ATG, respectively).⁶

Evolutionary Dynamics of Endogenous Retroviruses

Any ERV may become fixed in a host population over time, with subsequent selection for those that are least harmful or even beneficial.¹⁰ The host can benefit from ERV-mediated host gene regulation changes in a process called host exaptation. For example, host exaptation of ERVs is critical for proper placenta development and has occurred independently in multiple mammalian orders. Human ERV-W, which integrated between 25 and 40 million years ago in higher primates, translates a retroviral envelope protein syncytin. Syncytin exhibits placenta-specific expression, enabling tissue-specific fusion of cytotrophoblasts to form a protective, multinucleated cell layer critical for early human placental development.¹⁸

The host may excise the entire integrated viral genome leaving a single LTR, in a process called solo LTR formation. There is no mechanism to remove only the ERV without leaving behind a LTR (solo LTR formation) or deleting chromosomal DNA.¹⁹ Importantly, solo LTRs maintain the ability to alter

host gene expression and restructure existing host gene regulatory patterns. They may be secondary promoters, enhancers, or alternative promoters of a host gene; sometimes signaling a unique transcription start/stop site from the native state.^{10, 13}

***Odocoileus hemionus* (mule deer) and cervid ERVs**

Odocoileus hemionus (mule deer) have occupied a wide diversity of habitats throughout the species history. Today, large populations extend across western North America from northern Mexico to Canada. Even with such a large geographic range, mule deer exhibit low levels of population divergence, which is the accumulation of genetic differences between different populations of the same species. Their large population sizes, long generation times, and high mobility over large distances likely cause their low levels of population divergence. After the speciation event of mule deer from their sister species, white-tailed deer (*O. virginianus*), mule deer have undergone repeated ERV integrations by gammaretroviruses.²⁰

The recent acquisition of cervid endogenous retroviruses (CrERVs) in mule deer exhibits a high degree of insertional polymorphism, meaning some mule deer acquired these CrERVs while others have not.²⁰ One of these CrERVs, S2220, integrated beside a mule deer pre-mRNA splicing ISY1 gene in an antisense orientation; meaning the CrERV is transcribed in the direction opposite to that of the host ISY1 gene (Figure 6). In this study, we report two splice variants of a host gene initiated by the LTR of CrERV S2220. The integration of CrERV S2220 near a host gene serves as a model for other ERV integrations. Since CrERV S2220 exhibits insertional polymorphism, mule deer have three potential homozygous genetic states; the pre-integration site (State 1), CrERV S2220 provirus (State 2), or CrERV S2220 solo LTR (State 3) (Figure 6). Mule deer may also form any combination of these genetic states if heterozygous; State 1 & 2, State 1 & 3, or State 2 & 3.

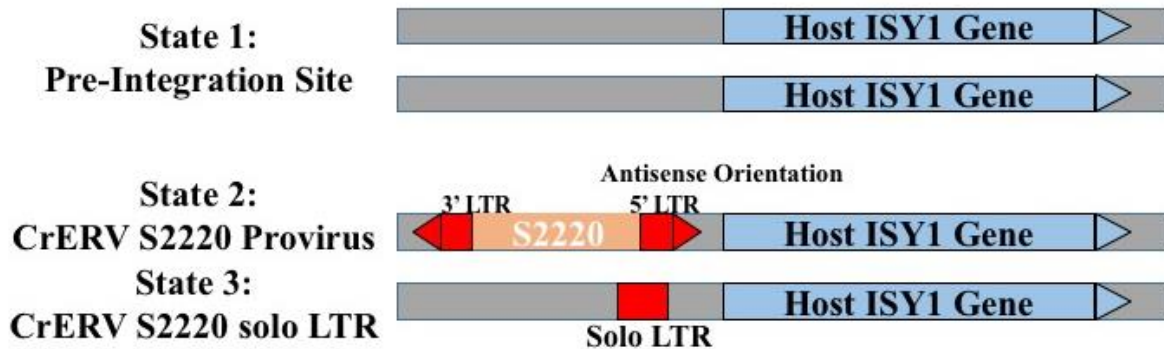


Figure 6: Endogenous Retrovirus Integration Nearby pre-mRNA Splicing ISY1 Gene

Mule deer exhibited three potential states for the CrERV S2220; the pre-integration site (State 1), CrERV S2220 provirus (State 2), or CrERV S2220 solo LTR (State 3). CrERV S2220 integrated beside a mule deer pre-mRNA splicing ISY1 gene in an antisense orientation; meaning the CrERV is transcribed in the direction opposite to that of the host ISY1 gene. The direction of transcription is indicated with an arrow.

Pre-mRNA Splicing ISY1 Gene Structure and Transcript Profile

The pre-mRNA splicing ISY1 gene encodes a peptide regulator of key proteins in the pre-mRNA splicing network, termed the spliceosome. This pre-mRNA splicing network is critical for the removal of introns from transcripts of eukaryotic cells.^{21, 22} The activity of the ISY1 gene is conserved in mammals and has 72 orthologues in eukaryotes. Since the mule deer ISY1 transcript profile is not known and will be investigated in this study, close relatives *Ovis Aries* (sheep) and *Bos Taurus* (domestic cow) ISY1 gene structure will be described.^{23, 24}

In sheep, the ISY1 gene has only one reported 906bp transcript that encodes a 302 amino acid residue protein. In Figure 7, the ISY1 gene (labelled green) displays the 12 coding exons (labelled red) that are joined to form a 906bp primary transcript.²³ In cow, the ISY1 gene has only one reported 1342bp transcript that encodes a 284 amino acid residue. In Figure 8, the ISY1 gene (labelled blue) displays the 11 coding exons (rectangles) that form the 1342bp primary transcript.²³ The remaining black rectangles and dashes designate exons and introns, respectively, from all cow ISY1 gene transcripts. Even though only one ISY1 gene transcript has been annotated, the transcript profile shows potential ISY1 gene

isoforms not yet annotated (Figure 8). Many of the sequences are truncated forms of the primary transcript, although a consistent potential isoform may splice exons 1 and 2.²⁴

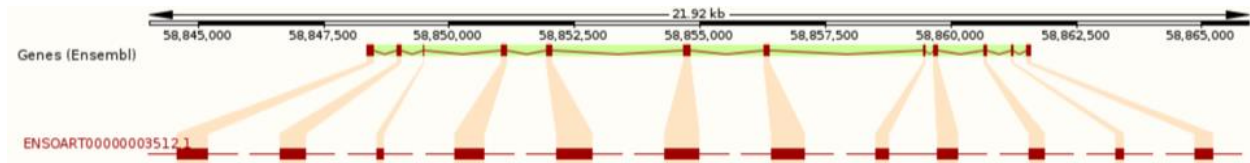


Figure 7: Primary Transcript of ISY1 Gene in Sheep²³

The sheep ISY1 gene model contains 12 coding exons that are joined to form a 906bp transcript. The primary ISY1 gene transcript contains all 12 coding exons as shown below the ISY1 gene model. The sheep ISY1 gene model (in green) displays the 12 coding exons as red rectangles and introns as red lines separating them.

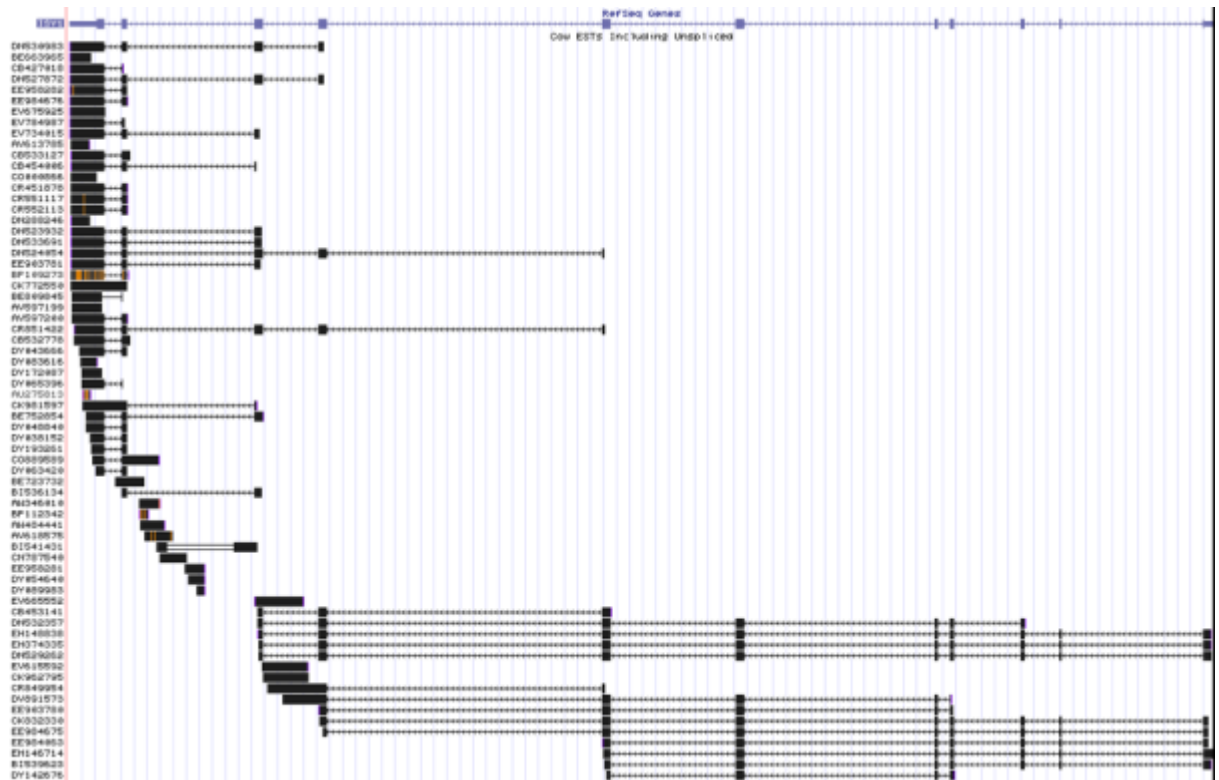


Figure 8: ISY1 Gene Transcript Profile in *Bos Taurus* from <http://genome.ucsc.edu>²⁴

The cow ISY1 gene model contains 11 coding exons that are joined to form a 1342bp transcript. The cow ISY1 gene model (in blue) displays the 11 coding exons as rectangles and introns as blue dashes separating them. The remaining black rectangles and dashes designate exons and introns, respectively, from partial cow ISY1 gene transcripts.

Research Question: Impact of CrERV S2220 on Mule Deer ISY1 Gene Expression

The aim of my research project is to determine the impact of CrERV S2220 on mule deer ISY1 gene expression. The null hypothesis is that there is no involvement of CrERV S2220 in the qualitative transcription profile of the ISY1 gene. An alternative hypothesis is the CrERV S2220 is an alternative, bidirectional promoter that initiates transcription of the ISY1 gene and alters its transcript profile. First, the prevalence and genotype of CrERV S2220 in 28 mule deer from Montana and Wyoming was determined. This screen determined the prevalence of different CrERV S2220 states in mule deer populations, and identified any selective pressures for any of the 3 homozygous states (State 1-3) or the 3 heterozygous states (State 1 & 2, State 1 & 3, or State 2 & 3), that differ between Montana and Wyoming mule deer.

Second, the host genome near the CrERV S2220 integration site was sequenced in the pre-integration site (State 1) and compared with mule deer from Montana or Wyoming and mule deer with or without CrERV S2220. This identifies if geographic location of the mule deer population or CrERV S2220 integration causes genetic sequence variation near the integration site. The final objective was to determine if the CrERV S2220 LTR is a promoter for the ISY1 gene. We confirmed our hypothesis that CrERV S2220 is an alternative, bidirectional promoter of the ISY1 gene and initiates two ISY1 splice variants.

Chapter 2

Methods

Mule Deer Sample Collection & Genomic DNA Extraction²⁰

28 mule deer retropharyngeal lymph node samples were utilized in our study from harvested mule deer brought through hunter check stations within Montana and Wyoming, as described by *Kamath et al.*²⁰ Genomic DNA was extracted from lymph node tissue using a phenol/chloroform extraction method.

CrERV S2220 Presence/Absence Screen

Extracted genomic DNA undergoes a polymerase chain reaction (PCR) with forward primer S2220aF and reverse primer S2220aR, which flanks the CrERV S2220 integration site (Figure 9). The expected DNA fragment length is 979bp for the pre-integration site, 1.5Kb for CrERV S2220 solo LTR, and 10Kb for CrERV S2220 integration. PCR conditions: 3-minute at 95°C for initial denaturation, 36 cycles of (30s at 94°C) for denaturation, (30s at 62°C) for annealing, (8 min at 72°C) for extension, and (10 min for 72°C) for the final extension.

25µL PCR reactions were performed with 12.5µL 2X GoTaq long master mix, 1.0µL S2220aF (10µM), 1.0µL S2220aR (10µM), 9.5µL nuclease-free water, and 1.0µL mule deer gDNA template (50 ng/µL). The PCR product was separated based on size by agarose (1%) gel electrophoresis. Prior to DNA sequencing, PCR products were purified by gel isolation using molecular biology grade agarose.⁸

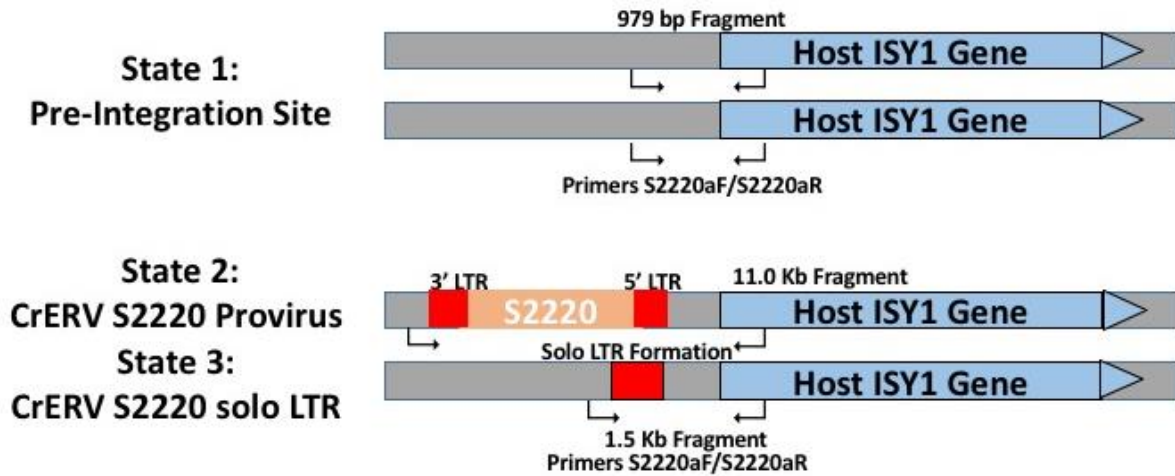


Figure 9: Primers S2220aF & S2220aR Flank CrERV S2220 Integration Site

Mule deer exhibited three potential states for CrERV S2220; the pre-integration site (State 1), the complete provirus (State 2), or its solo LTR (State 3). Primers S2220aF and S2220aR flank the CrERV S2220 integration site yielding a 979bp, 11Kb, or 1.5Kb size product for State 1, State 2, and State 3, respectively. The arrows indicate the location of primers S2220aF and S2220aR.

CrERV S2220 Population Structure

We determined the Hardy-Weinberg statistic of CrERV S2220 for all mule deer to identify CrERV S2220 frequency changes in the population. We stratified by geographic region (Montana or Wyoming) to identify any geographic difference in CrERV S2220 frequencies. Chi-squared statistical analysis was completed at a 99% confidence threshold.²⁵

Expected Genotype Frequencies of CrERV S2220:

$$\text{Frequency of } q: \left(\frac{\# \text{ mule deer } (q^2)}{\# \text{ total mule deer}} \right) + \left(\frac{\# \text{ mule deer } (2pq)}{2 * \# \text{ total mule deer}} \right), \text{ Frequency of } p: 1 - q$$

$$p^2 * 28 (\text{total mule deer}) = \text{expected number } p^2 \text{ mule deer}$$

$$2pq * 28 (\text{total mule deer}) = \text{expected number } 2pq \text{ mule deer}$$

$$q^2 * 28 (\text{total mule deer}) = \text{expected number } q^2 \text{ mule deer}$$

$$\text{Hardy-Weinberg Statistic Calculation} = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

O = total number of observed genotype, E = expected total number of genotype

Mule Deer ISY1 Gene and Transcript Structure

The host ISY1 gene flanking the CrERV S2220 integration site was amplified by PCR and the sequence of mule deer from Montana or Wyoming with CrERV S2220 were compared. Prior to sequencing, PCR products were purified with gel isolation using molecular biology grade agarose.⁸ A predicted mule deer exon structure of the ISY1 gene was generated through a comparative genomic analysis with close relatives, *Bos taurus* (domestic cow) and *Ovis aris* (sheep). RNA-seq generated the qualitative ISY1 gene transcript profile for the mule deer after the generation of cDNA from preserved mule deer RNA, as described.²⁶

Chapter 3

Results

CrERV S2220 Presence/Absence Screen

We conducted a CrERV S2220 screen in 28 mule deer samples to identify their genotype and determine if CrERV S2220 is fixed in the mule deer populations. The results of the CrERV S2220 presence/absence screen demonstrated that there were only two of the possible six genetic states present (Figure 10). Mule deer without CrERV S2220 (State 1) exhibited the pre-integration site with a 979bp DNA band, as shown in Sample 1 (Figure 10). Mule deer with CrERV S2220 always exhibited the full-length CrERV S2220 provirus (10Kb) and its solo-LTR (1.5Kb) (State 2/3), as shown in Sample 2 (Figure 10). No mule deer screened were homozygous for the CrERV S2220 provirus (State 2) or the CrERV S2220 solo LTR (State 3). No mule deer were heterozygous for the CrERV S2220 provirus and the pre-integration site (State 2/1) or CrERV S2220 solo LTR and the pre-integration site (State 3/1).

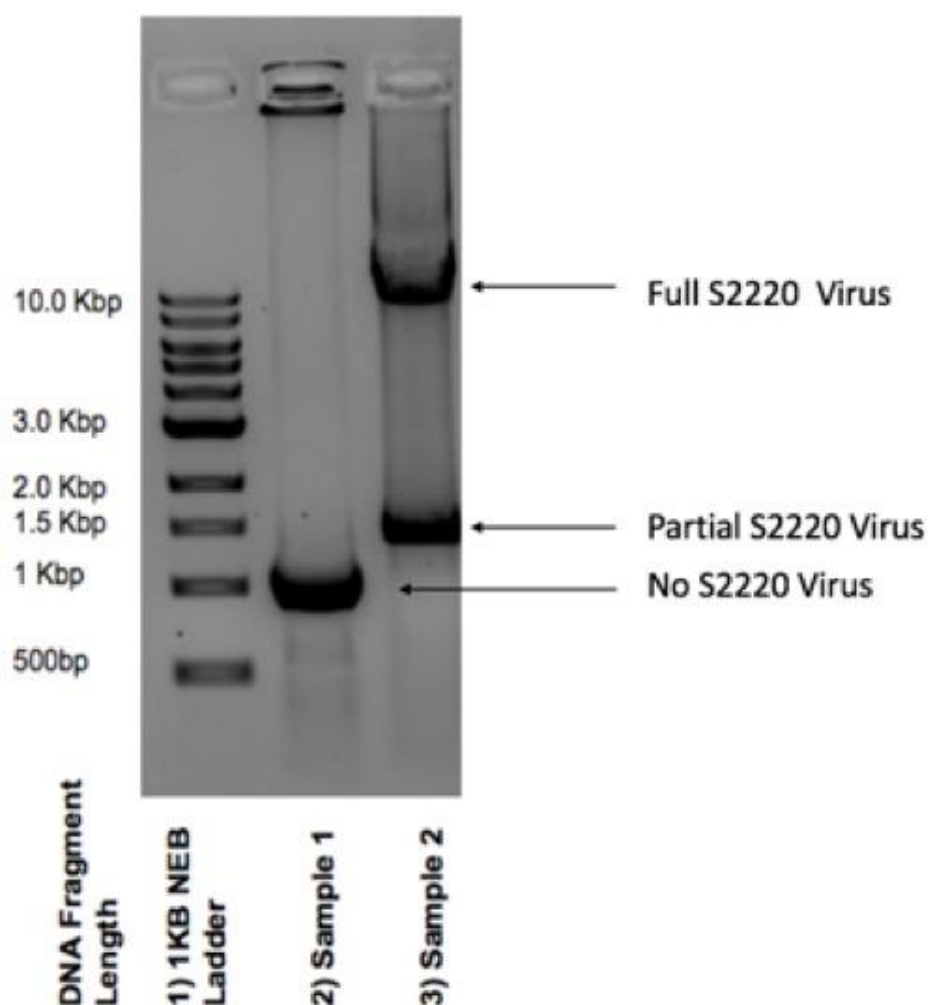


Figure 10: Agarose Gel (1%) of CrERV S2220 Presence/Absence Screen

Mule deer exhibited two potential genotypes for CrERV S2220; only the pre-integration site (Sample 1) or the CrERV S2220 provirus-solo LTR (Sample 2). The PCR products of the CrERV S2220 presence/absence screen are shown beside a reference 1Kb NEB ladder.

The results of the CrERV S2220 presence/absence screen show that CrERV S2220 was fixed in the mule deer population. Montana and Wyoming mule deer show a similar pattern of the provirus-solo LTR (State 2/3) being the only genotype other than the pre-integration site (Figure 11). Montana mule deer exhibited more of CrERV S2220 provirus-solo LTR (17/20) than the pre-integration site (3/20). Wyoming mule deer only exhibited the CrERV S2220 provirus-solo LTR (8/8) with no mule deer containing the pre-integration site.

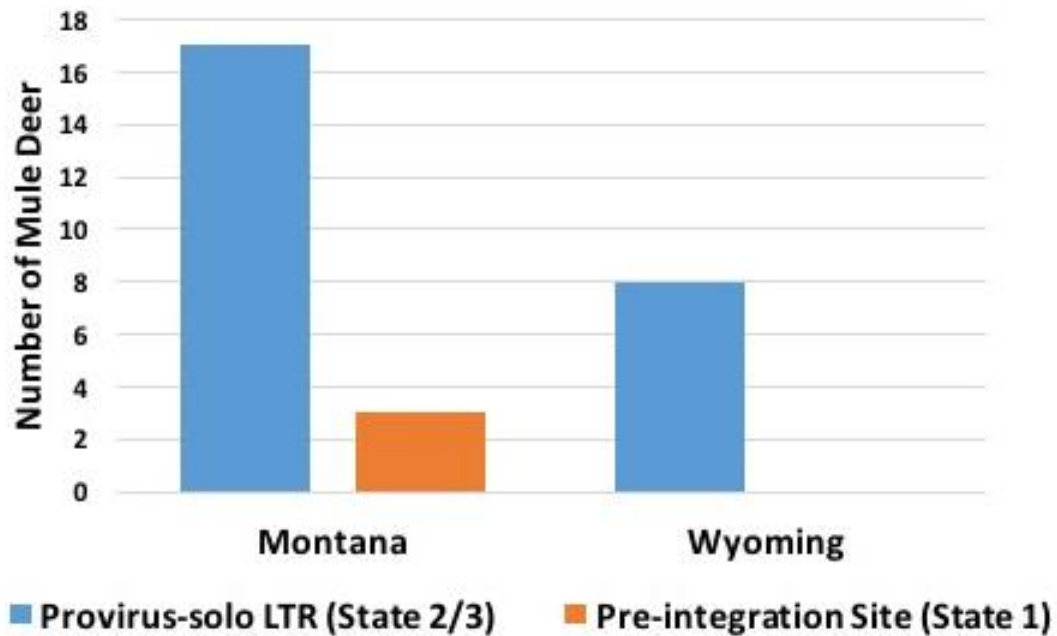


Figure 11: Distribution of CrERV S2220 in Montana and Wyoming Mule Deer

The results of the CrERV S2220 presence/absence screen were presented for Montana and Wyoming mule deer. Mule deer with the provirus-solo LTR genotype (in blue) were compared to mule deer with the pre-integration site (in orange).

We conducted Hardy-Weinberg equilibrium tests to identify any selective pressures for the CrERV S2220 occupied state, defined as the provirus or its solo LTR, or the pre-integration site (State 1). There were 25 mule deer with the occupied state, which was larger than the expected 22 mule deer samples (Table 1a). Since no mule deer with an occupied state (State 2/3) also contained the pre-integration site (State 1), this deviates from our expected 5.5 mule deer samples. Only 3 mule deer only contained the pre-integration site, which was larger than the expected 0.3 mule deer samples. The calculated Hardy-Weinberg value, 29.9, identifies that the occupied state (State 2 or 3) and the pre-integration site are not under Hardy-Weinburg equilibrium with >99% confidence. When stratified by geographic location, Montana mule deer exhibit a similar Hardy-Weinberg distribution as the combined mule deer population (Table 1b). The calculated Hardy-Weinberg value, 18.4, identifies that the Montana mule deer population is also not under Hardy-Weinburg equilibrium with >99% confidence.

a.	CrERV S2220 Genotype	Observed Mule Deer	Expected Mule Deer	HWE Value
	Occupied Site (State 2 or 3)	25	22	0.41
	Occupied/Pre-Integration Site (State 2 or 3 & 1)	0	5.5	5.5
	Pre-integration Site (State 1)	3	0.3	24
	Total			29.9

b.	CrERV S2220 Genotype	Observed Mule Deer	Expected Mule Deer	HWE Value
	Occupied Site (State 2 or 3)	17	14.5	0.43
	Occupied/Pre-Integration Site (State 2 or 3 & 1)	0	5	5
	Pre-integration Site (State 1)	3	0.5	13
	Total			18.4

Table 1: Hardy-Weinberg Table for CrERV S2220 Occupied /Pre-Integration Site

We generated Hardy-Weinberg tables to identify any selective pressures for the CrERV S2220 occupied state, defined as the provirus or its solo LTR, or the pre-integration site (State 1). Hardy-Weinberg tables were generated for the total (a) mule deer population and the Montana (b) mule deer population.

Of those mule deer with the occupied state (State 2 or 3), we conducted Hardy-Weinberg equilibrium tests to identify any selective pressures for the CrERV provirus or its solo LTR. Since no mule deer screened were homozygous for the CrERV S2220 provirus (State 2) or its solo LTR (State 3), there was less than the expected 6 mule deer samples (Table 2a). Every mule deer sample with the occupied state exhibited the provirus-solo LTR (State 2/3), which was more than the expected 13 mule deer samples. The calculated Hardy-Weinberg value, 23, identifies that the CrERV S2220 provirus and its solo LTR are not under Hardy-Weinburg equilibrium with >99% confidence. When stratified by geographic location, Montana and Wyoming mule deer populations exhibit a similar Hardy-Weinberg distribution as the combined mule deer population (Table 2b and Table 2c, respectively); still not under Hardy-Weinburg equilibrium with >99% confidence.

a.	CrERV S2220 Genotype	Observed Mule Deer	Expected Mule Deer	HWE Value
	Full CrERV S2220 (State 2)	0	6	6
	Provirus-solo LTR (State 2/3)	25	13	11
	CrERV S2220 solo LTR (State 3)	0	6	6
	Total			23

b.	CrERV S2220 Genotype	Observed Mule Deer	Expected Mule Deer	HWE Value
	Full CrERV S2220 (State 2)	0	4	4
	Provirus-solo LTR (State 2/3)	17	9	7
	CrERV S2220 solo LTR (State 3)	0	4	4
	Total			15

c.	CrERV S2220 Genotype	Observed Mule Deer	Expected Mule Deer	HWE Value
	Full CrERV S2220 (State 2)	0	2	2
	Provirus-solo LTR (State 2/3)	8	4	4
	CrERV S2220 solo LTR (State 3)	0	2	2
	Total			8.0

Table 2: Hardy-Weinberg Table for CrERV S2220 Provirus/Solo LTR

Of those mule deer with the occupied state (State 2 or 3), we conducted Hardy-Weinberg equilibrium tests to identify any selective pressures for the CrERV S2220 provirus or its solo LTR. We generated a Hardy-Weinberg table for the total (a), Montana (b), and Wyoming (c) mule deer populations.

Predicted ISY1 Gene Model in Mule Deer

The predicted ISY1 gene model in mule deer was generated through a comparative genomic analysis with close relatives, *Bos taurus* (domestic cow) and *Ovis aris* (sheep). Although the cow and sheep ISY1 gene exon structure was similar, there are several differences marked as * in Figure 12. The cow exon 1 begins 1779bp before the sheep exon 1 and contains the entire sheep exon 1 (60bp) with an additional 50bp. Cow and sheep exons 2-8 are homologous in location and sequence. The cow and sheep exon 9 has an identical start site and sequence except cow exon 9 has an additional 21bp on the 5' end. Sheep contain a unique exon 10 (21bp) not present in cow, which is predicted to be in the mule deer ISY1 gene model. The cow exon 10 and sheep exon 11 are homologous in location and sequence. Cow exon 11 and sheep exon 12 are homologous in location and sequence except an additional 6bp on the 5' end of the cow exon 11. They have an identical translation start and stop site at the beginning of exon 2 and the end of their final exon, respectively (Figure 12).

The predicted mule deer ISY1 gene model was a combination of the cow and sheep ISY1 gene exons. Mule deer are predicted to contain both the cow and sheep exon 1. Mule deer exons 3-8 are completely identical with cow and sheep exons 2-7, including the translational start site. Mule deer exon 9 is similar with cow and sheep exon 8 except it contains one amino acid substitution (H \Rightarrow Y). The end of mule deer exon 10 marks the translational stop site and contains multiple amino acid substitutions not found in cow or sheep exon 9 (Figure 12).

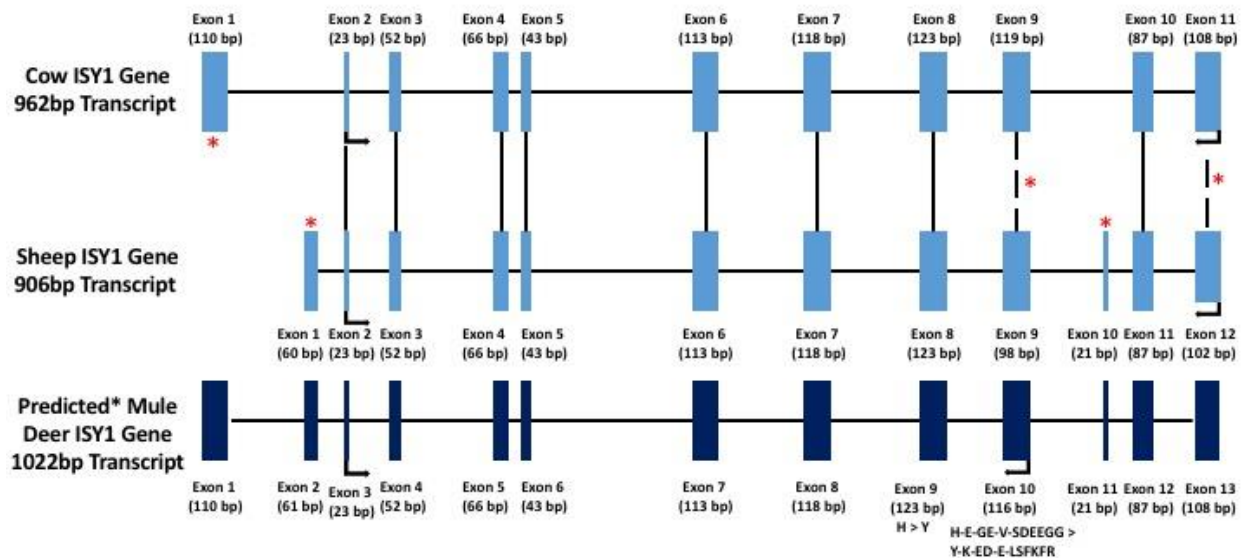


Figure 12: Predicted* Mule Deer Exon Profile of ISY1 Gene

The predicted ISY1 gene model in mule deer was generated through a comparative genomic analysis with close relatives, *Bos taurus* (domestic cow) and *Ovis aris* (sheep). The translational start and stop site of the ISY1 gene is indicated with horizontal arrows. The differences between the cow and sheep ISY1 gene model are marked as *. In mule deer, differences in the predicted ISY1 gene model impacting amino acid structure not found in cow or sheep are shown.

We conducted RNA-sequencing in mule deer to determine if the CrERV S2220 LTR initiates some ISY1 gene transcripts and identify potential isoforms of the ISY1 gene. After aligning these (~150bp) genetic fragments against our predicted mule deer ISY1 gene model, we identified the CrERV S2220 LTR initiates transcription in the direction of the ISY1 gene. There were no RNA transcripts that aligned to the native ISY1 gene promoter. There were multiple RNA transcripts spanning the CrERV S2220-host junction fragment, meaning they began in the CrERV S2220 LTR and ended in the host genome (Figure 13). There is no evidence for consecutive RNA transcription in exons 8 to 9 or exons 12

to 13. There was transcription in all mule deer exons, with the greatest transcript depth on exons 8, 10, and 12. This indicates exon 8, 10, and 12 are expected to be present in most ISY1 gene RNA transcripts.

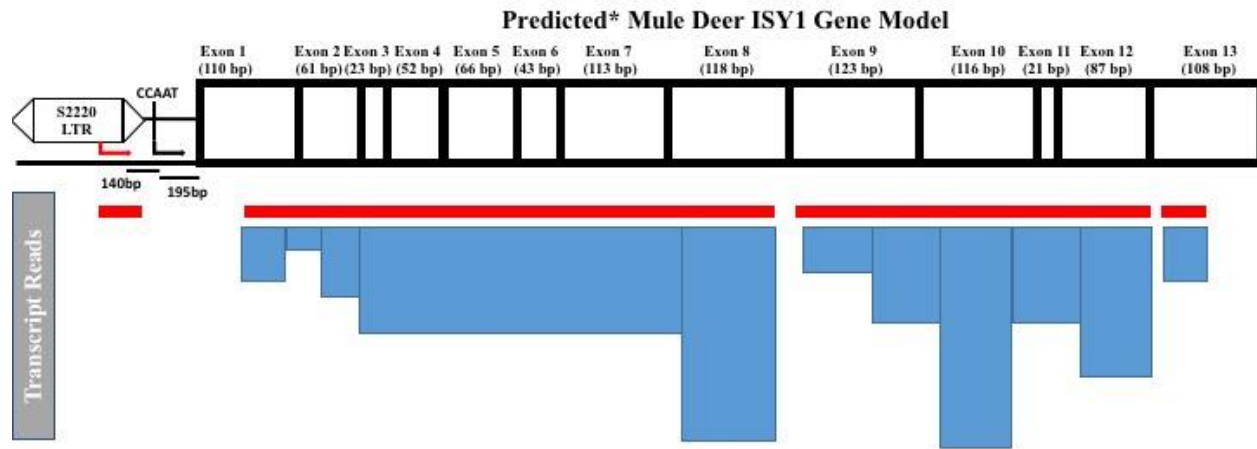


Figure 13: RNA Transcription Sequencing of the Mule Deer ISY1 Gene

The CrERV S2220 LTR is shown beside the predicted mule deer ISY1 gene model. The predicted and observed transcription start site are shown as black and red arrows, respectively. The CCAAT box marks the predicted native ISY1 gene transcription start site. The mule deer RNA transcript depth is shown beneath the corresponding region on the CrERV S2220 host-junction or the predicted ISY1 gene model. The red bar beneath the CrERV S2220 LTR and ISY1 gene model identifies all regions with any corresponding RNA transcripts.

We aligned the complete 150bp RNA transcripts initiated by the observed transcription start site in the CrERV S2220 LTR with the predicted mule deer ISY1 gene model (Figure 14). Since all of these RNA transcripts initiated from the CrERV S2220 LTR, the splicing patterns were identified when mule deer exons were present on the 5' end of those transcripts. The structure of mule deer ISY1 gene splice variant 1 and 2 are both shown in Figure 14. Both contain ISY1 gene exons 7, 8, and 12 although splice variant 1 initiates in the LTR splices directly to exon 5 while splice variant 2 splices directly to exon 7. There is no evidence for the inclusion of exon 11, which is unique to sheep and mule deer, in either ISY1 gene variant. Splice variant 2 also includes exon 10 that is not found in splice variant 1.

There were 14 total 150bp RNA transcripts that were ISY1 gene splice variant 1 or 2, indicated as blue lines in Figure 14. There are equivalent amounts of exon 8 and 10 containing RNA transcripts in splice variant 2, suggesting this splice form is dominant. Alternatively, ISY1 gene splice variant 1 only contains one RNA transcript spanning the entire transcript.

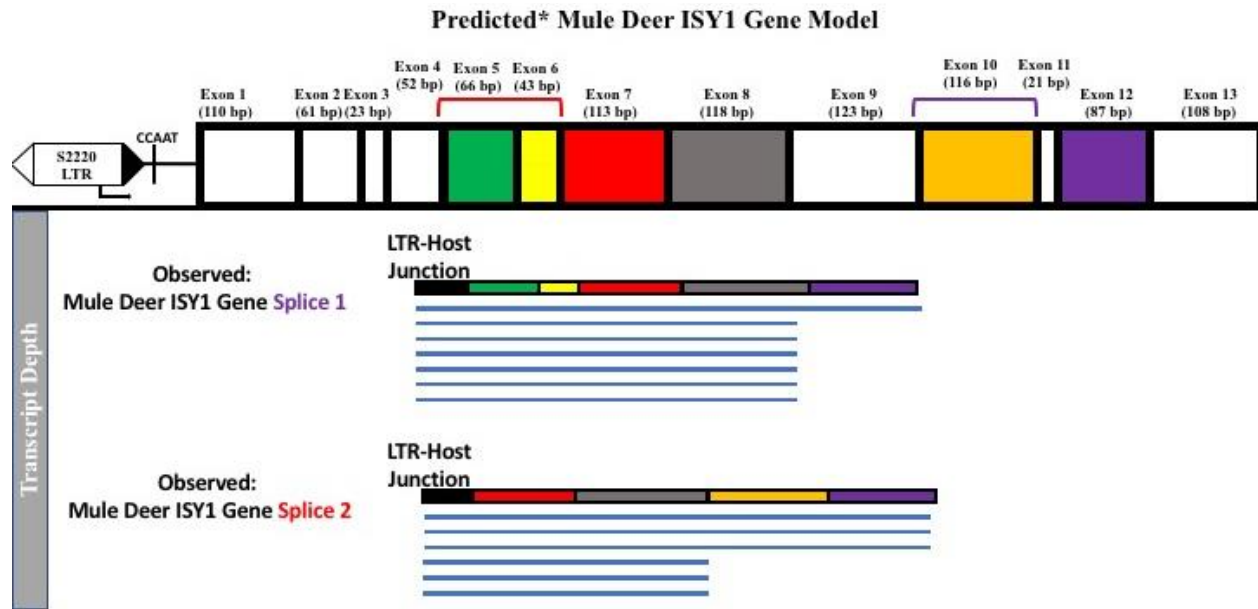


Figure 14: Observed Mule Deer ISY1 Gene Splice Variants

The CrERV S2220 LTR is shown beside the predicted mule deer ISY1 gene model. The black arrow identifies the transcription start site for all ISY1 gene RNA transcripts initiated from the CrERV S2220 LTR, while the CCAAT box identifies the native ISY1 gene transcription start site. There were two splicing variants found in RNA transcripts initiated from the CrERV S2220, labelled in purple and red respectively. The color of each splicing variant corresponds to the ISY1 gene exon contained, while the blue lines beneath represent an RNA transcript aligned to that region.

Chapter 4 Discussion

In this study, we reported the abundance and genotype of CrERV S2220 in Montana and Wyoming mule deer populations. The majority (25/28) of mule deer contained the occupied state (CrERV S2220 or its solo LTR), while only 3 of 28 screened mule deer contained the pre-integration site (Figure 11). To determine if the frequency of the occupied state was as predicted in the mule deer population, we conducted a Hardy-Weinberg equilibrium (HWE) analysis (Table 1). Our results found the CrERV S2220 occupied state/pre-integration site was not under HWE, meaning the frequency of the occupied state and pre-integration site are likely subject to other evolutionary forces.

The results of the CrERV S2220 presence/absence screen demonstrated that there were only two of the possible six genotypes present (Figure 10). Mule deer were either homozygous for the pre-integration site or contained CrERV S2220 provirus and its solo-LTR. Of the mule deer with the occupied state, we conducted an HWE test to identify any selective pressures for CrERV S2220 solo LTR formation (Table 2). Our results found the CrERV S2220 provirus and solo LTR formation were not under HWE, meaning their frequencies are changing in the mule deer population and likely subject to other evolutionary forces. The CrERV S2220 provirus/solo LTR state may be advantageous indicative of a strong balancing selection or the homozygous states (State 2 or 3) may be a lethal genotype.

CrERV S2220 integrated beside the host ISY1 gene in an antisense orientation, meaning it initiates CrERV transcription in the opposite direction of the host ISY1 gene. The CrERV S2220 LTR must be a bidirectional promoter to initiate host ISY1 gene transcription. Although human ERV LTRs have shown to activate transcription in either orientation in reporter assays, bidirectional promoters usually disrupt gene expression and may be excluded from gene annotation and thus underreported.^{27, 28, 29, 30, 31} The first experimentally verified example of a human gene pair controlled by a single ERV LTR promoter was the human ERV1 LTR. The ERV1 LTR acts as a bidirectional promoter for the human down syndrome critical region 4 (DSCR4) and DSCR8 genes. DSCR4 and DSCR8 include 4 splice variants, of which one novel splicing variant is initiated by the ERV1 LTR.²⁷

Since no cervid genomes are available, we were the first to compare the gene structure of the cervid ISY1 gene (or any other cervid gene). We found the ISY1 gene model is highly conserved in cervid genomes, with cow and sheep exons 2-8 and the predicted mule deer exons 3-9 being homologous in sequence and location (Figure 12). There are differences in the predicted mule deer exon structure that make the ISY1 gene amino acid sequence slightly different for mule deer, as shown on Figure 12. Mule deer exon 9 contains one amino acid substitution ($H \Rightarrow Y$), while mule deer exon 10 contains multiple amino acid substitutions not found in cow or sheep exon 9 ($H-E-GE-V-SDEEGG \Rightarrow Y-K-ED-E-LSFKFR$). One major difference in mule deer is the earlier translation stop site on exon 10, compared to cow and sheep that terminate translation at the end of their final exon. There was no evidence for mule deer ISY1 gene transcripts that contained exon 10 terminating at this translation stop site. The predicted mule deer ISY1 protein product to be a 217 amino acid residue, which is smaller than cow and sheep whose ISY1 gene has 284 and 302 amino acid residues, respectively. We have not confirmed if the ISY1 gene encodes functional proteins with the CrERV S2220 integration, although we are certain the ISY1 gene is actively transcribing RNA.

The predicted ISY1 gene transcription start site was at the CCAAT box, which is located between the CrERV S2220 LTR and the first predicted mule deer exon 1 (Figure 13). There were no RNA transcripts that aligned to the predicted native ISY1 gene promoter. We observed multiple RNA transcripts spanning the CrERV S2220-host junction fragment, which means they began in the CrERV S2220 LTR and end in the host genome (Figure 13). This indicates the TATA box in the LTR promoter is a stronger transcription start site than the native ISY1 gene promoter, so at least some RNA transcripts originate from the LTR while non originate in the predicted CCAAT box. There was a similar observation of ERV-initiated transcripts containing intronic portions in Evi-1 gene promoter (Figure 5). Our hypothesis that the CrERV S2220 LTR acts as an alternative, bidirectional promoter in mule deer ISY1 gene expression was confirmed. We identified two novel ISY1 gene splicing variants initiated from the CrERV S2220 LTR, which supports that a full ISY1 transcript (not just a non-coding RNA) originates in

the CrERV S2220 LTR (Figure 14). We suggest the impact of the CrERV S2220 LTR may confer a benefit to mule deer and is an example of host exaptation.

A limitation of our study was the sample size of 28 total mule deer within Montana and Wyoming. Increasing the sample size would give more accurate representations of the frequency of CrERV S2220 in the population, and increase chances of identifying mule deer homozygous for CrERV S2220 provirus or its solo LTR (if they exist). Other than increasing our mule deer sample size, the next logical experiment will identify if the CrERV S2220 LTR initiates a novel ISY1 gene transcript.

In our study, we identified the majority of mule deer contained the CrERV S2220 provirus and solo LTR, which is under balancing selection in the mule deer population. The mechanism for this balancing selection is likely due to CrERV S2220 LTR alternative, bidirectional promoter of the ISY1 gene. We were the first to compare the gene structure of the cervid ISY1 gene (or any other cervid gene), and future studies should will quantify the impact of the CrERV S2220 LTR on mule deer ISY1 gene transcription.

Appendix A

Supplemental Information

Primer Name	Primer Sequence
S2220_3kf	5' – GGC – TGC – TCC – CCT – TTT – GTC – 3'
S2220BF	5' – CTA – AGG – CTC – ACT – CTG – CAA – TGT – CTC – 3'
S2220dF	5' – ACG – TGG – GGG – TTA – ATC – AAA – GTC – TG – 3'
S2220fF	5' – TTC – CTG – GAT – GGG – TAA – ATG – TGC – TGA – 3'
S2220_HF	5' – GTG – GGT – GGA – GCC – GAC – TGA – CG – 3'
S2220GR	5' – ATG – GGG – AAG – CTG – AGC – CTG – AA – 3'
S2220ER	5' – TGG – CCT – CTC – AGT – CCT – GTC – CTA – CCC – 3'
S2220CR	5' – GTG – GCT – CGG – GAA – GGC – TAA – AAC – 3'
S2220LTRbR	5' – GCT – AGG – GCC – CCC – GCT – GTG – T – 3'

Table 3. Primers Utilized for Genetic Sequencing Near CrERV S2220 Integration Site

BIBLIOGRAPHY

1. "Retroviruses" *Retroviruses*, edited by John M. Coffin et al., Cold Spring Harbor, Cold Spring Harbor Laboratory Press, 1997. *National Center for Biotechnology Information*, www.ncbi.nlm.nih.gov/books/NBK19376/#_NBK19376_. Accessed 26 Feb. 2016.
2. Rein, A. *Retroviral RNA Packaging: A Review*. Vol. 9, Springer Vienna, 1983. Archives of Virology Supplementum.
3. Nicole, S., & Saïb, A. (2004). Early steps of retrovirus replicative cycle. *Retrovirology*, 1, 9. <http://doi.org/10.1186/1742-4690-1-9>
4. Zhang, J., Tang, L. Y., Li, T., Ma, Y., & Sapp, C. M. (2000). Most retroviral recombinations occur during minus-strand DNA synthesis. *Journal of Virology*, 74(5), 2313–22. <http://doi.org/10.1128/JVI.74.5.2313-2322.2000>
5. Cohen, C. J., Lock, W. M., & Mager, D. L. (2009). Endogenous retroviral LTRs as promoters for human genes: A critical assessment. *Gene*, 448(2), 105–114. <http://doi.org/10.1016/j.gene.2009.06.02>
6. Uren, A. G., Kool, J., Berns, A., & van Lohuizen, M. (2005). Retroviral insertional mutagenesis: past, present and future. *Oncogene*, 24(52), 7656–7672. <http://doi.org/10.1038/sj.onc.1209043>
7. Temin, H. M. (1982). Function of the retrovirus long terminal repeat. *Cell*, 28(1), 3–5. [http://doi.org/10.1016/0092-8674\(82\)90367-1](http://doi.org/10.1016/0092-8674(82)90367-1)
8. Skinner, L. M., Sudol, M., Harper, A. L., & Katzman, M. (2001). Nucleophile selection for the endonuclease activities of human, ovine, and avian retroviral integrases. *Journal of Biological Chemistry*, 276(1), 114–124. <http://doi.org/10.1074/jbc.M007032200>

9. Bushman, F., Lewinski, M., Ciuffi, A., Barr, S., Leipzig, J., Hannenhalli, S., & Hoffmann, C. (2005).
DNA integration. *Nature Reviews. Microbiology*, 3(11), 848–58.
<http://doi.org/10.1038/nrmicro1263>
10. Jern, P., & Coffin, J. M. (2008). Effects of retroviruses on host genome function. *Annu Rev Genet*, 42,
709–732. <http://doi.org/10.1146/annurev.genet.42.110807.091501>
11. Dewannieux M, Heidmann T. Endogenous retroviruses: acquisition, amplification and taming of
genome invaders. *Curr. Opin. Virol.* 2013 doi: 10.1016/j.coviro.2013.08.005.
12. Stoye, J. P. (2012). Studies of endogenous retroviruses reveal a continuing evolutionary saga. *Nature
Publishing Group*, 10(6), 395–406. <http://doi.org/10.1038/nrmicro2783>
13. Kassiotis, G. (2014). Endogenous retroviruses and the development of cancer. *Journal of Immunology
(Baltimore, Md. : 1950)*, 192(4), 1343–9. <http://doi.org/10.4049/jimmunol.1302972>
14. West, A. G., & Fraser, P. (2005). Remote control of gene transcription. *Human Molecular Genetics*,
14 (SPEC. ISS. 1), 101–111. <http://doi.org/10.1093/hmg/ddi104>
15. Morishita K, Parganas E, Matsugi T and Ihle JN. (1992). *Mol. Cell. Biol.*, 12, 183–189.
16. Kreider BL, Orkin SH and Ihle JN. (1993). *Proc. Natl. Acad. Sci. USA*, 90, 6454–6458.
17. Palmer S, Brouillet JP, Kilbey A, Fulton R, Walker M, Crossley M and Bartholomew C. (2001). *J.
Biol. Chem.*, 276, 25834–25840.
18. Frendo, J.-L., Olivier, D., Cheynet, V., Blond, J.-L., Bouton, O., Vidaud, M., ... Mallet, F. (2003).
Direct involvement of HERV-W Env glycoprotein in human trophoblast cell fusion and
differentiation. *Molecular and Cellular Biology*, 23(10), 3566–74.
<http://doi.org/10.1128/MCB.23.10.3566>
19. Johnson, W. E., & Coffin, J. M. (1999). Constructing primate phylogenies from ancient

- retrovirus sequences. *Proceedings of the National Academy of Sciences of the United States of America*, 96(18), 10254–60. <http://doi.org/10.1073/pnas.96.18.10254>
20. Kamath, P., Elleder, D., Bao, L., Cross, P., Powell, J., & Poss, M. (2013). The Population History of Endogenous Retroviruses in Mule Deer (*Odocoileus hemionus*). *Journal of ...*, 105(2), 1–15. <http://doi.org/10.5061/dryad.5c7c6>
 21. Villa, T., & Guthrie, C. (2005). The Isy1p component of the NineTeen Complex interacts with the ATPase Prp16p to regulate the fidelity of pre-mRNA splicing. *Genes and Development*, 19(16), 1894–1904. <http://doi.org/10.1101/gad.1336305>
 22. Konarska, M. M., & Query, C. C. (2009). Insights into the mechanisms of splicing: more lessons from the riboso ... Insights into the mechanisms of splicing : more lessons from the riboso ..., 1–3. <http://doi.org/10.1101/gad.1363105.ceosome>
 23. Aken, B. L., Ayling, S., Barrell, D., Clarke, L., Curwen, V., Fairley, S., ... Searle, S. M. J. (2016). The Ensembl Gene Annotation System. *Database : The Journal of Biological Databases and Curation*, 2016, baw093. <http://doi.org/10.1093/database/baw093>
 24. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The human genome browser at UCSC. *Genome Res.* 2002 Jun;12(6):996-1006.
 25. Relethford, J. H. (2012). Hardy–weinberg equilibrium. *Human Population Genetics*, (Stern 1965), 23–48. [http://doi.org/10.1662/0002-7685\(2000\)062\[0020:THWESH\]2.0.CO;2](http://doi.org/10.1662/0002-7685(2000)062[0020:THWESH]2.0.CO;2)
 26. Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews. Genetics*, 10(1), 57–63. <http://doi.org/10.1038/nrg2484>

- 27 Dunn, C. A., Romanish, M. T., Gutierrez, L. E., Van De Lagemaat, L. N., & Mager, D. L. (2006). Transcription of two human genes from a bidirectional endogenous retrovirus promoter. *Gene*, 366(2), 335–342. <http://doi.org/10.1016/j.gene.2005.09.003>
- 28 Feuchter, A., Mager, D., 1990. Functional heterogeneity of a large family of human LTR-like promoters and enhancers. *Nucleic Acids Res.* 18, 1261–1270
- 29 Domansky, A.N., Kopantzev, E.P., Snezhkov, E.V., Lebedev, Y.B., Leib-Mosch, C., Sverdlov, E.D., 2000. Solitary HERV-K LTRs possess bi-directional promoter activity and contain a negative regulatory element in the U5 region. *FEBS Lett.* 472, 191–195
- 30 Schulte, A.M., et al., 2000. Influence of the human endogenous retrovirus-like element HERV-E.PTN on the expression of growth factor pleiotrophin: a critical role of a retroviral Sp1-binding site. *Oncogene* 19, 3988–3998
- 31 Lee, W.J., Kwun, H.J., Jang, K.L., 2003. Analysis of transcriptional regulatory sequences in the human endogenous retrovirus W long terminal repeat. *J. Gen. Virol.* 84, 2229–2235.

ACADEMIC VITA OF KALEB BOGALE

ktb5179@psu.edu

EDUCATION

The Schreyer Honors College at The Pennsylvania State University
B.S. Biology neuroscience focus, *Honors in Biology

University Park, PA
May 2017

RESEARCH EXPERIENCE

Research Assistant under Dr. Mary Poss, The Penn State Department of Biology January 2015 - Present

- Completing undergraduate thesis determining the regulatory changes in gene expression when endogenous retroviruses integrate within host gene promoters.
- **Manuscript in preparation**, "Impact of Transcriptionally Active Endogenous Retroviruses on Mule Deer Gene Expression" Pending- June 2016

Research Assistant under Dr. Mary Poss & Dr. Steven Schiff, The Penn State Department of Biology

- Utilized well-established bacterial identification methods and computational tools to identify what bacterial agents were dominant in neonatal sepsis cases in Uganda. January 2014- December 2015
- Identified the correlation between Uganda rainfall and rates of neonatal sepsis from Ugandan hospital patient records.

Research Assistant under Dr. Douglas Coulter, Children's Hospital of Philadelphia June - August 2014

- Characterized the activity of semilunar granule cells in a temporal lobe epilepsy mouse model with multicellular calcium imaging of the hippocampus brain region.

Summer Research Internship Programs

- Summer Research Opportunity Program at The Pennsylvania State University (2016)
- Hershey Medical Center MD/PhD Summer Exposure Program (2015)
- Summer Undergraduate Internship Program at The University of Pennsylvania (2014)

HONORS/AWARDS

The Millennium Scholars Program, The Eberly College of Science 2013-Present

- A merit-based scholarship program designed to prepare students for the pursuit of doctoral degrees in science, technology, engineering and mathematics (STEM) disciplines.

The Eberly College of Science Undergraduate Thesis Research Grant (\$2,000) October 2017

PUBLICATIONS/PRESENTATIONS

Published Co-author, Separating Putative Pathogens from Background Contamination with Principal Orthogonal Decomposition: Evidence for *Leptospira* in the Ugandan Neonatal Septisome (2016). *Frontiers in Medicine*, 3(22)

Penn Honors Diversity (PHD) Symposium, Poster Presentation Award Winner September 2016

Annual Biomedical Research Conference for Minority Students, Travel Grant Awardee November 2015

Undergraduate Poster Presentation Award Winner, Eberly College of Science October 2014

Keynote Speaker, Inaugural HHMI Millennium Scholar & Meyerhoff Retreat October 2014

LEADERSHIP

Treasurer of the Undergraduate Research Society, <i>The Penn State University</i>	2014 – Present
Millennium Scholars Program Cohort Council Member, <i>The Penn State University</i>	May 2014 – May 2016

COMMUNITY SERVICE

State College Area High School VIPS Tutoring Program	August 2016- Present
Chemistry Tutor, Millennium Scholars Program	August 2015- Present

PHYSICIAN SHADOWING

Clinic Intern for University Health Services, <i>Pennsylvania State University</i>	August 2015 – May 2016
Clinic Observation Program, <i>Hershey Medical Center Outpatient Center</i>	April 2016 – May 2016
Hershey Medical Center Neurology Department Shadowing	June– August 2016