THE EXPRESSION AND DISTRIBUTION OF INSERTIONALLY POLYMORPHIC ENDOGENOUS RETROVIRUSES IN CANINE CANCER DERIVED CELL LINES.

Abigail S. Jarosz

A Thesis

Submitted to the Graduate College of Bowling Green
State University in partial fulfillment of
the requirements for the degree of

MASTER OF SCIENCE

August 2018

Committee:

Julia Halo-Wildschutte, Advisor

Raymond Larsen

Zhaohui Xu

ABSTRACT

Julia Halo-Wildschutte, Advisor

To our knowledge there are no current infectious retroviruses found in canines or wild canids. It has been previously thought that the canine reference genome consists only of about 0.15% of sequence of obvious retroviral origin, present as endogenous retroviruses (ERVs) within contemporary canids. In recent analyses of the canine reference genome, a few copies of ERVs were identified with features characteristic of recent integration, for example the presence of some ORFs and near-identical LTRs. Members of this group are referred to as CfERV-Fc1(a) and have been identified to have sequence similarity to the mammalian ERV-Fc/W groups. We have recently discovered and characterized a number of non-reference Fc1 copies in dogs and wild canids, and identified unexpectedly high levels of polymorphism among members of this ERV group. Some of the proviruses we have identified even possess complete or nearly intact open reading frames, identical LTRs, and derived phylogenetic clustering among other CfERV-Fc1(a) members. Based on LTR sequence divergence under an applied dog neutral mutation rate, it is thought these infections occurred within as recently as the last ~0.48 million years. There have been previous, but unsubstantiated, reports of reverse transcriptase activity as well as gamma-type C particles in tumor tissues of canines diagnosed with lymphoma. We hypothesize that expression of members of the CfERV-Fc1(a) lineage is responsible for those observations in canine cancers. We investigated the expression of individual proviruses in canine cancer cell lines, specifically the *pol* and *env* gene. There was expression of both genes in three canine-cancer derived cells lines that cluster with CfERV-Fc1(a) members. Clustering of these sequences also suggest that there is either a new sub lineage of CfERV-Fc1(a) or possibly missed polymorphic proviral insertions that are currently assembled as solo LTRs.

ACKNOWLEDGMENTS

I would first like to express my deepest appreciation for my advisor, Dr. Julia Halo, who is an exemplary role model. Her excitement for learning and research is contagious and without her constant support and patience, this project would not be possible. Along with Dr. Julia Halo, my committee members Dr. Raymond Larsen and Dr. Zhaohui Xu were instrumental to my success here at Bowling Green State University. Both always found time in their busy schedule to sit down and collaborate with me, for which I am extremely grateful.

I was fortunate enough to work with two talented undergraduate students, Erica Cech and Malika Day, on this project. It is rare to find people who are as intelligent and hardworking, while having the most beautiful souls, like Erica and Malika: thank you for everything. I would also like to express my gratitude to the rest of the Halo Lab for providing a fun and positive work environment which I looked forward to going to each day. A special thanks to Dwayne Michael Carter, II for the constant company in the lab, and providing the inspiration to push through whatever the days would bring.

I would finally, and most importantly, like to thank my parents, siblings, and my dog Sammy. To my mom who has supported me in everything I have ever decided to do; always willing to listen to me and help keep my spirits high. Also to my dad who was equally supportive and always challenging me to see things from different perspectives. Thank you both for the endless advice and motivation. And to my siblings Amy, Megan, and Matthew, thank you for being role models in life that showed me if you work hard enough, you can do anything and reminding me not to take myself too seriously. And finally, thank you Sammy, for always being excited to see me, constantly putting a smile on my face, and being my best friend through it all.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

Transposable Elements

Transposable elements are mobile genetic elements that have, or have had, the ability to be mobile in genomes, and presently account for ~42% of the human genome [1]. These elements are broadly classified based on their ability to be mobilized through either a DNA or RNA intermediate and as such are referred to as either DNA (~3% of the human genome) or RNA transposons (~39%), respectively, the latter referred to as "retrotransposons" to distinguish their mobilization via an RNA intermediate [2]. DNA transposons found in eukaryotic systems are considered Class 2 transposable elements which all contain a central transposase-coding region that is flanked by terminal inverted repeats (IRs). Class 2 DNA transposons are further divided into three subclasses: classic "cut and paste" transposons; those which use a rolling cycle replication mechanism; and those whose transposition mechanism is not yet fully understood [2, 3]. The mechanism of cut and paste is initiated by the transposase protein binding to the IRs and facilitates transesterification and excision of the transposon. The excised transposon can then be inserted into a new location of the genome; through the exact mechanism differs across various DNA transposon elements [3].

In contrast, retrotransposons are mobilized in the genome by a "copy and paste" method. All retrotransposons are mobilized via a transcribed RNA with the help of reverse transcriptase, which is an enzyme that is able to synthesize the conversion of single stranded RNA (ssRNA) to a double stranded DNA (dsDNA) "copy" of the transcribed element. The dsDNA copy is then inserted into a new position within its host genome. A result of the complete mechanism is the flanking of new all new retrotransposon insertions by target site duplications (TSDs) that are unique to the point of insertion of the new element in its respective host genome. Retrotransposons are further classified based on structural properties into non-long terminal repeats (LTR) and LTR elements (Figure 1). Non-LTR elements consist of long interspersed nuclear elements (LINEs) and short interspersed elements (SINEs). LINEs are “autonomous” elements, meaning they have the ability to self-mobilize throughout the genome by the mechanism of retrotransposition. Their retrotransposition mechanism, specifically referred to as target primed reverse transcription (TPRT), is catalyzed by enzymes encoded in the LINE. Within a retrotransposition-competent LINE, two LINE-encoded proteins referred to as ORF1p and ORF2p are responsible for this activity [4]. ORF1p has been shown to be involved in the mechanism of LINE retrotransposition, but its exact role is not fully defined [5]. ORF2p is necessary for retrotransposition, and contains reverse transcriptase and endonuclease activities for complete mobilization of the element [4]. SINEs, however, are nonautonomous elements that harbor only a promoter and short transcribed sequence, and have consequently evolved to rely on enzymatic activities of LINE element(s) for mobility throughout the genome. Generally, SINEs are classified by their origin from either tRNA sequences or from other RNAs such as the signal recognition particle component 7SL, the latter of which resulting in the still-active Alu elements in the human genome [6].

In contrast to non-LTR transposons, LTR retrotransposons possess flanking long terminal repeats that encode transcriptional regulatory motifs for the element within the host genome. These are respectively referred to as the 5’ and 3’ LTRs. Generally, LTR transposons are derived from retroviral infections of germline tissue(s), and thus replication of LTR retroelements predominantly occurs through infection of a retroviral particle, though certain classes of LTR elements are known to retrotranspose via intracellular mechanisms [4]. LTR retrotransposons derived from retroviral infections is the main focus of this thesis.

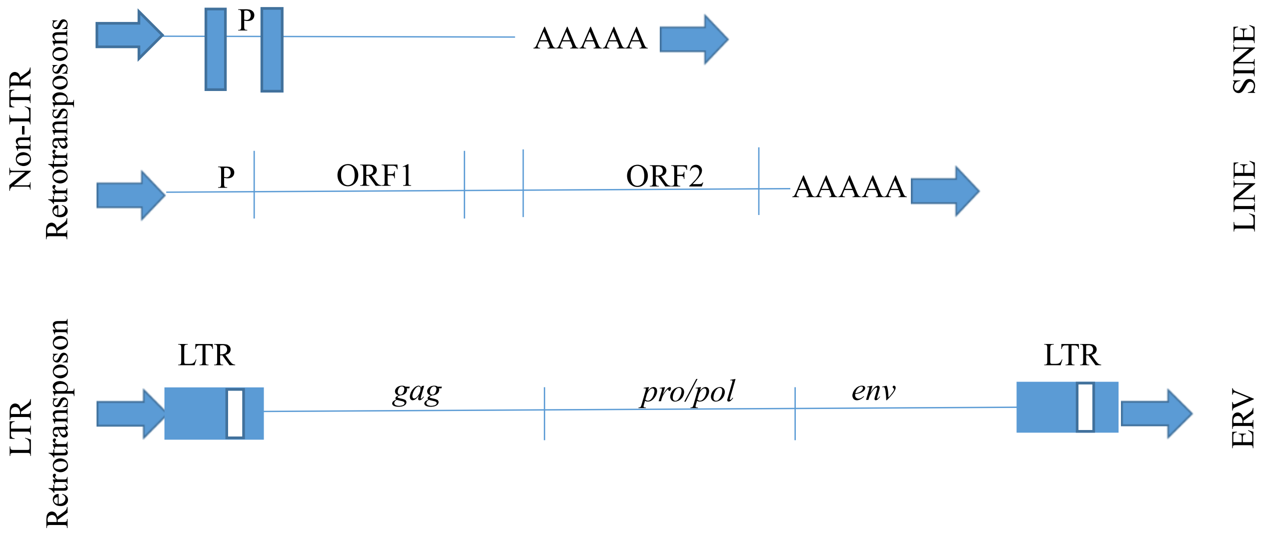


Figure 1. Depiction of Class I transposable elements.

Class I elements are divided into two overarching groups: non-LTR retrotransposons and LTR retrotransposons. Non-LTR retrotransposons are present as either an autonomous LINE or non-autonomous SINE. LTR retrotransposons consist of two long terminal repeats flanking the element, and are found in the form of endogenous retroviruses. Tandem site duplications are represented by the flanking arrows at the ends of each element.

Retrovirus Structure

Retroviruses are a unique type of enveloped virus that package their heritable information in the form of two positive RNA strands of genomic nucleic acid [7, 8]. The genome size of a retrovirus typically ranges from 7,000-11,000 base pairs and includes a 5'cap and poly-A tail. To be fully infectious the retrovirus's genomic information must i) undergo reverse transcription into a dsDNA copy, and the dsDNA copy must be ii) permanently integrated into the host genome. The copy in the host genome is referred to as a 'provirus', and differs from the corresponding RNA genome by the presence of long terminal repeats (LTRs) and short 4-6 bp target site duplications (TSDs) flanking the insertion (Figure 2). Because of the mechanism of reverse transcription, at the time of integration the LTRs are identical and contain *cis* regulatory elements, for example transcription factor binding sites (in the U3 region) and polyadenylation signal (in the R region) that are recognized by host cell transcriptional machinery. Retroviruses are divided over seven defined genera (Table 1), and are classified as either 'simple' or 'complex', the former possessing the canonical viral genes only and the latter accessory genes of various functions. More specifically, a simple retrovirus [7, 9, 10] encodes the viral genes *gag*, *pro/pol*, and *env* flanked by the 5' and 3' LTRs (Figure 3). The additional genes present in a complex retrovirus typically overlap with the canonical viral genes and have role(s) in viral replication and infection efficacy, as well as manipulation and/or evasion of the host.

As a family, the *Retroviridae* include viruses of the genera *alpha*-, *beta*-, *gamma*-, *delta*-, *epsilon*-, *spuma*-, and *lentiviruses* [9] (See Table 1). Some of the major research topics in the Halo lab, including my own research projects (see Results and Discussion), are focused on a unique gamma-like ERV lineage that we have identified as a recently expanded group within the canine. Gammaretroviruses are simple retroviruses considered to be a Type-C mammalian retrovirus [9, 11]. Various pathogenic gamma-retroviruses have been found in a variety of mammals which cause malignancies, immunosuppression, and neurological disease as well [9, 11]. Virions in the genera are distinguished by their condensed central core, with spikes that are barely visible [11].

Diagrams of the Gammaretroviral genome structure and a representative particle are provided in Figure 2 and 3, respectively. In all retroviruses, the first gene following the 5' LTR is the *gag* gene, named as an acronym of Group AntiGene, that encodes a polyprotein responsible for building the inner structure of a virus particle and includes the matrix, capsid and nucleocapsid [4]. In a mature retroviral particle, the matrix lines the inside of the surrounding envelope as well as the inner capsid. Following the matrix is the capsid which functions to protect the 'core' of the virus, within which the two single strands of viral genomic RNA are located. The two strands of RNA are further protected by subunits of nucleocapsid that functions to coat the genome copies within the virion. The *pol,* or polymerase gene, is downstream of *gag* and encodes a polyprotein including the enzymes responsible for reverse transcription of the RNA genome and integration of the resulting viral dsDNA into the host nuclear genome: reverse transcriptase and integrase. Another function encoded within the *pol* gene is protease, which is responsible for the cleavage of the viral polyproteins following budding from its host cell during maturation of the viral particle into an infectious virion. Reverse transcriptase catalyzes the reverse transcription of the ssRNA genome into dsDNA, a hallmark of all retroviruses. Newly reverse transcribed DNA is permanently integrated into the host's genome by integrase, after which the provirus may then be transcribed by the host cell machinery. The final viral gene is *env*, or envelope, which encodes the transmembrane (TM) and surface unit (SU) glycoproteins. The mature *env* glycoproteins form inner and outer envelope proteins on the virion and aid in the binding to the host cell receptor and fusion of the lipid bilayers of the virion and host cell at infection.

Table 1. Characteristics and representatives of the *Retroviridae*.

	Basic Structure	**Group**	**Class**	**Example**
Alpha	Simple	Type-C	Class II	ALV
Beta	Simple/Complex	Type-B,D	Class II	MMTV
Gamma	Simple	Type-C	Class I	MLV
Delta	Complex	Type-C like	Class II	BLV
Epsilon	Complex	Type-C	Class I	Walley dermal sarcoma virus
Lenti	Complex	N/A	Class II	HIV
Spuma	Complex	Type-C like	Class III	Human Foamy virus

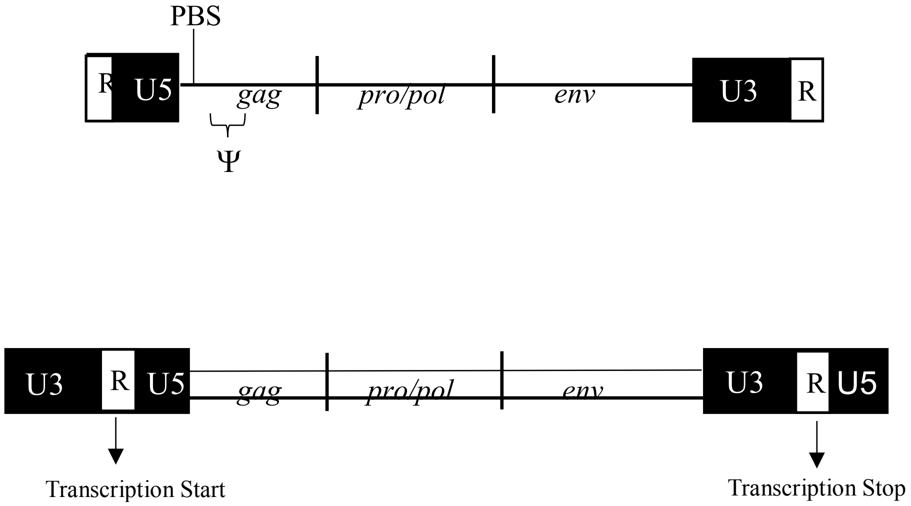


Figure 2. Unintegrated RNA genome and its integrated provirus.

A. The unintegrated (+)ssRNA genome after uncoating of the viral core. The primer binding site (PBS) is located just upstream of the *gag* gene, followed by the encapsidation sequence (Ψ) that aids in packaging of the RNA genome strands into newly formed virions during budding from the host cell. Flanking the proviral genes are sequences required for LTR formation during reverse transcription: the R sequence with the U5 region at the 5' end and U3 region with R sequence at the 3' end (also refer to Figure 7).

B. The corresponding provirus once integrated into the host genome as dsDNA. After reverse transcription, two identical LTRs are formed at both the 5' and 3' end of the provirus.

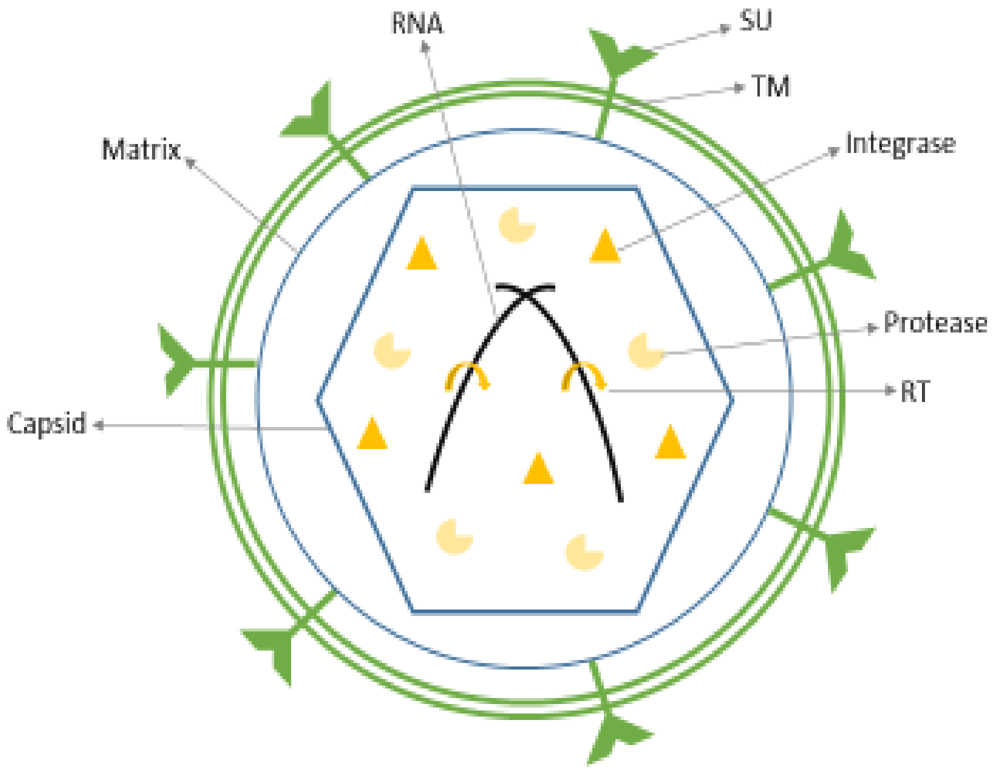


Figure 3. Schematic representation of a retrovirus particle.

Details are found in the text. The arrows indicate the structural and enzymatic components. The dimerized ssRNA genomes are represented as two linked, curved lines, and protected by the nucleocapsid (not shown). Abbreviations are as followed: SU, surface unit; TU, transmembrane domain; RT, reverse transcriptase.

Entry into Host Cell

The replication cycle of a typical retrovirus is depicted in Figure 4. A mature retroviral particle infects a new host cell by highly specific interactions of the *env* SU domain with its receptor on the host cell. The SU portion of Env is anchored to the envelope surrounding the viral core by the TM portion of Env. By virtue of the SU proteins displayed on the virion, the virus is able to bind to and infect a naive host cell [7, 12]. However, retroviral particles of some classes that lack a SU domain have been shown to bind to a cell by binding to the cells through interactions with Heparan Sulfates, though the particles are not able to actually infect the cell [13]. Binding of SU to its receptor triggers a cascade of conformational changes that facilitates fusion of the viral and host cell membranes [9, 10] and allows the capsid to enter the host cell cytoplasm. Once in the host cell, the capsid is uncoated, freeing the RNA genome for its reverse transcription and subsequent integration into the host genome.

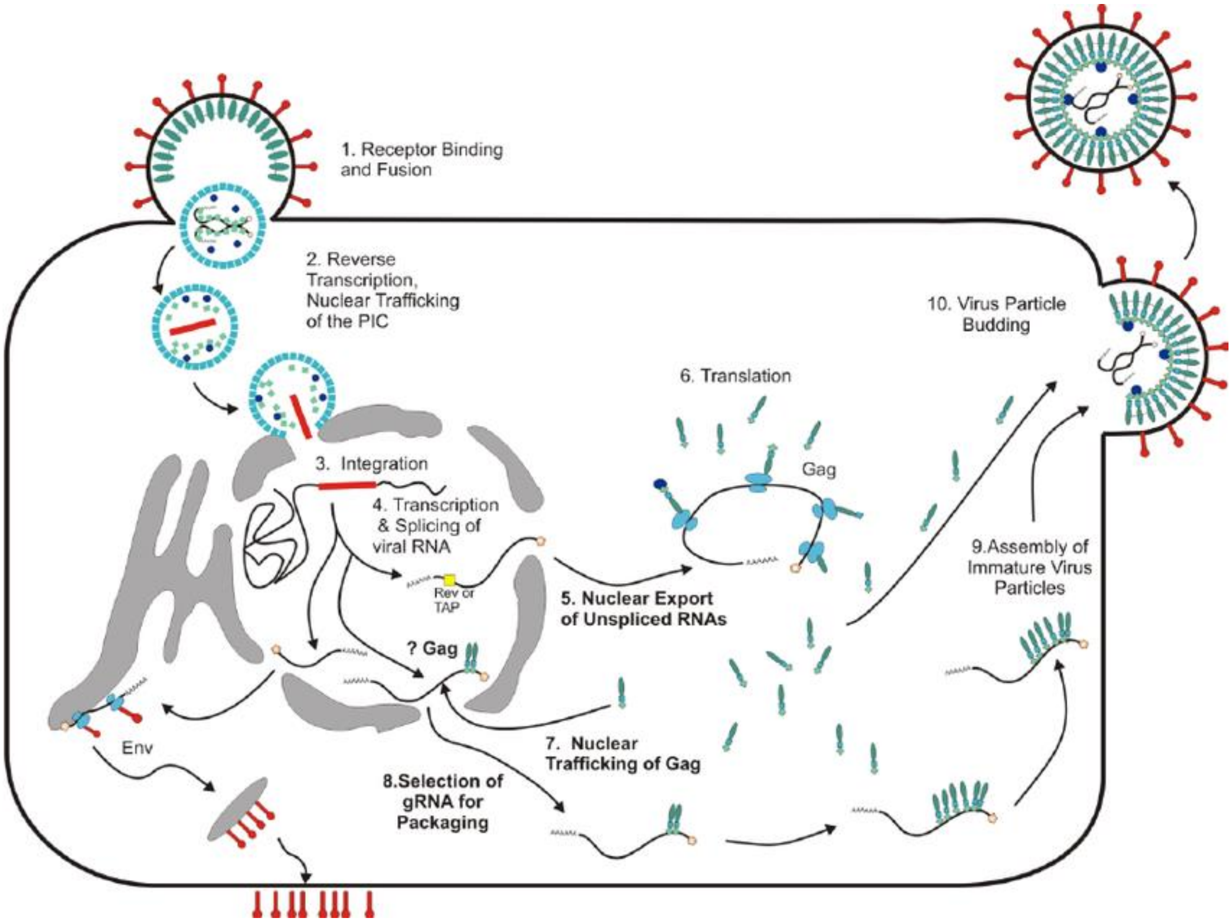

1. Receptor Binding and Fusion
2. Reverse Transcription, Nuclear Trafficking of the PIC
3. Integration
4. Transcription & Splicing of viral RNA
Rev or TAP
? Gag
5. Nuclear Export of Unspliced RNAs
6. Translation
Gag
7. Nuclear Trafficking of Gag
8.Selection of gRNA for Packaging
9.Assembly of Immature Virus Particles
10. Virus Particle Budding
Env

Figure 4. Schematic representation of retroviral replication cycle.

Steps of retroviral replication are detailed further in the text. From "Nuclear Trafficking of Retroviral RNAs and Gag Proteins during Late Steps of Replication" Stake et al, 2013[14].

Reverse Transcription

A detailed version of reverse transcription is provided in Figure 5 and described as follows. A primer binding site (PBS), just downstream of the 5'LTR, is hybridized by a tRNA molecule that is specific to the infecting retrovirus, and functions for the reverse transcriptase (RT) as a primer to initiate reverse transcription (Figure 5). DNA synthesis continues until RT reaches the 5' end of the genomic RNA where the first "strong stop" occurs. An RNaseH function (within RT itself) mediates degradation of the genomic RNA, after which RT "switches strands" causing the negative stranded DNA to be annealed to the 3'end of the viral sequence. This is accomplished by the R region that was already synthesized at the 5' end is now complementary to the R region on the 3' end (also refer to Figure 5). Negative strand synthesis resumes on the RNA template strand, accompanied by RNaseH digestion of the RNA in subsequently formed RNA:DNA hybrid. The polypurine tract (PPT) portion of the viral RNA is resistant to RNaseH digestion, allowing for this section on the positive strand to act as second primer. A second "strong stop" occurs when the PBS is reverse transcribed. After the RNaseH removes the RNA encoding the PBS, the positive strand of DNA is exposed. Annealing of the negative and positive strands at the complementary PBS segments occurs, constituting a second strand transfer of RT. DNA synthesis can then be completed, with the positive and negative strands acting as templates for one another.

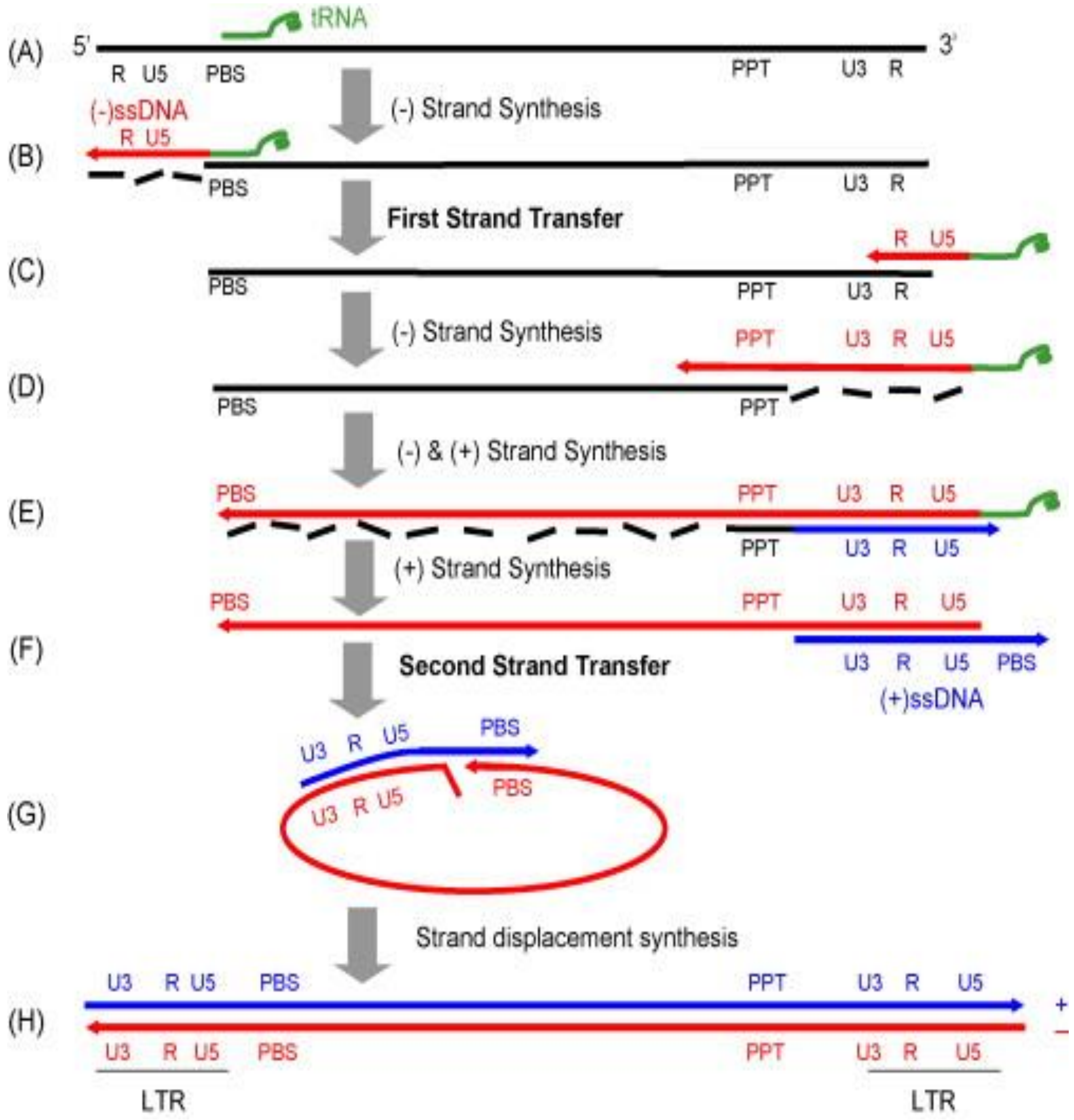


Figure 5. Process of Reverse Transcription.

Details are provided in text. Abbreviations are as followed: PBS, Primer binding site; PPT, polypurine tract. From “Strand transfer events during HIV-1 reverse transcription”, Basu et al., 2008. [15].

Integration

Along with the synthesis of linear, viral dsDNA, the two identical LTRs each made up of the U3, R, and U5 regions are generated. The LTR structures are coupled with a specific nucleoprotein complex, making up what is referred to as the pre-integration complex. While in the cytoplasm, integrase cleaves the viral DNA at either of the two 3' termini, which in all retroviruses produces the resulting terminal sequences 5'-TG…CA-3'[4]. Cleavage of this site provides a 3'-OH group to act as an attachment site for the provirus to the host DNA during integration. The viral nucleoprotein complex is transferred to the nucleus, which typically occurs during mitosis when the nuclear membrane is disassembled. However, some retroviruses have the ability to ' 'piggy-back' existing host machinery, relying on active transport to enter [11]. Once the pre-integration complex has accessed the nuclear DNA, binding of the host DNA by integrase viral DNA complex occurs, mediated by attack of the free 3'-OH groups on the viral DNA to the phosphodiester bonds on the target DNA. The energy of the newly broken phosphodiester bonds on the host DNA is transferred in order to form the new bonds between the viral and host DNA [11]. Extending from the 3'-OH group, DNA synthesis by host cell machinery fills in the gaps flanking the viral DNA. The viral DNA is permanently integrated into the host's genomic DNA. The overall mechanism is summarized by Figure 6.

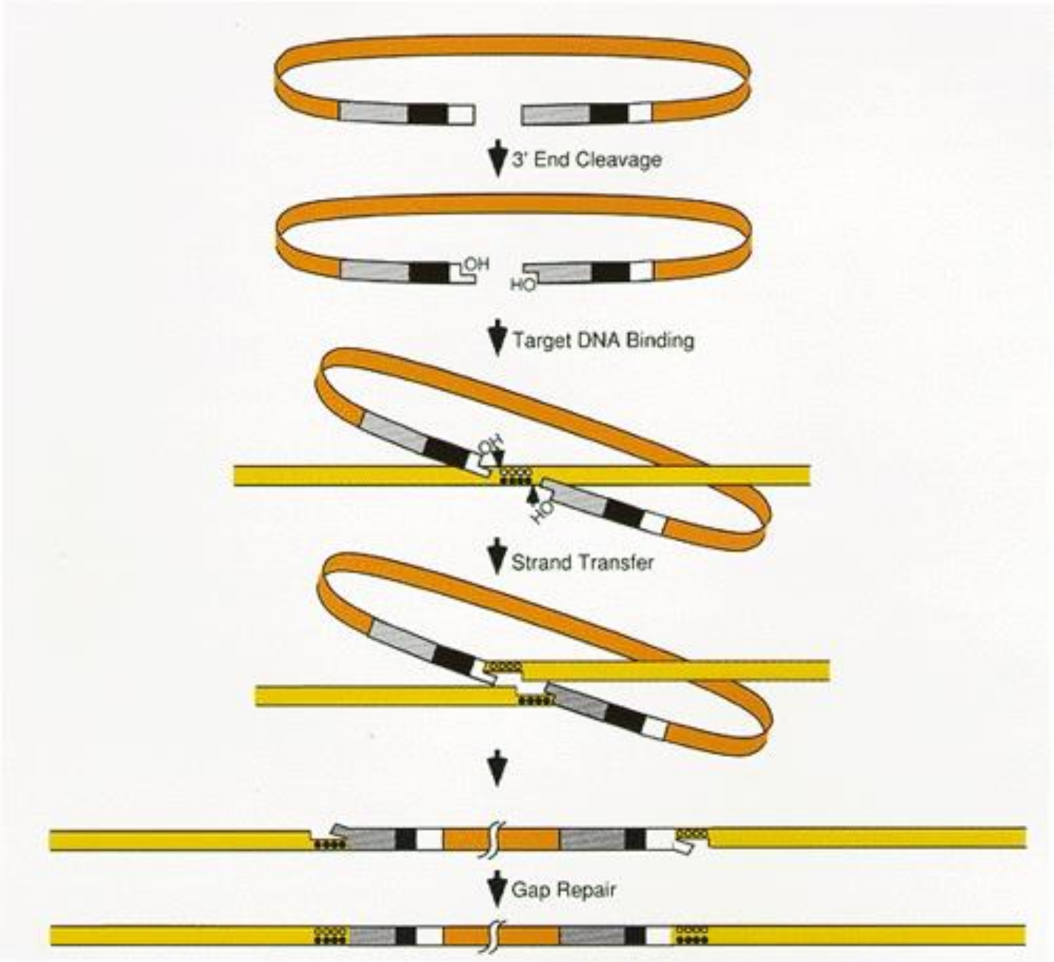


Figure 6. Steps of proviral integration into host genome.

Details are provided in text. Image from *Retroviruses*, Coffin 1997 [11].

Synthesis and Assembly

Retroviral transcription is regulated by RNA polymerase II, which is responsible for synthesizing cellular mRNAs in the host cell. Once the provirus is integrated into the host cell, RNA polymerase II is recruited to the TATA box (within the LTR U3 region), the provirus's core promotor element. Transcription of the provirus begins at the beginning of the R region of the 5' LTR, and proceeds to the end of the R region of the 3'LTR. Just as the majority of other host mRNA transcripts, a 5' methyl cap is added along with a 3' poly-A-tail. In order for retroviral gene expression to elicit a productive viral infection, there must be both spliced and unspliced transcripts transferred to the cytoplasm [11] (Figure 7). Typically, simple retroviruses have a splice donor site in the proviral leader sequence and a splice accepter site just before the *env* coding region creating *env* specific mRNA (Figure 7). A ratio of *gag-pol* mRNA and *env* mRNA must be retained in order for efficient replication of the retrovirus. This ratio of transcripts is usually determined by *cis*-activating sequences found in the retrovirus.

The polyprotein for *gag* is translated from an unspliced transcript on a free ribosome within the host cytoplasm, and the translated product directs the budding of newly formed retroviral particles from the cell. Like *gag*, the *pro-pol* genes are likewise synthesized from unspliced transcript, but are not independent from *gag*, initially creating *gag-pro* or *gag-pro-pol* polyprotein. Further translation of *pro* and *pol* is accomplished via the combination of a read-through and frameshift of the viral mRNA. The *gag* segment of the polyprotein is responsible for directing the Pro and Pol segments to the site to viral assembly and further mediates the segments into newly forming particles. Gag is also responsible for binding to genomic RNA through the nucleocapsid domain and its recruitment to the new viral particle [16]. This complex then associates with the plasma membrane through the matrix domain. Association of the matrix and plasma membrane initiates assembly of the viral core on the cytosolic side.

The *env* mRNA coding for the SU and TM being the product of a spliced transcript are translocated through the membrane of the rough endoplasmic reticulum. The *env* mRNA is anchored into the membrane by a hydrophobic region located near the carboxyl terminus. Through vesicular transport, it is carried through the Golgi apparatus, where a cellular protease cleaves the polyprotein into the SU and the TM products. This cleavage is necessary in order to activate the hydrophobic fusion peptide found on the amino terminus of the TM. Following cleavage, it is then transported to the plasma membrane and exposed on the outside of the cell. Env proteins reach the site of budding by lateral movement allowing assembly of particles containing *gag*, *gag-pro-pol*, and viral ssRNA. Newly assembled particles then bud off the host cell's surface, meaning the lipid membrane layer surrounding the capsid is acquired during budding from the previous host cell [17].

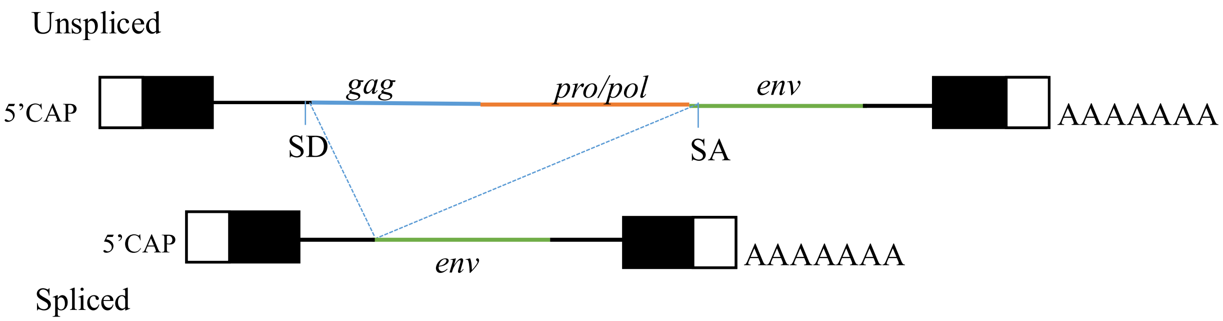


Figure 7. Transcripts produced from proviral sequence.

A. Messenger RNA for *gag* and *gag-pol*, showing where the splice sites are coded to create the *env* transcript

B. The spliced messenger RNA for *env* transcript, excising out the *gag-pol* region.

Maturation

Maturation of the viral particle either begins while budding occurs or immediately thereafter. During maturation, the viral protease, present within the virus, cleaves the Gag polyprotein into its separate domains. The matrix domain remains anchored to the plasma membrane. The nucleocapsid remains complexed with the genomic RNA but condenses into a compact orbicular mass. The capsid, however, undergoes a major morphological transformation from an immature state of being more or less spherical to forming distinct geometries including cylindrical, polyhedral and conical [16]. Only following maturation is the virion infectious.

Host Defense

Both the virus and host are continually evolving mechanisms to counter the other. As such, host cells have evolved various mechanisms to restrict steps of the virion replication process. Since the virus is reliant on host machinery to replicate and produce more infectious virus, most described host defense restriction mechanisms focus on: receptor intervention; restriction during uncoating; and viral assembly including inhibition of virus release from the cell surface [18]. Receptor interference occurs when an exogenous retrovirus is inhibited from binding its specific receptor due to the expression of an retrovirus pre-existing in the cell, resulting in receptor sequestration, becoming known as ‘super infection’. This phenomenon is sometimes mediated by ‘endogenous’ viral forms that result from germline infection (this topic is detailed in sections below). Restriction during uncoating has been found both in primates, with the help of TRIM5 proteins, as well as in mice with Fv1 proteins. Both proteins restrict viruses during uncoating by targeting the capsid region of Gag, however the exact mechanism remains unknown [19]. For example, an endogenous form of the Jaagsietke sheep retrovirus (enJSRV) inhibits its exogenous counterpart by restricting viral assembly. Specifically, if two copies of enJSRV are found with mutated *gag* sequences that encode dominant negative proteins, these proteins then interfere with the late stages of the retroviral life cycle [20]. Inhibiting the release of the virus from the host' surface has been shown in HIV-1 in humans, where a gene product known as tetherin is able to stop enveloped viral release. Tetherin works in a relatively simplistic mechanism in which it physically tethers the viral particles to the host's cellular membrane [21]. If the host cell fails to prevent the integration of the element, then methylation of the integrant can act to silence expression [22].

Spread of Retroviruses

Retroviruses are often transmitted horizontally (*i.e.*, individual to individual) through bodily fluids including those exchanged during both mating and breast feeding. Horizontal transfer is not limited to interspecies transmission, but also subjected to cross species transmissions [23]. This is frequently observed between predators and prey exchanging blood to blood transmission through biting or open wound contact. Although retroviruses usually infect somatic cells and spread through horizontal transfer, on occasion the retrovirus can access a germ cell or germ tissues. If infection occurs in the germline, the provirus has the potential to be passed vertically, in a more or less Mendelian fashion, from the host to its progeny, and is then termed an endogenous retrovirus (ERV). Owing to the much lower neutral substitution rate while replicated as a permanent part of the host genome, ERVs are considered to represent "fossil" forms of their once-exogenous fast-evolving counterparts. With the unique feature of the ERVs having identical LTRs along with adopting the mutation rate of their host, comparative sequence analysis can be used to determine the time of infection. Once an ERV is introduced into the germline, it may be amplified in that genome beyond the initial infection. If there is expression of one of these ERVs in a normal cell, the resulting viral RNA has the potential to contribute to infectious virions. Amplification may also occur if there are multiple ERVs complementing each other in trans or as aided by the infection of the cell by a new exogenous retrovirus in a process referred to as complementation in *trans*. Retrotransposition can also occur in *cis*, if the element with *env* loss relying on *gag* and *pol* [24]. With the removal of the *env* gene, the viral sequence evades the host defense, such as tetherin discussed above, which inhibits budding of the virus for reinfection. Additional ERV amplification can also be assisted if an ERV transcript piggybacks off a LINE-encoded enzymes; this mechanism of amplification is strictly intracellular and referred to as retrotransposition in *cis*.

Evolutionary selection and Polymorphism

ERVs that have been recently integrated in the host bear a strong resemblance to their exogenous counterparts. This resemblance implies the ability to remain pathogenic, or have some function(s), at least for some time. In regards to integration site, there appears to be a preference to the nucleotide sequence directly flanking insertion, hypothesized to provided proper manipulation of the DNA to produce a secondary structure that is optimal for integrase [25, 26]. Although the direct flanking regions may show sequence preference, there appears to be little to no inclination for integration site distant from it [26]. The location of integration in the genome is more or less random, implying that the chance integration at an orthologous position and endogenized of a provirus in two species is negligible. Therefore, it can be assumed that if a provirus is found in two species, it must have been integrated into the shared ancestor before speciation occurred.

ERVs can be detrimental to their host, but also have the capability to be beneficial and even essential to normal physiology of the host species. In general, ERV insertions are considered to be subjected to drift and/or selection, occasionally resulting in the loss or eventual fixation of the insertion. Unlike their exogenous counterparts that evolve rapidly, ERVs evolve at the neutral substitution rate of the new host's genome. Given enough time after initial infection, an ERV has the potential to eventually reach fixation in a species' genome [9, 27, 28]. If the integrated retrovirus is mutated in such a way that it is 'dead on arrival', or even minimally infectious, the sequence may not be purged from the genome right away, thereby permitting its potential for transmission to the offspring of the host. For recently integrated and apparently 'non-harmful' ERVs, it is common to see insertionally polymorphic copies both between and within individuals, the former based on presence/absence over individuals within a population and the latter being present in a heterozygous state within a single genome. For ERVs not causing substantial harm, or that might have a sub-deleterious effect, it will take longer for the ERV to be selected against and removed from the population. Neutral ERVs, or those integrants without any apparent effect to the host, may increase in population frequency over time or be lost by drift.

Significantly, there are several examples of ERVs that have offered the host some benefit and so have been co-opted for host functions [4]. In cases where these ERVs are beneficial to the host, these insertions are positively selected for and can reach a state of fixation at a rapid pace. An example of this is syncytin, a viral-derived protein expressed in mammalian placentas in convergent evolution [29]. This developmental function of the placenta has evolved independently across several mammalian clades and distinct ERV lineages by virtue of the viral-derived Env receptor-binding and fusogenic properties. For example in humans, the fusogenic property displayed by a HERV-W *env* gene was co-opted and utilized by the host the function for placental fusing into the uterine wall [30]. Owing to its advantage to the host, the HERV-W *env* was subjected to strong purifying selection in evolution of the primate lineage leading to contemporary humans. Due to the highly repetitive nature of the LTRs, and the fact that they are identical at the time of integration, there is a tendency of the LTRs to recombine. As a consequence, the inner provirus sequence is removed leaving behind a single LTR, termed a ‘solo’ LTR [12, 31] (Figure 8). This is the most common deletion among ERVs and in fact, solo LTRs outnumber other ERVs [32]. Although the functional genes for the viruses has been excised out, the promoter for the element remains in the genome; such solo LTRs have also been utilized for transcriptional affects or as enhancer elements over the evolution of many mammalian species [33].

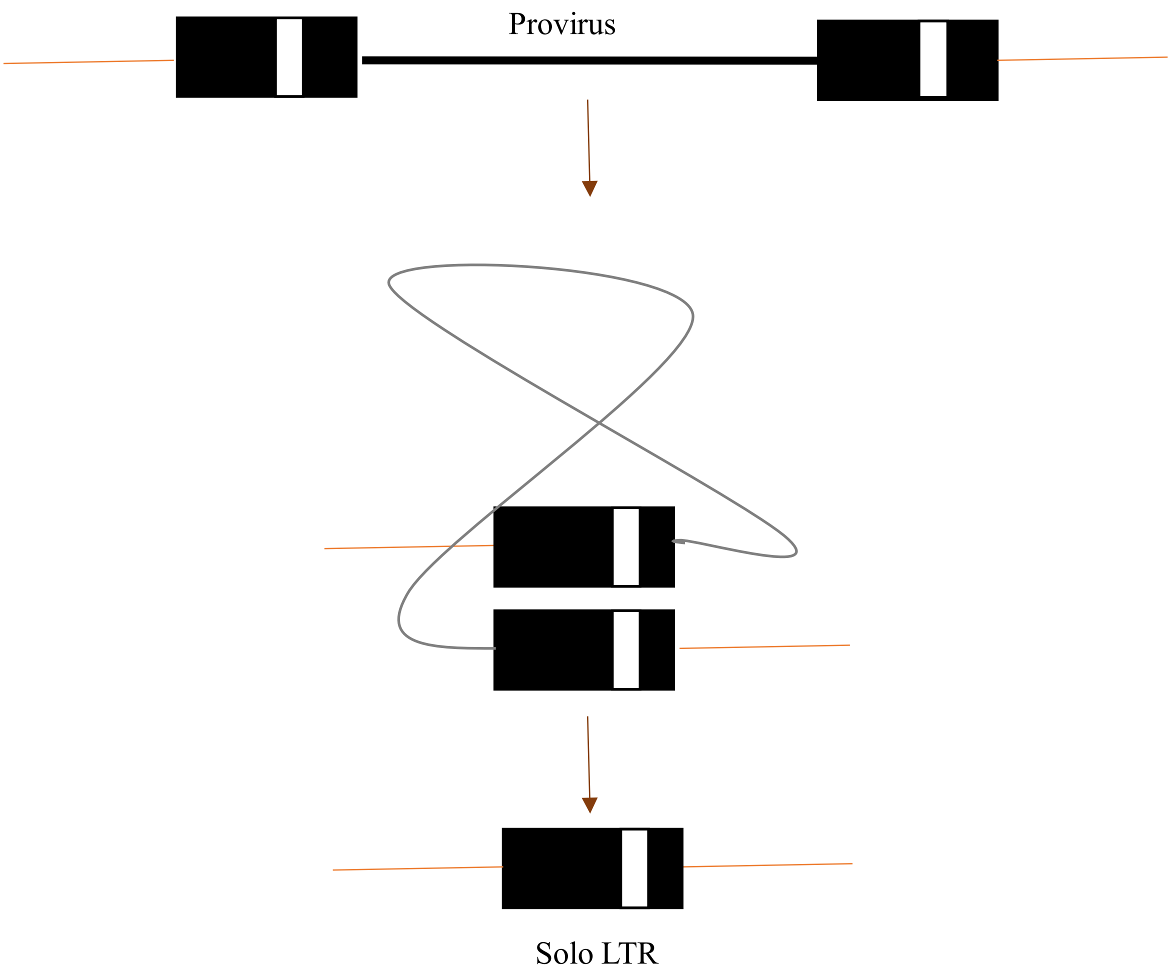


Figure 8. Formation of a Solo LTR.

Due to highly repetitive nature of the LTRs and the selection against harmful proviral insertions, corresponding 5’ and 3’ tend to recombine with one another. This leads to the excision of viral genes, leaving behind a single LTR, termed a solo LTR.

ERVs and Disease

The discovery of cancer-causing retroviruses began with the identification of Rous sarcoma virus, in which an oncogenic retrovirus was infecting chickens and inducing sarcomas [34]. Discovery of this virus causing cancer fueled the interest in retrovirology, leading to the discovery additional disease-causing viruses. Referred to as “tumor viruses” at the time, the viruses were shown to cause disease from having acquired host proto-oncogenic genes, that likely resulted from reverse transcription of co-packaged host mRNAs within a virion. Therefore, infection by the virus resulted in transformation of the cell via direct expression of the oncogene following its integration as a physical segment of a provirus. Animal models developed based on this research were pivotal in laying the foundation modern cancer research; however, with the exception of the rare aforementioned oncogenic retrovirus, exogenous retroviruses have been shown to cause cancer predominantly through gene disruption/affects via insertional mutagenesis. This characteristic is shared amongst the exogenous retroviruses with some ERVs that retain infectivity, which was first uncovered in mice during this exploration period in the 1960s and early 1970s [35]. While researching the exogenous form of murine leukemia virus (MLV), an endogenous form was revealed, which was also shown to cause the development of thymic lymphomagenesis in mice [35]. ERVs have been further linked to the development of disease in human and other mammals, including cancers, yet, their exact role(s) in this development remains unclear.

ERVs account for >8% of the human genome and almost all appear as highly mutated into remnants of their former exogenous counter parts [27]. However, some ERVs still have intact, or nearly intact, genes that code for their viral function [36]. Expression of these particular ERVs have led to observations of retrovirus like particles [37], reverse transcriptase activity, and antibody responses in a multitude of diseases [38]. Some cancers have been consistently linked to presence of ERVs in humans including; seminomas [39], testicular cancer [40], certain breast cancers [41-43], renal cancer [44] and leukemia/lymphomas [43, 45, 46]. HERV-K, a ‘young’ beta-like ERV group found in humans (so termed for its use of a tRNA^{lys} to initiate reverse transcription) that is estimated to have infected the germline as recent as within the ~150 thousand years [28, 47]. HERV-K has been shown to be expressed in human cancers as well as during HIV-infection. This expression has been shown to have tissue specificity by analysis of expressed HERV-K proviral RNAs, though the consequences of such expression remain unclear [43]. In the same study, a subset of expressed RNAs were also identified that did not appear to have a matched annotated locus, indicating the presence of transcriptionally active proviruses at uncharacterized loci [43]. Given their relatively recent integration, the presence of some intact open reading frames is not necessarily surprising and even suggests that these insertions may be offering some benefit to the host [23, 48]. Studies have suggested the immunosuppressive domain of the *env* promotes tumor growth through suppression of anti-tumor immunity [49]. In the context of the research presented here, it is important to recognize the HERV-W and HERV-Fc elements, two distinct gamma-like lineages present in the human genome, each of which possesses one or two proviruses with some intact genes and thus resemble recently integrated HERVs, despite their insertion over tens of millions of years ago [27]. Both activation and functional viral products from these human ERV lineages have been detected in tissues associated with certain diseases as well as in normal physiology [50].

Solo LTRs have also been shown to possess the potential to disrupt normal physiology [32]. Although the functional genes for the virus have been excised, the promoter for the element remains in the genome (also refer to Figure 8). Integration of proviruses can occur seemingly random in the genome [4] and therefore can be integrated in or near a proto-oncogene. Since proviruses, as well as solo LTRs, encode a promoter that is recognized by the host, the insertion may cause a proto-oncogenes to become transcriptionally active [9, 31, 38]. Mobility of these viruses also allows for the possibility of the integration within a gene via insertional mutagenesis, interrupting the protein coding regions. If this occurs in a tumor suppressing gene, the ability of the cells to halt tumor growth can be jeopardized. Since the link between ERV expression and development of disease is not fully clear, understanding the biology and evolution of closely related ERVs will help us better understand their role in human health and disease.

CfERV-Fc1(a)

As opposed to human and other mammals, the dog displays a substantially lower ERV presence, representing just 0.15% of their genome [51]. To this day, there have been no confirmed infectious exogenous retroviruses in the dog, or any candid. However, ERVs present in the canine genome (the reference genome from a boxer breed dog, referred to as CanFam3.1) confirm that retroviruses clearly infected canine ancestors. The vast majority of ERVS in the CanFam3.1 genome are of ancient origin, however there are features of one or two proviruses that suggest their relatively recent integration. Specifically, these few apparent recent integrants possess some open reading frames as well as high nucleotide identity between the 5' and 3' LTR [51]. In 2016, Diehl *et al* examined this ERV lineage, 'ERV-Fc', in a mammalian-wide analysis including the Caniformia, and classified the latter integrants as ERV-Fc1-derived (for its use of a tRNA^{phe} to initiate reverse transcription). Invasion of the Caniformia germline appeared to take place roughly 20 million years ago, from a virus that appeared as a recombinant of two gamma-like lineages [23]. Namely, the *gag*, *pol* and flanking LTRs have been derived from an Fc lineage that had, at some point, assimilated an *env* gene most closely related to ERV-W (syncytin-like). A sub lineage, CfERV-Fc1(a) invaded canid ancestor by an unknown crosspieces transmission (possibly from a ferret or now extinct source) [23]. After initial infection, multiple germline invasions followed until at least the last 1-2 million years ago. The majority of CfERVs found in the *Canis familaris* CanFam3.1 boxer genome appear to be older insertions that are severely mutated and presumed as fixed amongst canids. The current CfERV-Fc1 sequenced found in Repbase, based on reference proviruses, show open reading frames for both *gag* (~2.0kb) and *pol* (~3.5kb). A deletion in *env* deletion present while three reference insertions have full sequences and chrX:50,661,636 has an open reading frame. CfERV-Fc1 will be the focus of this thesis. This ERV differs from the typical gamma or gamma-like retrovirus by generating target site duplications of 5 base pairs instead of the usual 4 base pairs [23].

Using whole genome Illumina data from >100 dog breeds, semi-feral dogs, wolves, and wild canids, and using an anchor mapping strategy, my advisor Dr. Halo was able to infer at least 59 CfERV-Fc1 insertions that are not present in the Boxer reference, and another group of at least 11 reference elements deleted from the sequenced genomes of the same >100 samples (unpublished data; manuscript in prep). To improve support for the identification of the new insertions, the read data subsets of the samples were combined and analyzed utilizing BAM and Retroseq [52, 53]. Any of the initial candidate calls that were within 500 bp range of a reference insertion were disregarded due to the possibility of a false call. With the resulting potential insertions, a *de novo* assembly was applied to supporting read pairs obtained for each site in order to reconstruct individual LTR-genome junction for each site. A total of 59 new insertions were identified and 35 formally validated; the others were not validated due to limited DNA availability. Dr. Halo then identified and validated full-length proviruses for 8 of the 35 sites. The remainder of sites were confirmed to have a solo LTR present, and a few sites had both solo LTR and provirus alleles present, in addition to the pre-integration site. We are currently in collaborative work to screen additional sites in samples for which DNA was previously unavailable (*i.e.*, wild canids); we anticipate additional insertions, including full-length elements, will be characterized.

This data demonstrated that the level of insertionally polymorphic young CfERV-Fc1 integrants greatly exceeds that of HERV-K in humans and its highly variable presence in contemporary canines is reminiscent of disease causing gamma-like ERVs such as MLV in mice [54]. In our recent findings, young copies of the gamma-like lineage CfERV-Fc1 have been identified that have sequence similarity to the mammalian ERV-Fc/W groups in the boxer reference for canines [10]. Specifically, the *gag* and *pol* genes are most similar to ERV-Fc and the *env* gene is most similar to ERV-W, the human syncytin1 gene [23]. The CfERVFc1 group also possesses a unique feature of being a recombinant of ERV-Fc and ERV-W, both present as related forms in humans implying the CfERV-Fc1 lineage shares a common ancient source [23, 55]. Investigating this recombinant ERV will lead to a better understanding of the link between ERVs to health and disease and ERV-host interactions. In fact, actively circulating retroviruses, or potentially mobilized ERVs, in other species present the potential for cross species transmission that have overwhelming consequences given the naivety of the new human host to the virus. The gain of viral genes in a ‘new’ background has the potential to offer new properties to the resulting recombinant; in particular gain of a new *env* gene may alter or expand the tropism of the virus to cross species transmission.

Some of the CfERV-Fc1 proviruses we have identified possess either completely intact or nearly intact open reading frames. Upon further investigation of these proviruses, we have shown the LTRs of individual proviruses possess a low number of base changes, and several are even identical, suggesting infection within at least the last ~1.5 my ago, making Fc1 the most recent retroviral lineage to invade the canine germline. We also identified solo LTRs that are identical to some of the proviral LTRs, suggesting closely related Fc1 haplotypes were infecting canine ancestors over a similar timeframe. Additionally, the *env* gene present in the recently found proviruses show to be specific to the canine species again suggesting a more recent infection of exogenous retrovirus than was previously theorized. These observations lead us to believe that these are recent insertions, and should therefore be the most likely candidate ERVs to have impacted on the genome structure of canines.

Previously, there have been reports of reverse transcriptase activity as well as gamma-type C particles in canine lymphoma tumor tissues [56], but these reports were never substantiated. Since there are no known active exogenous retroviruses found in the canine model, and the appearance of only highly mutated insertions, the source of this expression remains a mystery. We hypothesize these observations are directly associated with endogenous Fc1 in those samples. Understanding the association of these ERVs within the canine host should help us understand what these CfERVs have had on canine physiology and evolution. Dogs and humans both share similarities in their pathologies in spontaneous tumors which include parallels between the molecular profile, histology, genetics and response to treatment [57]. This puts the dog a unique niche to help provide a model that allows us to investigate these viruses and a possible link to disease. As part of my contributions to this research focus in the Halo lab, I have analyzed Fc1 expression in tumor-derived and normal canine cell lines, and I have begun to analyze the expression of individual Fc1 proviral loci in tumors and matched normal tissues to help us better understand the biological implications of endogenous retroviruses in the canine [57].

MATERIALS AND METHODS

Genetic Material

The source of all genetic material in this project comes from cancer derived cell line bought from ATCC and tumor tissues from our collaborators at Ridgeway Pet Hospital (Bowling Green, OH). The four cell lines being included in this study include: A72, DH82, D17 and MDCK. Each cell line provides different sources and a unique disease or 'normal' state (Table 2). For all experiments, cells were grown and maintained at 37°C with an atmosphere of 5% CO_2**,** with the exception of A72 which require an atmosphere of 0% CO_2**.** Each cell type requires different complete growth medias for culture. A72 will be maintained in Leibovitz media with a final concentration of 10% Fetal Bovine Serum (FBS) and 2% PenStrep. DH82 cells require Eagle's minimal essential medium, with 15% heat-inactivated FBS and 2% PenStrep. Both D17 and MDCK cells will be cultured in Eagle's minimal essential medium with a final concentration of 10% FBS and 2% PenStrep. FBS is added to provide embryonic growth promoting factors along with necessary hormones and attachment factors. Addition of PenStrep antibiotics prevents bacterial contamination during tissue culture. All the cells were subcultured at about 80% confluency in a 1:4 ratio and will be performed in a sterile, air controlled hood.

It is from these cells where genomic DNA was extracted using the NucleoSpin gDNA extraction kit (Machery-Nagel), yielding ~150ng/ul of dsDNA. RNA extraction was simultaneously performed on the same cell lines using NucleoZol RNA extraction kit (Machery-Nagel) followed by immediate reverse transcription using the M-MuLV Reverse Transcriptase kit and protocols as suggested by the manufacturer (New England Biolabs). All remaining RNA was stored at -80°C and the newly reverse transcribed cDNA stored at -20°C. A *Taq* PCR reaction was ran with the newly reverse transcribed cDNA alongside the negative control for reverse transcription using GAPDH primers as a control. GAPDH is a good candidate to test the cDNA because it is involved in the process of glycolysis and is highly expressed in our cell types and tissues. Including this step allowed us to examine the quality of the extracted RNA. cDNA was only be used in further analyses if the PCR shows no gDNA background in the RNA reaction well.

Table 2. Description of Canine derived Cell lines used in the study.

	Breed	**Tissue**	**Disease**	**Morphology**
A72	Golden Retriever	Unknown	Tumor	Fibroblast
DH82	Golden Retriever	Unknown	Malignant	Microphage-like
D17	Dalmatian	Bone	Osteosarcoma	Epithelial
MDCK	Unknown	Kidney	Normal	Epithelial

Primer Design for Expression

Primers were designed to amplify viral gene expression using the Primer 3 program (http://frodo.wi.mit.edu/primer3). An alignment was generated in BioEdit of the full-length proviruses (including all present in the CanFam3.1 reference genome and from new discoveries by the Halo lab) and annotated, from which 300-400 base pair segments corresponding to the 3'end of *pol* and 3'end of *env* were identified. One set of primers were aimed to amplify a highly conserved region of *pol*, which is the most conserved gene of retroviruses. Another set of primers placed inside the common deletion within the *env* gene; therefore, the only *env* amplified should come from proviruses with a complete *env* gene. The primers generated were further assessed using an *in silico* PCR over the CanFam3.1 reference genome (https://genome.ucsc.edu/). This step allowed us to obtain both the size of the amplified product and predicted melting temperature of the primer pairs.

cDNA Amplification and Cloning

We utilized extracted RNA to observe and analyze the expression patterns of individual Fc1 proviruses in cell based tissue culture. cDNA was obtained through reverse transcription of the cell line's extracted RNA, and utilized as a template in a 50uL Invitrogen *Taq* based PCR reaction with 10x buffer, 2.5uM dNTPs, 10Um each primer, 2.5uM MgCl2, and x *Taq*. Reactions were performed in an Eppendorf Mastercyler under the following conditions: initial denaturation was at 95°C for 2 minutes followed by 35 cycles of 95°C for 30sec, 59°C for 30 seconds for annealing of primers, with an extension time of 1:15 minutes at 72°C. For the final extension, the temperature was set at 72°C for 3 mins. 10uL of the PCR reaction was then assessed by gel electrophoreses using a 1% agarose in 1 x TBE.

Once PCR products are visualized from cDNA, captured sequences were determined through being cloned into a TOPO vector. Before ligation of the product, the remaining PCR products were cleaned using Nucleospin Gel and PCR clean-up reagents and protocols (Machery-Nagel). The concentrations of the cleaned up PCR products were analyzed using a Nanodrop Lite (Thermo Fisher) in order to determine a 3:1 insert to vector ratio with the NEB kit manual calculation (www.neb.com/E1202). Ligations were allowed to run at 25°C for 10 minutes, followed by ice incubation for 2 minutes. 1 ul of the ligation was inserted into 25 ul of NEB 10-beta Competent *E. coli,* following the protocols provided by NEB. The resulting transformants were plated on Amp$_{150}$ LB plates and incubated at 37° overnight.

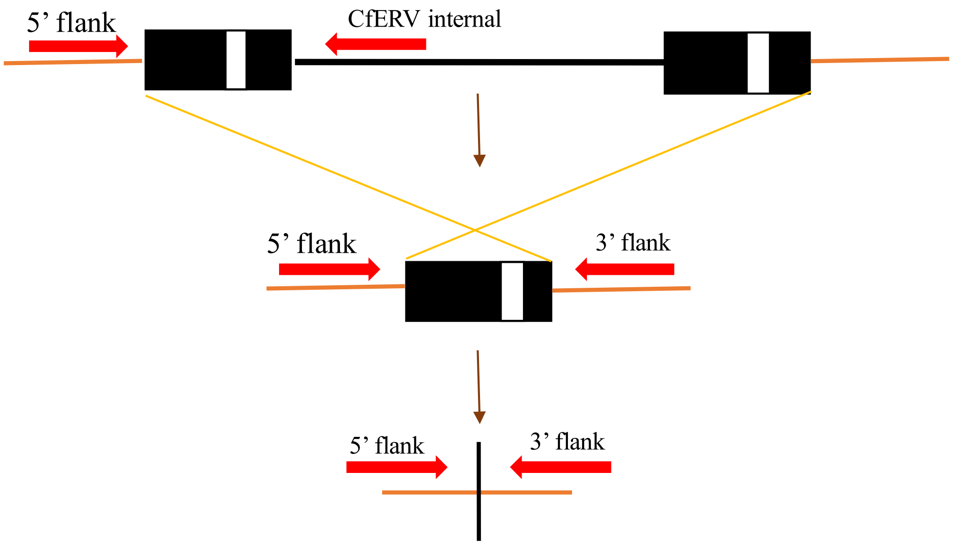
Colonies were selected and grown in 5 ml LB supplemented with Amp$_{100}$ overnight at 37°C. Using a Nucleospin miniplasmid kit (Machery-Nagel), the individual plasmids were purified to ensure the removal of cellular materials and the bacterial cell wall. Purified plasmids were then used as the template in an NEB *Taq* PCR reaction 10x buffer, 2.5uM dNTPs, 0.2uM each primer, and 0.125units/50uL reaction *Taq*. Vector-specific primers were used as recommended by NEB to amplify the inserted sequence. PCR reactions were then performed in an Eppendorf Mastercyler with the following conditions: initial denaturation was at 95°C for 30sec followed by 30 cycles of 95°C for 30sec, 59°C for 30 seconds for annealing of primers, with an extension time of 1 minutes at 68°C. For the final extension, temperature was set at 68°C for 5 mins. 10uL of the PCR reaction was then observed through electrophoreses using a 1% agarose in 1 x TBE. Once reproducible bands of the correct size were amplified, the PCR product was cleaned up and sequenced at University of Chicago Capillary Sequencing Center allowing individual sequences to be sequenced.

Bioinformatics Tools

Obtained sequences were analyzed using the BLAST-like alignment tool (https://genome.ucsc.edu/index.html) to the CanFam3.1 reference genome in order to confirm the correct sequence was transformed and therefore, aligned with sequences that BLAT to the same gene. The results were first aligned manually in BioEdit Sequence Alignment Editor as well as the MEGA 7 program. When 50 sequences were obtained for each gene from each cell line, the group was further aligned with the polymorphic proviruses found in that cell line based on genotyping results (also see below). A neighbor joining tree is computed in MEGA7. Findings will help out line details concerning the individual Fc1 proviruses that are expressed in the tested tissues.

Genotyping Primer Design

For each insertion, primers were designed to flank the predicted breakpoint based on the CanFam3.1 reference genome sequence flanking both the 5' and 3' LTR junctions. Amplification of either a pre-integration site or a solo LTR were detected by the flanking primers. Additional primers were designed to hybridize within the proviral leader sequence of the 5' end of the *gag* gene. These primers were within the provirus 5' untranslated region (outside of but near the 5' LTR), ranging from base 506 to 2210 from the start of the consensus CfERV Fc1. PCR of the 5' and 3' LTRs was used to infer the presence of a provirus as well as the orientation of the insertion. In the case we find any new proviruses in this screening, sequencing will be performed to reach 4-6x coverage over the full-length insertion allele, with precedence placed on the provirus for sites that are found to include both insertion alleles (Figure 9).



Primers were designed based of the CanFam(3.1) reference genome. Primers were designed to flank the specific proviral insertion at the 5’ and 3’ end; this would amplify either an “empty site” if there is no insertion present or a solo LTR. An additional primer was designed to be situated with in the insertion, there for if a provirus is present, the internal primer paired with a flanking primer would result in amplification of partial provirus.

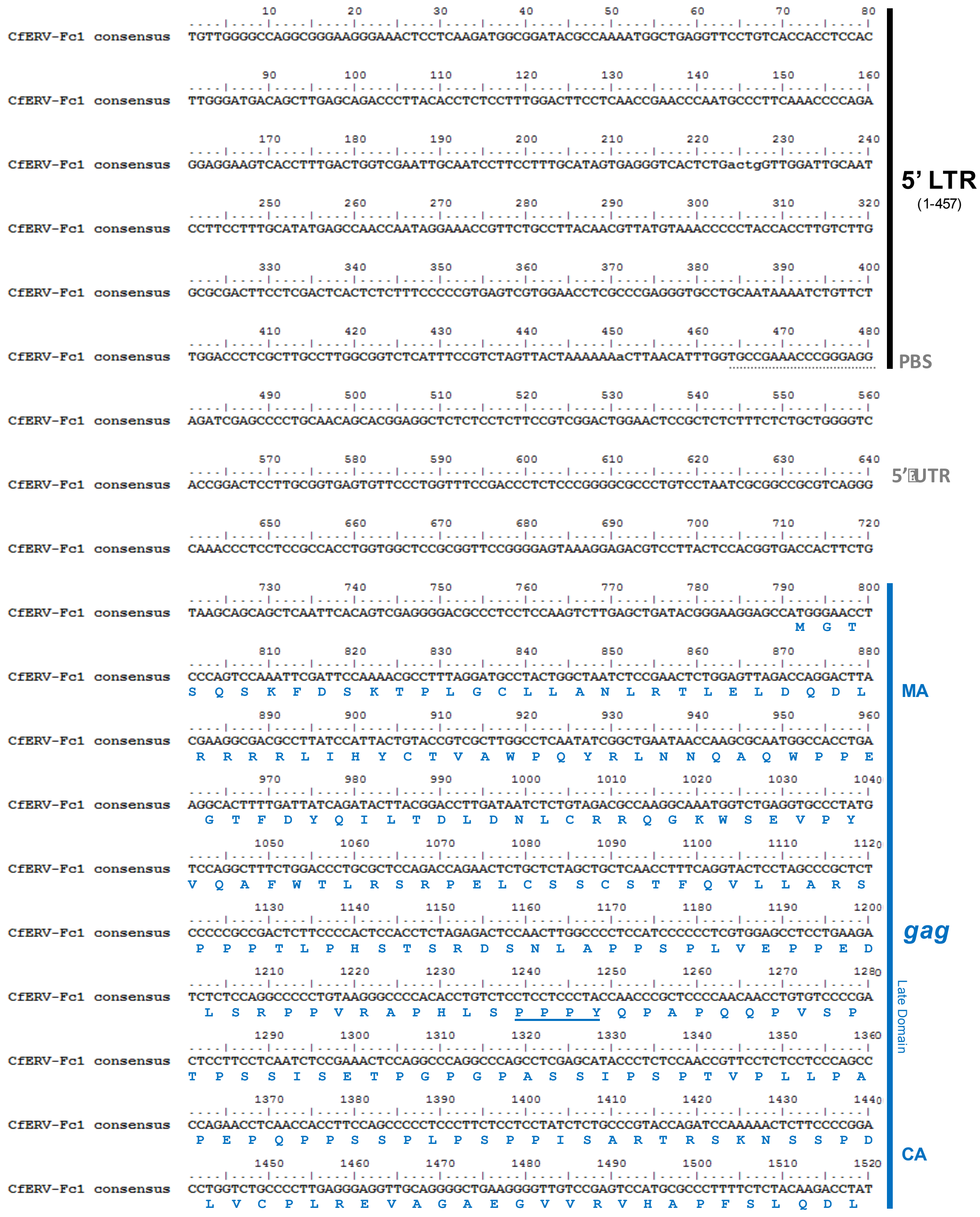
Genotyping PCR

Initial PCR was run with the two flanking primers of the provirus or solo LTR to examine the presence of a solo LTR or pre-integration site is present. The internal primer was then ran with the designated flanking primer in order to determine if a provirus is present in the genome, yielding a product ~1,000bp. Cell line gDNA was used as the template in an Invitrogen *Taq* based PCR reaction with 10x buffer, 2.5uM dNTPs, 10uM each primer, 2.5uM MgCl2, and x *Taq*. Reactions were performed in an Eppendorf Mastercyler with the following conditions: initial denaturation was at 95°C for 2 minutes followed by 35 cycles of 95°C for 30sec, 59°C for 30 seconds for annealing of primers, with an extension time of 1:15 minutes at 72°C. For the final extension, temperature was set at 72°C for 3 mins. 10uL of the PCR reaction was than observed through electrophoreses using a 1% agarose in 1 x TBE. This information will also help in improving the accuracy of the phylogenetic trees; because the recently integrated proviruses are at a polymorphic state, the expressed genes can be compared to only proviruses that are found to be present in that particular cell line or tumor genome.

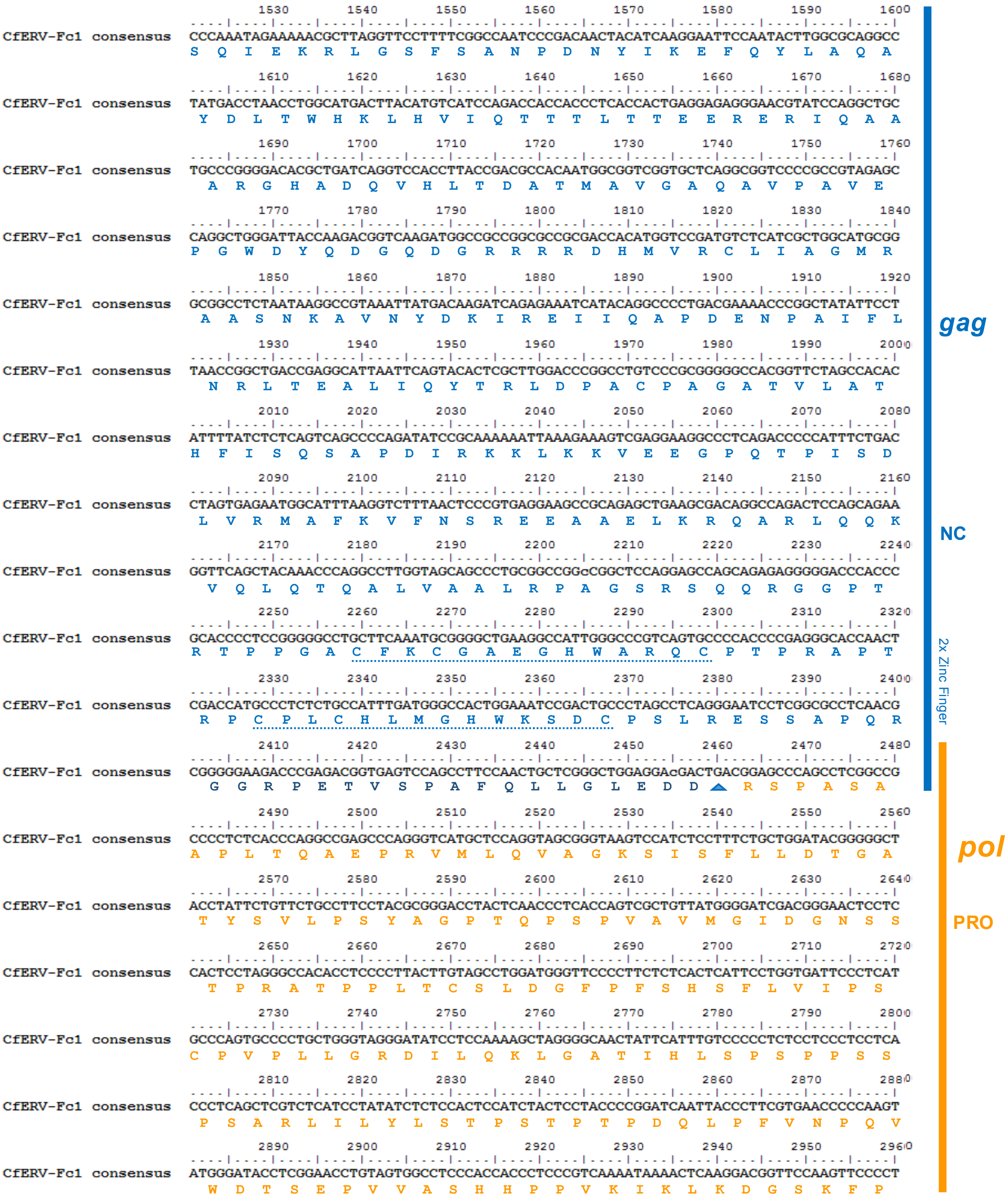
RESULTS

CfERV-Fc1(a)$_{\text{CON}}$

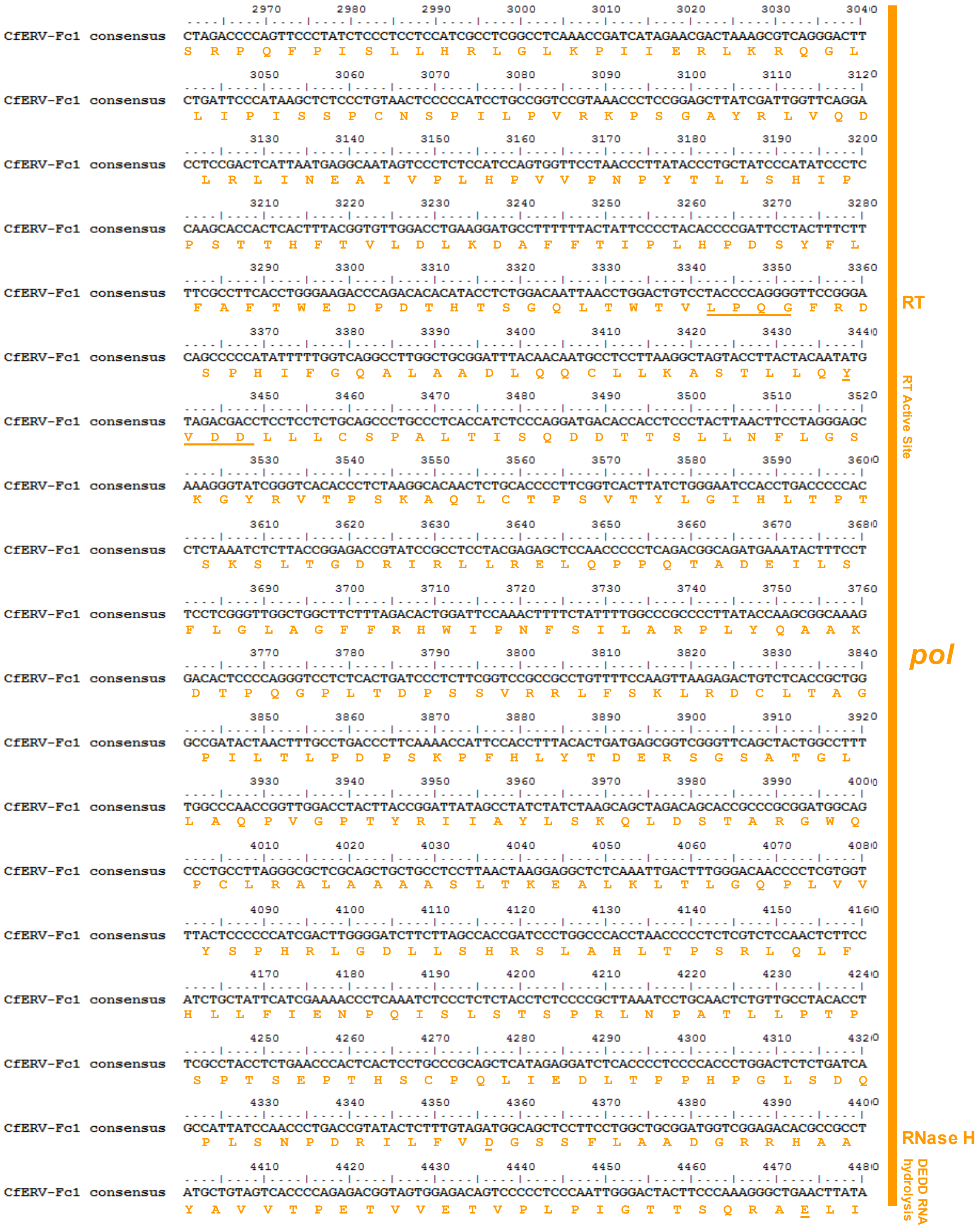
Based off of the 19 proviral sequences in *Canis familaris* a consensus was generated in silico from the nucleotide alignment based off the most represented nucleotide at each position and termed CfERV-Fc1(a)$_{\text{CON}}$ (Figure 10). The 19 proviruses included eleven present in the CanFam3.1 reference and eight non-reference insertions. CfERV-Fc1(a)$_{\text{CON}}$ had complete open reading frames in *gag* (~1.67kb), *pol* (~3.54kb) and *env* (~1.73kb). The proviral coding sequence was flanked by two identical LTRs, possessing a GAA anticodon that acts as a binding site for reverse transcriptase along with the expected 5'-TG…CA-3'. In the *gag* coding region contained all the expected structural motifs for the matrix, capsid, and nucleocapsid. Matrix encoded the expected motifs including both the PPPY late domain involved in particle release and the N-terminal glycine site of myristoylation that facilitated *Gag*-cell membrane association. Although in the majority of gamma like retroviruses code for three RNA binding zinc finger motif of CCHC-type domains, two where identified in nuclocapsid of the CfERV-Fc1(a)$_{\text{CON}}$. Characteristic of the gammaretroviral organization, the beginning of the *pol* gene begins immediately after the gag stop codon remaining in the same reading frame. Conserved motifs for protease, reverse transcriptase (including the LPQG and YVDD motifs), Rnase H (catalytic DEDD center of RNA Hydrolysis as well), and integrase (the DDX35E protease resistant core and N terminal HHCC DNA binding motif) were identified. *Env* had an alternate ORF that overlaps the 3' end of *pol*. The predictive product included the RRKR furin cleavage site of the SU and TM were present. As expected, the CWIC, and CX6CC motifs essential for SU-TM interaction, the immunosuppressive domain and the RD114-and-D-type receptor binding motif [55] were present as well.



```
                              10        20        30        40        50        60        70        80
                      ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
CfERV-Fc1 consensus   TGTTGGGGCCAGGCGGGAAGGGAAACTCCTCAAGATGGCGGATACGCCAAAATGGCTGAGGTTCCTGTCACCACCTCCAC

                              90       100       110       120       130       140       150       160
                      ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
CfERV-Fc1 consensus   TTGGGATGACAGCTTGAGCAGACCCTTACACCTCTCCTTTGGACTTCCTCAACCGAACCCAATGCCCTTCAAACCCCAGA

                             170       180       190       200       210       220       230       240
                      ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
CfERV-Fc1 consensus   GGAGGAAGTCACCTTTGACTGGTCGAATTGCAATCCTTCCTTTGCATAGTGAGGGTCACTCTGactgGTTGGATTGCAAT

                             250       260       270       280       290       300       310       320
                      ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
CfERV-Fc1 consensus   CCTTCCTTTGCATATGAGCCAACCAATAGGAAACCGTTCTGCCTTACAACGTTATGTAAACCCCCTACCACCTTGTCTTG

                             330       340       350       360       370       380       390       400
                      ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
CfERV-Fc1 consensus   GCGCGACTTCCTCGACTCACTCTCTTTCCCCCGTGAGTCGTGGAACCTCGCCCGAGGGTGCCTGCAATAAAATCTGTTCT

                             410       420       430       440       450       460       470       480
                      ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
CfERV-Fc1 consensus   TGGACCCTCGCTTGCCTTGGCGGTCTCATTTCCGTCTAGTTACTAAAAAAaCTTAACATTTGGTGCCGAAACCCGGGAGG

                             490       500       510       520       530       540       550       560
                      ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
CfERV-Fc1 consensus   AGATCGAGCCCCTGCAACAGCACGGAGGCTCTCTCCTCTTCCGTCGGACTGGAACTCCGCTCTCTTTCTCTGCTGGGGTC

                             570       580       590       600       610       620       630       640
                      ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
CfERV-Fc1 consensus   ACCGGACTCCTTGCGGTGAGTGTTCCCTGGTTTCCGACCCTCTCCCGGGGCGCCCTGTCCTAATCGCGGCCGCGTCAGGG

                             650       660       670       680       690       700       710       720
                      ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
CfERV-Fc1 consensus   CAAACCCTCCTCCGCCACCTGGTGGCTCCGCGGTTCCGGGGAGTAAAGGAGACGTCCTTACTCCACGGTGACCACTTCTG

                             730       740       750       760       770       780       790       800
                      ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
CfERV-Fc1 consensus   TAAGCAGCAGCTCAATTCACAGTCGAGGGGACGCCCTCCTCCAAGTCTTGAGCTGATACGGGAAGGAGCCATGGGAACCT
                                                                                            M  G  T

                             810       820       830       840       850       860       870       880
                      ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
CfERV-Fc1 consensus   CCCAGTCCAAATTCGATTCCAAAACGCCTTTAGGATGCCTACTGGCTAATCTCCGAACTCTGGAGTTAGACCAGGACTTA
                      S  Q  S  K  F  D  S  K  T  P  L  G  C  L  L  A  N  L  R  T  L  E  L  D  Q  D  L

                             890       900       910       920       930       940       950       960
                      ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
CfERV-Fc1 consensus   CGAAGGCGACGCCTTATCCATTACTGTACCGTCGCTTGGCCTCAATATCGGCTGAATAACCAAGCGCAATGGCCACCTGA
                       R  R  R  R  L  I  H  Y  C  T  V  A  W  P  Q  Y  R  L  N  N  Q  A  Q  W  P  P  E

                             970       980       990      1000      1010      1020      1030      1040
                      ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
CfERV-Fc1 consensus   AGGCACTTTTGATTATCAGATACTTACGGACCTTGATAATCTCTGTAGACGCCAAGGCAAATGGTCTGAGGTGCCCTATG
                        G  T  F  D  Y  Q  I  L  T  D  L  D  N  L  C  R  R  Q  G  K  W  S  E  V  P  Y

                            1050      1060      1070      1080      1090      1100      1110      1120
                      ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
CfERV-Fc1 consensus   TCCAGGCTTTCTGGACCCTGCGCTCCAGACCAGAACTCTGCTCTAGCTGCTCAACCTTTCAGGTACTCCTAGCCCGCTCT
                      V  Q  A  F  W  T  L  R  S  R  P  E  L  C  S  S  C  S  T  F  Q  V  L  L  A  R  S

                            1130      1140      1150      1160      1170      1180      1190      1200
                      ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
CfERV-Fc1 consensus   CCCCCGCCGACTCTTCCCCACTCCACCTCTAGAGACTCCAACTTGGCCCCTCCATCCCCCCTCGTGGAGCCTCCTGAAGA
                       P  P  P  T  L  P  H  S  T  S  R  D  S  N  L  A  P  P  S  P  L  V  E  P  P  E  D

                            1210      1220      1230      1240      1250      1260      1270      1280
                      ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
CfERV-Fc1 consensus   TCTCTCCAGGCCCCCTGTAAGGGCCCCACACCTGTCTCCTCCTCCCTACCAACCCGCTCCCCAACAACCTGTGTCCCCGA
                        L  S  R  P  P  V  R  A  P  H  L  S  P  P  P  Y  Q  P  A  P  Q  Q  P  V  S  P

                            1290      1300      1310      1320      1330      1340      1350      1360
                      ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
CfERV-Fc1 consensus   CTCCTTCCTCAATCTCCGAAACTCCAGGCCCAGGCCCAGCCTCGAGCATACCCTCTCCAACCGTTCCTCTCCTCCCAGCC
                      T  P  S  S  I  S  E  T  P  G  P  G  P  A  S  S  I  P  S  P  T  V  P  L  L  P  A

                            1370      1380      1390      1400      1410      1420      1430      1440
                      ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
CfERV-Fc1 consensus   CCAGAACCTCAACCACCTTCCAGCCCCCTCCCTTCTCCTCCTATCTCTGCCCGTACCAGATCCAAAAACTCTTCCCCGGA
                       P  E  P  Q  P  P  S  S  P  L  P  S  P  P  I  S  A  R  T  R  S  K  N  S  S  P  D

                            1450      1460      1470      1480      1490      1500      1510      1520
                      ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
CfERV-Fc1 consensus   CCTGGTCTGCCCCTTGAGGGAGGTTGCAGGGGCTGAAGGGGTTGTCCGAGTCCATGCGCCCTTTTCTCTACAAGACCTAT
                        L  V  C  P  L  R  E  V  A  G  A  E  G  V  V  R  V  H  A  P  F  S  L  Q  D  L
```



```
                          1530      1540      1550      1560      1570      1580      1590      1600
                    ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
CfERV-Fc1 consensus CCCAAATAGAAAAACGCTTAGGTTCCTTTTCGGCCAATCCCGACAACTACATCAAGGAATTCCAATACTTGGCGCAGGCC
                    S  Q  I  E  K  R  L  G  S  F  S  A  N  P  D  N  Y  I  K  E  F  Q  Y  L  A  Q  A

                          1610      1620      1630      1640      1650      1660      1670      1680
                    ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
CfERV-Fc1 consensus TATGACCTAACCTGGCATGACTTACATGTCATCCAGACCACCACCCTCACCACTGAGGAGAGGGAACGTATCCAGGCTGC
                     Y  D  L  T  W  H  K  L  H  V  I  Q  T  T  T  L  T  T  E  E  R  E  R  I  Q  A  A

                          1690      1700      1710      1720      1730      1740      1750      1760
                    ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
CfERV-Fc1 consensus TGCCCGGGGACACGCTGATCAGGTCCACCTTACCGACGCCACAATGGCGGTCGGTGCTCAGGCGGTCCCCGCCGTAGAGC
                      A  R  G  H  A  D  Q  V  H  L  T  D  A  T  M  A  V  G  A  Q  A  V  P  A  V  E

                          1770      1780      1790      1800      1810      1820      1830      1840
                    ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
CfERV-Fc1 consensus CAGGCTGGGATTACCAAGACGGTCAAGATGGCCGCCGGCGCCGCGACCACATGGTCCGATGTCTCATCGCTGGCATGCGG
                    P  G  W  D  Y  Q  D  G  Q  D  G  R  R  R  R  D  H  M  V  R  C  L  I  A  G  M  R

                          1850      1860      1870      1880      1890      1900      1910      1920
                    ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
CfERV-Fc1 consensus GCGGCCTCTAATAAGGCCGTAAATTATGACAAGATCAGAGAAATCATACAGGCCCCTGACGAAAACCCGGCTATATTCCT
                     A  A  S  N  K  A  V  N  Y  D  K  I  R  E  I  I  Q  A  P  D  E  N  P  A  I  F  L

                          1930      1940      1950      1960      1970      1980      1990      2000
                    ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
CfERV-Fc1 consensus TAACCGGCTGACCGAGGCATTAATTCAGTACACTCGCTTGGACCCGGCCTGTCCCGCGGGGGCCACGGTTCTAGCCACAC
                      N  R  L  T  E  A  L  I  Q  Y  T  R  L  D  P  A  C  P  A  G  A  T  V  L  A  T

                          2010      2020      2030      2040      2050      2060      2070      2080
                    ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
CfERV-Fc1 consensus ATTTTATCTCTCAGTCAGCCCCAGATATCCGCAAAAAATTAAAGAAAGTCGAGGAAGGCCCTCAGACCCCCATTTCTGAC
                    H  F  I  S  Q  S  A  P  D  I  R  K  K  L  K  K  V  E  E  G  P  Q  T  P  I  S  D

                          2090      2100      2110      2120      2130      2140      2150      2160
                    ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
CfERV-Fc1 consensus CTAGTGAGAATGGCATTTAAGGTCTTTAACTCCCGTGAGGAAGCCGCAGAGCTGAAGCGACAGGCCAGACTCCAGCAGAA
                     L  V  R  M  A  F  K  V  F  N  S  R  E  E  A  A  E  L  K  R  Q  A  R  L  Q  Q  K

                          2170      2180      2190      2200      2210      2220      2230      2240
                    ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
CfERV-Fc1 consensus GGTTCAGCTACAAACCCAGGCCTTGGTAGCAGCCCTGCGGCCGGcCGGCTCCAGGAGCCAGCAGAGAGGGGGACCCACCC
                      V  Q  L  Q  T  Q  A  L  V  A  A  L  R  P  A  G  S  R  S  Q  Q  R  G  G  P  T

                          2250      2260      2270      2280      2290      2300      2310      2320
                    ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
CfERV-Fc1 consensus GCACCCCTCCGGGGGCCTGCTTCAAATGCGGGGCTGAAGGCCATTGGGCCCGTCAGTGCCCCACCCCGAGGGCACCAACT
                    R  T  P  P  G  A  C  F  K  C  G  A  E  G  H  W  A  R  Q  C  P  T  P  R  A  P  T
                                      ..........................................

                          2330      2340      2350      2360      2370      2380      2390      2400
                    ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
CfERV-Fc1 consensus CGACCATGCCCTCTCTGCCATTTGATGGGCCACTGGAAATCCGACTGCCCTAGCCTCAGGGAATCCTCGGCGCCTCAACG
                     R  P  C  P  L  C  H  L  M  G  H  W  K  S  D  C  P  S  L  R  E  S  S  A  P  Q  R
                           ..........................................

                          2410      2420      2430      2440      2450      2460      2470      2480
                    ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
CfERV-Fc1 consensus CGGGGGAAGACCCGAGACGGTGAGTCCAGCCTTCCAACTGCTCGGGCTGGAGGAcGACTGACGGAGCCCAGCCTCGGCCG
                      G  G  R  P  E  T  V  S  P  A  F  Q  L  L  G  L  E  D  D  ▲  R  S  P  A  S  A

                          2490      2500      2510      2520      2530      2540      2550      2560
                    ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
CfERV-Fc1 consensus CCCCTCTCACCCAGGCCGAGCCCAGGGTCATGCTCCAGGTAGCGGGTAAGTCCATCTCCTTTCTGCTGGATACGGGGGCT
                    A  P  L  T  Q  A  E  P  R  V  M  L  Q  V  A  G  K  S  I  S  F  L  L  D  T  G  A

                          2570      2580      2590      2600      2610      2620      2630      2640
                    ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
CfERV-Fc1 consensus ACCTATTCTGTTCTGCCTTCCTACGCGGGACCTACTCAACCCTCACCAGTCGCTGTTATGGGGATCGACGGGAACTCCTC
                     T  Y  S  V  L  P  S  Y  A  G  P  T  Q  P  S  P  V  A  V  M  G  I  D  G  N  S  S

                          2650      2660      2670      2680      2690      2700      2710      2720
                    ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
CfERV-Fc1 consensus CACTCCTAGGGCCACACCTCCCCTTACTTGTAGCCTGGATGGGTTCCCCTTCTCTCACTCATTCCTGGTGATTCCCTCAT
                      T  P  R  A  T  P  P  L  T  C  S  L  D  G  F  P  F  S  H  S  F  L  V  I  P  S

                          2730      2740      2750      2760      2770      2780      2790      2800
                    ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
CfERV-Fc1 consensus GCCCAGTGCCCCTGCTGGGTAGGGATATCCTCCAAAAGCTAGGGGCAACTATTCATTTGTCCCCCTCTCCTCCCTCCTCA
                    C  P  V  P  L  L  G  R  D  I  L  Q  K  L  G  A  T  I  H  L  S  P  S  P  P  S  S

                          2810      2820      2830      2840      2850      2860      2870      2880
                    ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
CfERV-Fc1 consensus CCCTCAGCTCGTCTCATCCTATATCTCTCCACTCCATCTACTCCTACCCCGGATCAATTACCCTTCGTGAACCCCCAAGT
                     P  S  A  R  L  I  L  Y  L  S  T  P  S  T  P  T  P  D  Q  L  P  F  V  N  P  Q  V

                          2890      2900      2910      2920      2930      2940      2950      2960
                    ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
CfERV-Fc1 consensus ATGGGATACCTCGGAACCTGTAGTGGCCTCCCACCACCCTCCCGTCAAAATAAAACTCAAGGACGGTTCCAAGTTCCCCT
                      W  D  T  S  E  P  V  V  A  S  H  H  P  P  V  K  I  K  L  K  D  G  S  K  F  P
```



```
                             2970      2980      2990      3000      3010      3020      3030      3040
                    ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
CfERV-Fc1 consensus CTAGACCCCAGTTCCCTATCTCCCTCCTCCATCGCCTCGGCCTCAAACCGATCATAGAACGACTAAAGCGTCAGGGACTT
                    S  R  P  Q  F  P  I  S  L  L  H  R  L  G  L  K  P  I  I  E  R  L  K  R  Q  G  L
                             3050      3060      3070      3080      3090      3100      3110      3120
                    ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
CfERV-Fc1 consensus CTGATTCCCATAAGCTCTCCCTGTAACTCCCCCATCCTGCCGGTCCGTAAACCCTCCGGAGCTTATCGATTGGTTCAGGA
                     L  I  P  I  S  S  P  C  N  S  P  I  L  P  V  R  K  P  S  G  A  Y  R  L  V  Q  D
                             3130      3140      3150      3160      3170      3180      3190      3200
                    ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
CfERV-Fc1 consensus CCTCCGACTCATTAATGAGGCAATAGTCCCTCTCCATCCAGTGGTTCCTAACCCTTATACCCTGCTATCCCATATCCCTC
                      L  R  L  I  N  E  A  I  V  P  L  H  P  V  V  P  N  P  Y  T  L  L  S  H  I  P
                             3210      3220      3230      3240      3250      3260      3270      3280
                    ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
CfERV-Fc1 consensus CAAGCACCACTCACTTTACGGTGTTGGACCTGAAGGATGCCTTTTTTACTATTCCCCTACACCCCGATTCCTACTTTCTT
                    P  S  T  T  H  F  T  V  L  D  L  K  D  A  F  F  T  I  P  L  H  P  D  S  Y  F  L
                             3290      3300      3310      3320      3330      3340      3350      3360
                    ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
CfERV-Fc1 consensus TTCGCCTTCACCTGGGAAGACCCAGACACACATACCTCTGGACAATTAACCTGGACTGTCCTACCCCAGGGGTTCCGGGA
                     F  A  F  T  W  E  D  P  D  T  H  T  S  G  Q  L  T  W  T  V  L  P  Q  G  F  R  D
                             3370      3380      3390      3400      3410      3420      3430      3440
                    ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
CfERV-Fc1 consensus CAGCCCCCATATTTTTGGTCAGGCCTTGGCTGCGGATTTACAACAATGCCTCCTTAAGGCTAGTACCTTACTACAATATG
                      S  P  H  I  F  G  Q  A  L  A  A  D  L  Q  Q  C  L  L  K  A  S  T  L  L  Q  Y
                             3450      3460      3470      3480      3490      3500      3510      3520
                    ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
CfERV-Fc1 consensus TAGACGACCTCCTCCTCTGCAGCCCTGCCCTCACCATCTCCCAGGATGACACCACCTCCCTACTTAACTTCCTAGGGAGC
                    V  D  D  L  L  L  C  S  P  A  L  T  I  S  Q  D  D  T  T  S  L  L  N  F  L  G  S
                             3530      3540      3550      3560      3570      3580      3590      3600
                    ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
CfERV-Fc1 consensus AAAGGGTATCGGGTCACACCCTCTAAGGCACAACTCTGCACCCCTTCGGTCACTTATCTGGGAATCCACCTGACCCCCAC
                     K  G  Y  R  V  T  P  S  K  A  Q  L  C  T  P  S  V  T  Y  L  G  I  H  L  T  P  T
                             3610      3620      3630      3640      3650      3660      3670      3680
                    ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
CfERV-Fc1 consensus CTCTAAATCTCTTACCGGAGACCGTATCCGCCTCCTACGAGAGCTCCAACCCCCTCAGACGGCAGATGAAATACTTTCCT
                      S  K  S  L  T  G  D  R  I  R  L  L  R  E  L  Q  P  P  Q  T  A  D  E  I  L  S
                             3690      3700      3710      3720      3730      3740      3750      3760
                    ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
CfERV-Fc1 consensus TCCTCGGGTTGGCTGGCTTCTTTAGACACTGGATTCCAAACTTTTCTATTTTGGCCCGCCCCTTATACCAAGCGGCAAAG
                    F  L  G  L  A  G  F  F  R  H  W  I  P  N  F  S  I  L  A  R  P  L  Y  Q  A  A  K
                             3770      3780      3790      3800      3810      3820      3830      3840
                    ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
CfERV-Fc1 consensus GACACTCCCCAGGGTCCTCTCACTGATCCCTCTTCGGTCCGCCGCCTGTTTTCCAAGTTAAGAGACTGTCTCACCGCTGG
                     D  T  P  Q  G  P  L  T  D  P  S  S  V  R  R  L  F  S  K  L  R  D  C  L  T  A  G
                             3850      3860      3870      3880      3890      3900      3910      3920
                    ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
CfERV-Fc1 consensus GCCGATACTAACTTTGCCTGACCCTTCAAAACCATTCCACCTTTACACTGATGAGCGGTCGGGTTCAGCTACTGGCCTTT
                      P  I  L  T  L  P  D  P  S  K  P  F  H  L  Y  T  D  E  R  S  G  S  A  T  G  L
                             3930      3940      3950      3960      3970      3980      3990      4000
                    ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
CfERV-Fc1 consensus TGGCCCAACCGGTTGGACCTACTTACCGGATTATAGCCTATCTATCTAAGCAGCTAGACAGCACCGCCCGCGGATGGCAG
                    L  A  Q  P  V  G  P  T  Y  R  I  I  A  Y  L  S  K  Q  L  D  S  T  A  R  G  W  Q
                             4010      4020      4030      4040      4050      4060      4070      4080
                    ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
CfERV-Fc1 consensus CCCTGCCTTAGGGCGCTCGCAGCTGCTGCCTCCTTAACTAAGGAGGCTCTCAAATTGACTTTGGGACAACCCCTCGTGGT
                     P  C  L  R  A  L  A  A  A  A  S  L  T  K  E  A  L  K  L  T  L  G  Q  P  L  V  V
                             4090      4100      4110      4120      4130      4140      4150      4160
                    ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
CfERV-Fc1 consensus TTACTCCCCCCATCGACTTGGGGATCTTCTTAGCCACCGATCCCTGGCCCACCTAACCCCCTCTCGTCTCCAACTCTTCC
                      Y  S  P  H  R  L  G  D  L  L  S  H  R  S  L  A  H  L  T  P  S  R  L  Q  L  F
                             4170      4180      4190      4200      4210      4220      4230      4240
                    ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
CfERV-Fc1 consensus ATCTGCTATTCATCGAAAACCCTCAAATCTCCCTCTCTACCTCTCCCCGCTTAAATCCTGCAACTCTGTTGCCTACACCT
                    H  L  L  F  I  E  N  P  Q  I  S  L  S  T  S  P  R  L  N  P  A  T  L  L  P  T  P
                             4250      4260      4270      4280      4290      4300      4310      4320
                    ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
CfERV-Fc1 consensus TCGCCTACCTCTGAACCCACTCACTCCTGCCCGCAGCTCATAGAGGATCTCACCCCTCCCCACCCTGGACTCTCTGATCA
                     S  P  T  S  E  P  T  H  S  C  P  Q  L  I  E  D  L  T  P  P  H  P  G  L  S  D  Q
                             4330      4340      4350      4360      4370      4380      4390      4400
                    ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
CfERV-Fc1 consensus GCCATTATCCAACCCTGACCGTATACTCTTTGTAGATGGCAGCTCCTTCCTGGCTGCGGATGGTCGGAGACACGCCGCCT
                      P  L  S  N  P  D  R  I  L  F  V  D  G  S  S  F  L  A  A  D  G  R  R  H  A  A
                             4410      4420      4430      4440      4450      4460      4470      4480
                    ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
CfERV-Fc1 consensus ATGCTGTAGTCACCCCAGAGACGGTAGTGGAGACAGTCCCCCTCCCAATTGGGACTACTTCCCAAAGGGCTGAACTTATA
                    Y  A  V  V  T  P  E  T  V  V  E  T  V  P  L  P  I  G  T  T  S  Q  R  A  E  L  I
```

```
                               4490      4500      4510      4520      4530      4540      4550      4560
                     ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
CfERV-Fc1 consensus  GCTCTCACCAGGGCTCTACATCTATCTAAGGGACAACGAGTCACCATCTACACCGACTCAAAATATGCCTATCTCATCGT
                      A  L  T  R  A  L  H  L  S  K  G  Q  R  V  T  I  Y  T  D  S  K  Y  A  Y  L  I  V

                               4570      4580      4590      4600      4610      4620      4630      4640
                     ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
CfERV-Fc1 consensus  TCATACTCATTCCGTCCTCTGGCAGGAGCGGGGATTTTTAACCACCAAGGGGACGCCTATAGTAAATGGACCTCTCATTG
                       H  T  H  S  V  L  W  Q  E  R  G  F  L  T  T  K  G  T  P  I  V  N  G  P  L  I

                               4650      4660      4670      4680      4690      4700      4710      4720
                     ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
CfERV-Fc1 consensus  CCAAATTGCTTGAGGCCCTTAGTCTGCCCACTGAGGTTGCAATCGTTCACTGTAGGGGCCATCAGACTTCTAAAGACATG
                     A  K  L  L  E  A  L  S  L  P  T  E  V  A  I  V  H  C  R  G  H  Q  T  S  K  D  M

                               4730      4740      4750      4760      4770      4780      4790      4800
                     ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
CfERV-Fc1 consensus  GTCTCCATAGGAAATAATAAGGCCGACTCAGTGGCCAGGGAAACGGCCTTAAGTAACCCAATATCTCCCATCCTCTTCCT
                      V  S  I  G  N  N  K  A  D  S  V  A  R  E  T  A  L  S  N  P  I  S  P  I  L  F  L

                               4810      4820      4830      4840      4850      4860      4870      4880
                     ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
CfERV-Fc1 consensus  TAATACCCCTCATCGACCTTTCTACTCTATAAAGGAAACTCAAGCCCTCCAGGCCCTGGGAGGAAAGGCAGAAAGTAAAG
                       N  T  P  H  R  P  F  Y  S  I  K  E  T  Q  A  L  Q  A  L  G  G  K  A  E  S  K

                               4890      4900      4910      4920      4930      4940      4950      4960
                     ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
CfERV-Fc1 consensus  GATGGATTTACATCCGAGGGAAGATTGCCCTCCCAGAAAACCTGGCCCATACCCTAATTACTGATATCCACCAATCTCTC
                     G  W  I  Y  I  R  G  K  I  A  L  P  E  N  L  A  H  T  L  I  T  D  I  H  Q  S  L

                               4970      4980      4990      5000      5010      5020      5030      5040
                     ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
CfERV-Fc1 consensus  CATATTGGCCCAAGGGCACTGAACCAGTTTCTCCAGCCCCTGTTTTACTATCCATCCCTACCTAAGGTGATTGAGGCTGT
                      H  I  G  P  R  A  L  N  Q  F  L  Q  P  L  F  Y  Y  P  S  L  P  K  V  I  E  A  V

                               5050      5060      5070      5080      5090      5100      5110      5120
                     ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
CfERV-Fc1 consensus  CCATAGGGCTTGTAAAACCTGCTCGGCCGTAAATGCACAGGGAGGAATCCGCAGGCCGGGGCCTAACCATCAGCTCCGAG
                       H  R  A  C  K  T  C  S  A  V  N  A  Q  G  G  I  R  R  P  G  P  N  H  Q  L  R

                               5130      5140      5150      5160      5170      5180      5190      5200
                     ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
CfERV-Fc1 consensus  GCCATCAGCCCGGTGAAGACTGGCAGCTGGACTTCACTCACATGCCCCGCCATAAAGCCTTTCGTTATCTACTGACTTTG
                     G  H  Q  P  G  E  D  W  Q  L  D  F  T  H  M  P  R  H  K  A  F  R  Y  L  L  T  L

                               5210      5220      5230      5240      5250      5260      5270      5280
                     ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
CfERV-Fc1 consensus  GTTGATACTTTTACAGGATGGATTGAGGCATACCCCACAGCCAGAGAGACTGCAGATGTGGTGGCCACAATCCTCATCGA
                      V  D  T  F  T  G  W  I  E  A  Y  P  T  A  R  E  T  A  D  V  V  A  T  I  L  I  E

                               5290      5300      5310      5320      5330      5340      5350      5360
                     ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
CfERV-Fc1 consensus  GCACATCATCCCGAGGTTTGGGTTACCCCGGACCCTACAGTCAGACAACGGGCCGGCATTTATCTCCAGTGTGACCCAAC
                       H  I  I  P  R  F  G  L  P  R  T  L  Q  S  D  N  G  P  A  F  I  S  S  V  T  Q

                               5370      5380      5390      5400      5410      5420      5430      5440
                     ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
CfERV-Fc1 consensus  AGGTGGCCGAGAGCCTCAACATTACCTGGAAGCTGCACATCCCCTACCACCCTCAGTCTTCGGGTAAGGTGGAAAGGGCC
                     Q  V  A  E  S  L  N  I  T  W  K  L  H  I  P  Y  H  P  Q  S  S  G  K  V  E  R  A

                               5450      5460      5470      5480      5490      5500      5510      5520
                     ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
CfERV-Fc1 consensus  AACGGGCTACTTAAGGCTCAACTCACTAAACTTACCCTGGAGACTCGCCTGTCGTGGCCCACACTGTTACCTATAGCTCT
                      N  G  L  L  K  A  Q  L  T  K  L  T  L  E  T  R  L  S  W  P  T  L  L  P  I  A  L

                               5530      5540      5550      5560      5570      5580      5590      5600
                     ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
CfERV-Fc1 consensus  CACCAGACTCCGGGCgTCCCCCCGAGGACCATCAGGTTTGAGTCCCTTTGAGTTACTGTATGGTCGGCCCTTCCTTATCA
                       T  R  L  R  A  S  P  R  G  P  S  G  L  S  P  F  E  L  L  Y  G  R  P  F  L  I

                               5610      5620      5630      5640      5650      5660      5670      5680
                     ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
CfERV-Fc1 consensus  ACCACAACCTCCCGGCCATTCCTCCACCTCTTTTATCCTATCTGCCTTACCTTACCCTCCTCCGGGCCCTCTTGAGAGCC
                     N  H  N  L  P  A  I  P  P  P  L  L  S  Y  L  P  Y  L  T  L  L  R  A  L  L  R  A

                               5690      5700      5710      5720      5730      5740      5750      5760
                     ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
CfERV-Fc1 consensus  CATGCTGATGCTGTCATTCCGGCCCCGACCGACAATGCCTCCGAGGAAGCCTCTCCACGAGACTTATCCCCAGGGGACCA
                      H  A  D  A  V  I  P  A  P  T  D  N  A  S  E  E  A  S  P  R  D  L  S  P  G  D  Q
                       M  L  M  L  S  F  R  P  R  P  T  M  P  P  R  K  P  L  H  E  T  Y  P  Q  G  T

                               5770      5780      5790      5800      5810      5820      5830      5840
                     ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
CfERV-Fc1 consensus  GGTCCTTCTCAGAAATCTGCAGCCAGGCTCCCTGCAGACCCGATGGACCGGACCTCACACGGTCATCCTCACCACCCCAA
                       V  L  L  R  N  L  Q  P  G  S  L  Q  T  R  W  T  G  P  H  T  V  I  L  T  T  P
                     R  S  F  S  E  I  C  S  Q  A  P  C  R  P  D  G  P  D  L  T  R  S  S  S  P  P  Q

                               5850      5860      5870      5880      5890      5900      5910      5920
                     ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
CfERV-Fc1 consensus  CCGCGGCAAAATTGCTGGGACATACGGCATGGGTGCACATCAACAACCTTAAACGGGCACCCACAGGTATCGAATGGACC
                     T  A  A  K  L  L  G  H  T  A  W  V  H  I  N  N  L  K  R  A  P  T  G  I  E  W  T
                      P  R  Q  N  C  W  D  I  R  H  G  C  T  S  T  T  L  N  G  H  P  Q  V  S  N  G  P

                               5930      5940      5950      5960      5970      5980      5990      6000
                     ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
CfERV-Fc1 consensus  TCCCAGATGGTGGGACCCACCAAACTCCGCCTGGCTAGGGCTCCTTCTCATACTTCTCCTGAGCCCCCCGATCCAAGCGG
                      S  Q  M  V  G  P  T  K  L  R  L  A  R  A  P  S  H  T  S  P  E  P  P  D  P  S  G
                       P  R  W  W  D  P  P  N  S  A  W  L  G  L  L  L  I  L  L  L  S  P  P  I  Q  A
```

RNase H

H_2C_2 DNA Binding

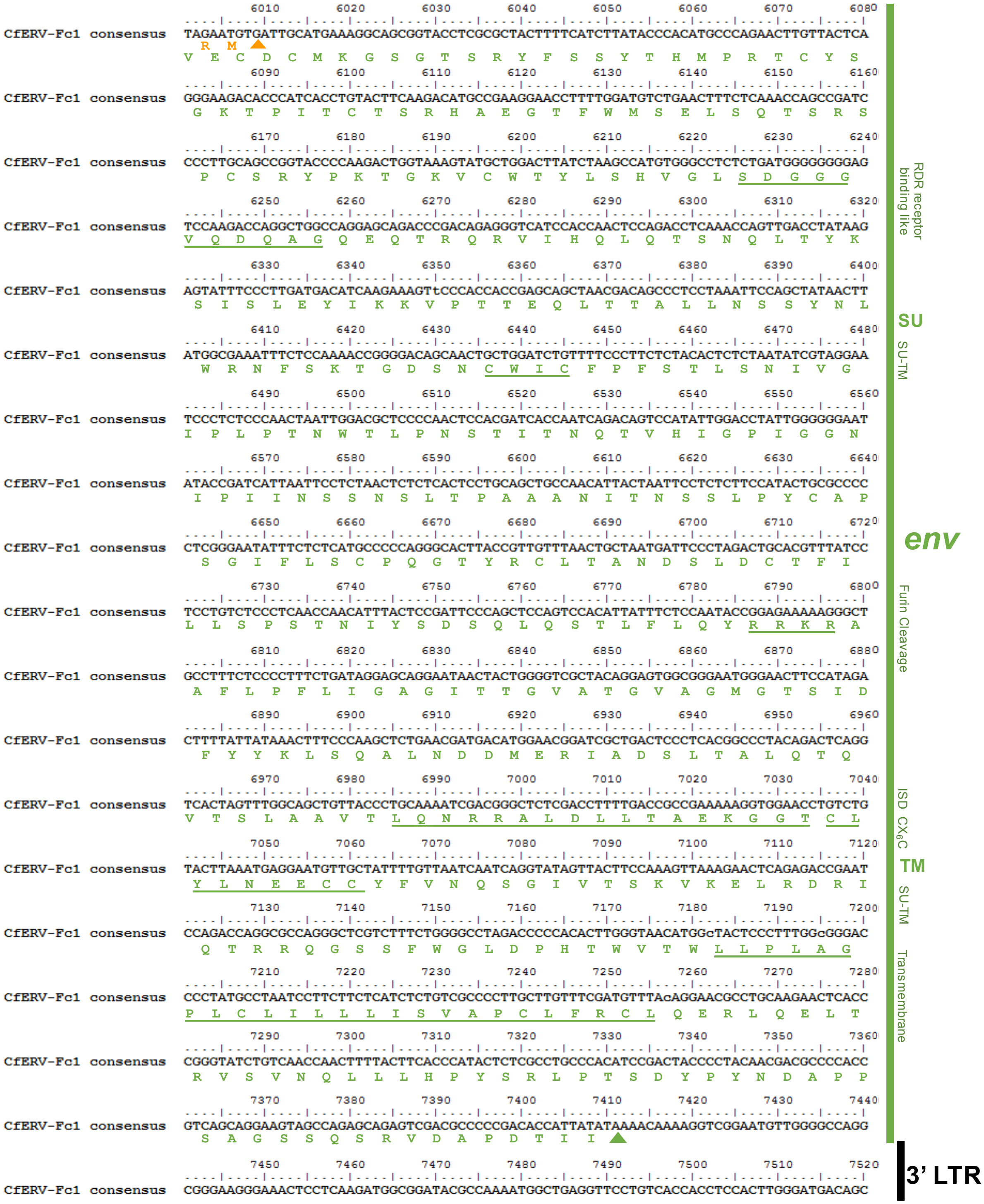
pol

IN

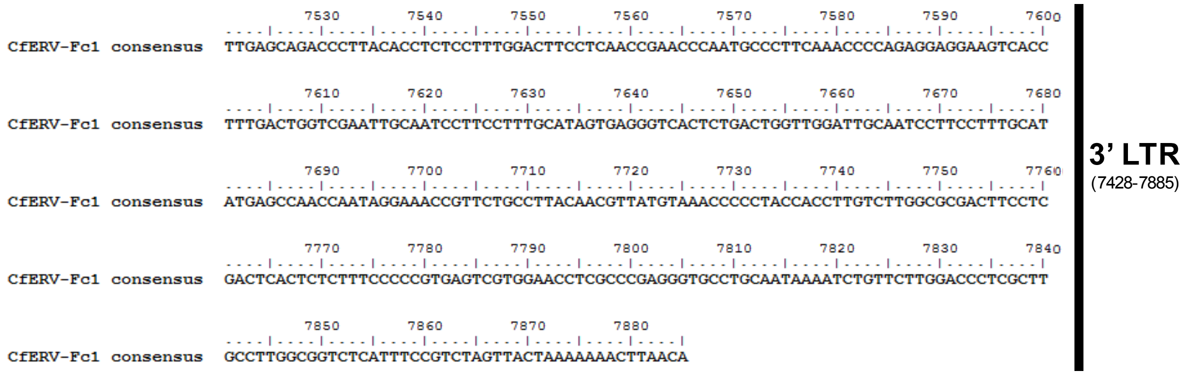
$DDX_{35}E$ protease resistant core

env

SU



```
                        6010      6020      6030      6040      6050      6060      6070      6080
CfERV-Fc1 consensus  TAGAATGTGATTGCATGAAAGGCAGCGGTACCTCGCGCTACTTTTCATCTTATACCCACATGCCCAGAACTTGTTACTCA
                     V  E  C  D  C  M  K  G  S  G  T  S  R  Y  F  S  S  Y  T  H  M  P  R  T  C  Y  S
                        6090      6100      6110      6120      6130      6140      6150      6160
CfERV-Fc1 consensus  GGGAAGACACCCATCACCTGTACTTCAAGACATGCCGAAGGAACCTTTTGGATGTCTGAACTTTCTCAAACCAGCCGATC
                      G  K  T  P  I  T  C  T  S  R  H  A  E  G  T  F  W  M  S  E  L  S  Q  T  S  R  S
                        6170      6180      6190      6200      6210      6220      6230      6240
CfERV-Fc1 consensus  CCCTTGCAGCCGGTACCCCAAGACTGGTAAAGTATGCTGGACTTATCTAAGCCATGTGGGCCTCTCTGATGGGGGGGGAG
                       P  C  S  R  Y  P  K  T  G  K  V  C  W  T  Y  L  S  H  V  G  L  S  D  G  G  G
                        6250      6260      6270      6280      6290      6300      6310      6320
CfERV-Fc1 consensus  TCCAAGACCAGGCTGGCCAGGAGCAGACCCGACAGAGGGTCATCCACCAACTCCAGACCTCAAACCAGTTGACCTATAAG
                     V  Q  D  Q  A  G  Q  E  Q  T  R  Q  R  V  I  H  Q  L  Q  T  S  N  Q  L  T  Y  K
                        6330      6340      6350      6360      6370      6380      6390      6400
CfERV-Fc1 consensus  AGTATTTCCCTTGATGACATCAAGAAAGTtCCCACCACCGAGCAGCTAACGACAGCCCTCCTAAATTCCAGCTATAACTT
                      S  I  S  L  E  Y  I  K  K  V  P  T  T  E  Q  L  T  T  A  L  L  N  S  S  Y  N  L
                        6410      6420      6430      6440      6450      6460      6470      6480
CfERV-Fc1 consensus  ATGGCGAAATTTCTCCAAAACCGGGGACAGCAACTGCTGGATCTGTTTTCCCTTCTCTACACTCTCTAATATCGTAGGAA
                       W  R  N  F  S  K  T  G  D  S  N  C  W  I  C  F  P  F  S  T  L  S  N  I  V  G
                        6490      6500      6510      6520      6530      6540      6550      6560
CfERV-Fc1 consensus  TCCCTCTCCCAACTAATTGGACGCTCCCCAACTCCACGATCACCAATCAGACAGTCCATATTGGACCTATTGGGGGGAAT
                     I  P  L  P  T  N  W  T  L  P  N  S  T  I  T  N  Q  T  V  H  I  G  P  I  G  G  N
                        6570      6580      6590      6600      6610      6620      6630      6640
CfERV-Fc1 consensus  ATACCGATCATTAATTCCTCTAACTCTCTCACTCCTGCAGCTGCCAACATTACTAATTCCTCTCTTCCATACTGCGCCCC
                      I  P  I  I  N  S  S  N  S  L  T  P  A  A  A  N  I  T  N  S  S  L  P  Y  C  A  P
                        6650      6660      6670      6680      6690      6700      6710      6720
CfERV-Fc1 consensus  CTCGGGAATATTTCTCTCATGCCCCCAGGGCACTTACCGTTGTTTAACTGCTAATGATTCCCTAGACTGCACGTTTATCC
                       S  G  I  F  L  S  C  P  Q  G  T  Y  R  C  L  T  A  N  D  S  L  D  C  T  F  I
                        6730      6740      6750      6760      6770      6780      6790      6800
CfERV-Fc1 consensus  TCCTGTCTCCCTCAACCAACATTTACTCCGATTCCCAGCTCCAGTCCACATTATTTCTCCAATACCGGAGAAAAAGGGCT
                     L  L  S  P  S  T  N  I  Y  S  D  S  Q  L  Q  S  T  L  F  L  Q  Y  R  R  K  R  A
                        6810      6820      6830      6840      6850      6860      6870      6880
CfERV-Fc1 consensus  GCCTTTCTCCCCTTTCTGATAGGAGCAGGAATAACTACTGGGGTCGCTACAGGAGTGGCGGGAATGGGAACTTCCATAGA
                      A  F  L  P  F  L  I  G  A  G  I  T  T  G  V  A  T  G  V  A  G  M  G  T  S  I  D
                        6890      6900      6910      6920      6930      6940      6950      6960
CfERV-Fc1 consensus  CTTTTATTATAAACTTTCCCAAGCTCTGAACGATGACATGGAACGGATCGCTGACTCCCTCACGGCCCTACAGACTCAGG
                       F  Y  Y  K  L  S  Q  A  L  N  D  D  M  E  R  I  A  D  S  L  T  A  L  Q  T  Q
                        6970      6980      6990      7000      7010      7020      7030      7040
CfERV-Fc1 consensus  TCACTAGTTTGGCAGCTGTTACCCTGCAAAATCGACGGGCTCTCGACCTTTTGACCGCCGAAAAAGGTGGAACCTGTCTG
                     V  T  S  L  A  A  V  T  L  Q  N  R  R  A  L  D  L  L  T  A  E  K  G  G  T  C  L
                        7050      7060      7070      7080      7090      7100      7110      7120
CfERV-Fc1 consensus  TACTTAAATGAGGAATGTTGCTATTTTGTTAATCAATCAGGTATAGTTACTTCCAAAGTTAAAGAACTCAGAGACCGAAT
                      Y  L  N  E  E  C  C  Y  F  V  N  Q  S  G  I  V  T  S  K  V  K  E  L  R  D  R  I
                        7130      7140      7150      7160      7170      7180      7190      7200
CfERV-Fc1 consensus  CCAGACCAGGCGCCAGGGCTCGTCTTTCTGGGGCCTAGACCCCCACACTTGGGTAACATGGcTACTCCCTTTGGcGGGAC
                       Q  T  R  R  Q  G  S  S  F  W  G  L  D  P  H  T  W  V  T  W  L  L  P  L  A  G
                        7210      7220      7230      7240      7250      7260      7270      7280
CfERV-Fc1 consensus  CCCTATGCCTAATCCTTCTTCTCATCTCTGTCGCCCCTTGCTTGTTTCGATGTTTAcAGGAACGCCTGCAAGAACTCACC
                     P  L  C  L  I  L  L  L  I  S  V  A  P  C  L  F  R  C  L  Q  E  R  L  Q  E  L  T
                        7290      7300      7310      7320      7330      7340      7350      7360
CfERV-Fc1 consensus  CGGGTATCTGTCAACCAACTTTTACTTCACCCATACTCTCGCCTGCCCACATCCGACTACCCCTACAACGACGCCCCACC
                      R  V  S  V  N  Q  L  L  L  H  P  Y  S  R  L  P  T  S  D  Y  P  Y  N  D  A  P  P
                        7370      7380      7390      7400      7410      7420      7430      7440
CfERV-Fc1 consensus  GTCAGCAGGAAGTAGCCAGAGCAGAGTCGACGCCCCCGACACCATTATATAAAACAAAAGGTCGGAATGTTGGGGCCAGG
                       S  A  G  S  S  Q  S  R  V  D  A  P  D  T  I  I
                        7450      7460      7470      7480      7490      7500      7510      7520
CfERV-Fc1 consensus  CGGGAAGGGAAACTCCTCAAGATGGCGGATACGCCAAAATGGCTGAGGTTCCTGTCACCACCTCCACTTGGGATGACAGC
```



```
                            7530      7540      7550      7560      7570      7580      7590      7600
                      ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
CfERV-Fc1 consensus   TTGAGCAGACCCTTACACCTCTCCTTTGGACTTCCTCAACCGAACCCAATGCCCTTCAAACCCCAGAGGAGGAAGTCACC

                            7610      7620      7630      7640      7650      7660      7670      7680
                      ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
CfERV-Fc1 consensus   TTTGACTGGTCGAATTGCAATCCTTCCTTTGCATAGTGAGGGTCACTCTGACTGGTTGGATTGCAATCCTTCCTTTGCAT

                            7690      7700      7710      7720      7730      7740      7750      7760
                      ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
CfERV-Fc1 consensus   ATGAGCCAACCAATAGGAAACCGTTCTGCCTTACAACGTTATGTAAACCCCCTACCACCTTGTCTTGGCGCGACTTCCTC

                            7770      7780      7790      7800      7810      7820      7830      7840
                      ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
CfERV-Fc1 consensus   GACTCACTCTCTTTCCCCCGTGAGTCGTGGAACCTCGCCCGAGGGTGCCTGCAATAAAATCTGTTCTTGGACCCTCGCTT

                            7850      7860      7870      7880
                      ....|....|....|....|....|....|....|....|....|
CfERV-Fc1 consensus   GCCTTGGCGGTCTCATTTCCGTCTAGTTACTAAAAAAACTTAACA
```

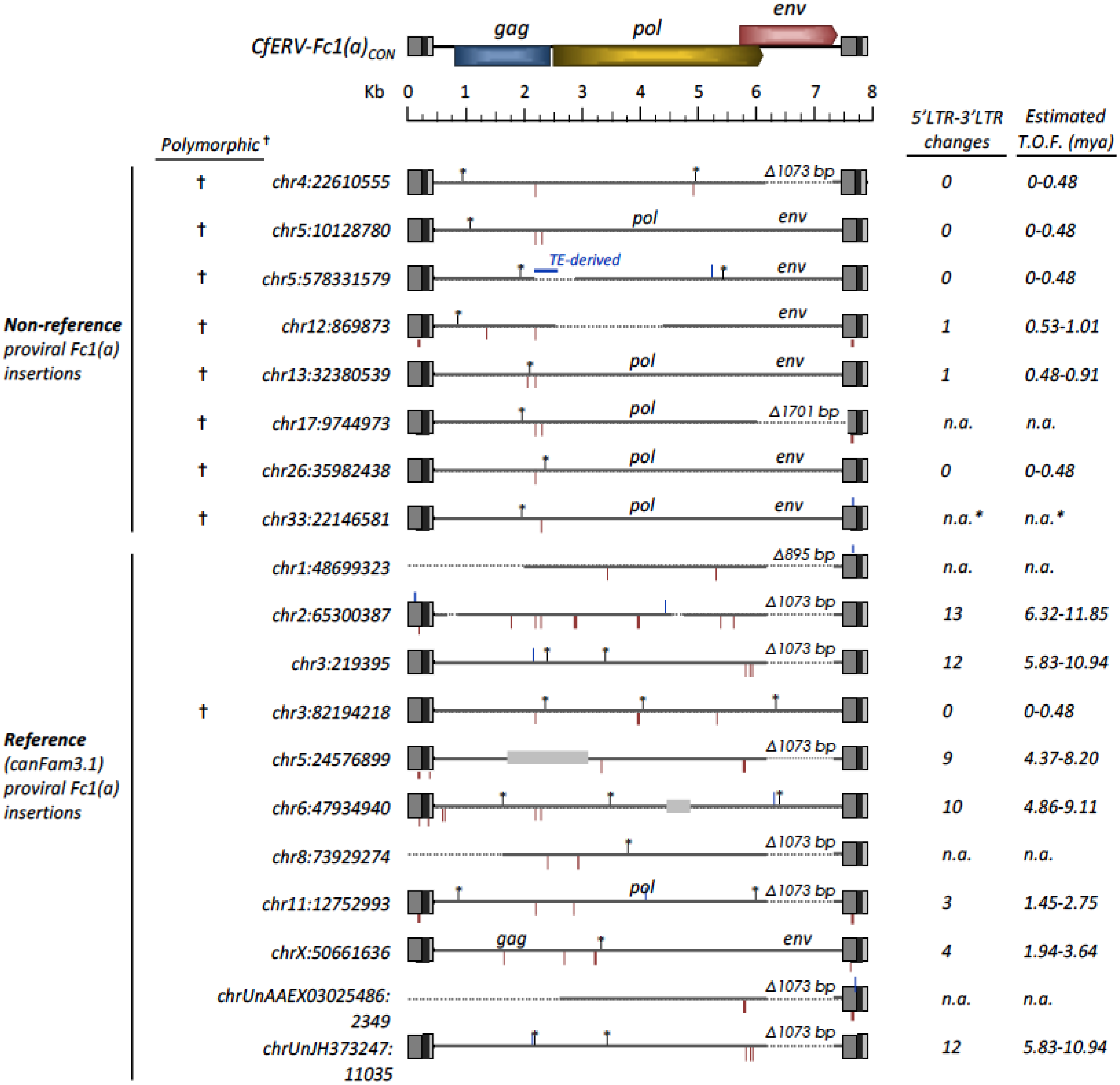
Figure 10. Annotation of CfERV-Fc1(a)CON.

CfERV-Fc1(a) genetic sequence is represented with the corresponding amino acid below the center (second base of the codon) in the correct reading frame. Colored bars are located on the right side which corresponds to either the LTR or gene coded for by the region. (LTR is represented by black, *gag* is represented in blue, *pol* represented in yellow, and *env* represented in green)

Individual CfERV-Fc1 proviruses

Properties identified in the CfERV-Fc1(a) in individual Fc1(a) proviruses were examined as well (Figure 11). In contrast to the CfERV-Fc1(a) consensus provirus, none of the nineteen individuals insertions had a complete open reading frame of the *gag* gene. Of the three genes, *gag* experienced more inactivating mutations, some of which were shared frame shifts leading to premature stops. Of all the proviruses, Chr3:82,194,218 and Chr26:35,982,438 contained the longest *gag* reading frame, sharing a premature stop codon with in the first zinc finer domain of the nucleocapsid. Absence of both zinc fingers would interrupt the ability of the gag gene to encapsidize the viral RNA, being that the zinc fingers provide the packaging signal. In total, there were six proviruses with a complete reading frame in *pol*. All six appeared to have the above mentioned domains, including RT, RnaseH, and integrase. There were no changes to indicate an altering of function.

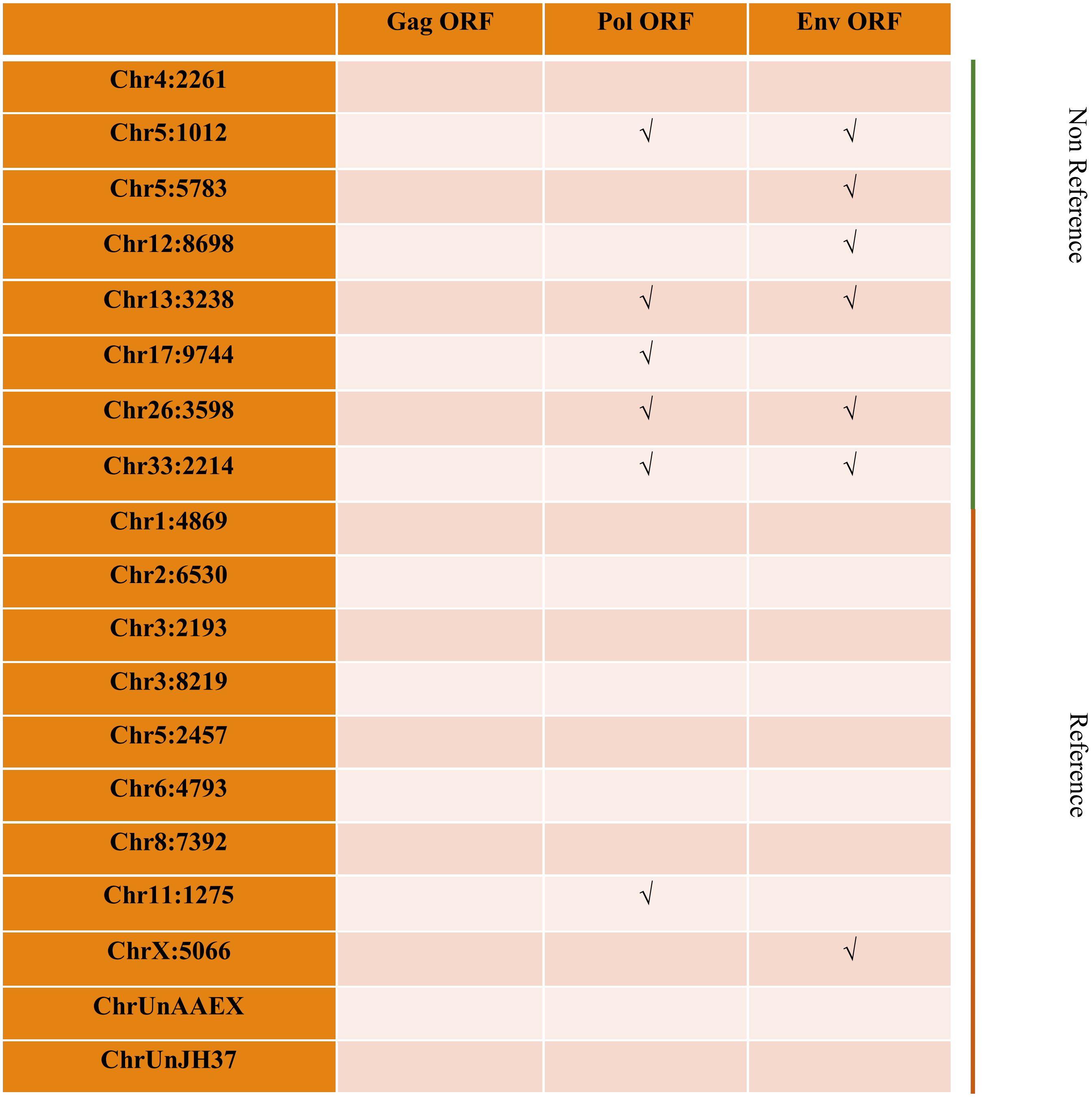
A reading frame for the *env* gene was present in seven of the proviruses. Of the seven proviruses, the fusion peptide, TM region and ISD showed to agree with what was present in the CfERV-Fc1(a)$_{\mathrm{CON}}$. Upon investigation of the *env* gene, we identified a common deletion present amongst eight of the proviruses that spans a 1,037 bp segment of *env*. As a consequence, the majority of the internal portions of SU and TM are eliminated in proviruses that possess the deletion. Portions missing include the coding regions for the RDR receptor binding domain, motifs involved in SU-TM interactions and the transmembrane domain. With the absence of these domains, the *env* gene would be unable to serve its canonical viral functions (*i.e.*, receptor binding and membrane fusion). This deletion is nearly ubiquitous amongst the older proviruses with the exception of chrX:50,661,636, in which a complete *env* open reading frame is present. Both chr3:82,194,218 and chr6:47,934,940, also lack this common deletion but both contain mutations leading to a premature stop codon.



	Polymorphic†		5'LTR-3'LTR changes	Estimated T.O.F. (mya)
Non-reference *proviral Fc1(a) insertions*	†	*chr4:22610555*	*0*	*0-0.48*
	†	*chr5:10128780*	*0*	*0-0.48*
	†	*chr5:578331579*	*0*	*0-0.48*
	†	*chr12:869873*	*1*	*0.53-1.01*
	†	*chr13:32380539*	*1*	*0.48-0.91*
	†	*chr17:9744973*	*n.a.*	*n.a.*
	†	*chr26:35982438*	*0*	*0-0.48*
	†	*chr33:22146581*	*n.a.**	*n.a.**
Reference *(canFam3.1) proviral Fc1(a) insertions*		*chr1:48699323*	*n.a.*	*n.a.*
		chr2:65300387	*13*	*6.32-11.85*
		chr3:219395	*12*	*5.83-10.94*
	†	*chr3:82194218*	*0*	*0-0.48*
		chr5:24576899	*9*	*4.37-8.20*
		chr6:47934940	*10*	*4.86-9.11*
		chr8:73929274	*n.a.*	*n.a.*
		chr11:12752993	*3*	*1.45-2.75*
		chrX:50661636	*4*	*1.94-3.64*
		chrUnAAEX03025486:2349	*n.a.*	*n.a.*
		chrUnJH373247:11035	*12*	*5.83-10.94*

Figure 11. Annotation of individual insertion site.

Annotated individual CfERV-Fc1 proviruses are depicted. Insertions are grouped into those found in the canFam3.1 (reference) and those not found in canFam3.1 (non reference). If insertions are polymorphic in canines, a + symbol appears before the insertion site. Present open reading frames are noted by the name of the gene above the bar. Number of differences between the 5'LTR and 3'LTR noted directly to the left of the insertion followed by the estimated time of infection. Picture from Halo, 2018 (unpublished). Individual motifs in the coding regions are underlined and labeled at the right of the figure.



	Gag ORF	Pol ORF	Env ORF	
Chr4:2261				Non Reference
Chr5:1012		√	√	
Chr5:5783			√	
Chr12:8698			√	
Chr13:3238		√	√	
Chr17:9744		√		
Chr26:3598		√	√	
Chr33:2214		√	√	
Chr1:4869				Reference
Chr2:6530				
Chr3:2193				
Chr3:8219				
Chr5:2457				
Chr6:4793				
Chr8:7392				
Chr11:1275		√		
ChrX:5066			√	
ChrUnAAEX				
ChrUnJH37				

Figure 12. Open reading frames present in proviral insertions.

Open reading frames were found in proviral insertions are noted by check marks is specific gene column. Proviruses are grouped by insertions found in the CanFam3.1 reference genome, marked by the side orange bar, and insertions not found in the CanFam3.1 referenced genome termed Non Reference and noted by the green bar.

Expression of ERVs

Fifty sequences each for the *pol* gene and *env* gene, respectively, were obtained from cell lines A72, DH82 and D17. Amplification of *pol* and *env* was in cell line MDCK was not achieved despite repeated attempts. Using MEGA7, Neighbor-Joining trees were created for the *pol* and *env* sequences obtained from each line. For *pol* sequences each gene found in the cell lines were first aligned with a decrepit provirus found on chr8:16,833,81 which was chosen to be used as the outgroup in the alignments being part of a related but distinct different subfamily known as CfERV-Fc1(b). The phylogenetic relationship of the *pol* sequences between individual CfERV-Fc1(a) insertion compared to the CfERV-Fc1(b) is represented in Figure 13. The neighbor joining tree shows the two lineages in distinct groups from one another. Two insertions in CfERV-Fc1, including chr3:21,939,61 and chrUN:JH,373,247, although group with the CfERV-Fc1(a), appear to be slightly diverged from the rest of the lineage. The insertion at site chr12:86,987,5 was left out of the phylogenetic trees comparing the *pol* gene because without chr8:16,833,81 present it was acting as an outgroup; our sequence inspection suggested a gene conservation at this site in *pol*. Gene conversion is a mechanism of homologous recombination that involves the unidirectional transfer of genetic sequences from one sequence to a homologous site, which masks what the insertions true sequences was in its original state.

Pol Gene

We used primers to amplify the conserved region of *pol* encoding RNase H activity in which to build phylogenies. A72 sequences in a centrally rooted tree divided into two main groups (Figure 14). One group of sequences aligned most closely to provirus found on chr3:21,939,61 and site chrUN:JH,373,247. Of note, these two proviruses originated from the same insertion which arose from a segmental duplication event. A small group of sequences including 9A, 21A, and 27A grouped within the CfERV-Fc1 proviruses. Sequences 31A, 33A,40A, and 48A all had identical sequence as well as sequences 42A, 4A, 5A, and 18A being the largest groups of such. Groups of identical sequences were also found between: 23A and 11A; 34A and 42A; 4A, 5A, and 18A.

The DH82 centrally rooted tree (Figure 15) mimicked that of A72, showing sequences from the cell line were most closely related to proviruses at chr3: 21,939,61 and chrUN:JH,373,247. Interestingly sequence 17 from the DH82 cell line grouped with three polymorphic proviruses including those at site chr13:32,380,542,1, chr5: 10,129,759, and chr26: 359,824,40. Only one of those proviruses that is found in cell line DH82 is provirus at site chr13:32,380,542,1. Two more sequences grouped closely with the cluster including 43DH82 and 39DH82. There were three more groups that included identical sequences indicating amplification of the same insertion, the biggest group was shown to be 31DH82, 28DH82, 9DH82, and 4DH82. Other identical groups included: 35DH82 and 36DH82; 30DH82 and 33DH82.

Similar to cell lines A72 and DH82, D17 centrally rooted tree (Figure 16) showed amplified sequences to be most closely related to the insertions at chr3: 21,939,61 and site chrUN:JH,373,247. One sequence, 8D17 grouped directly with the two insertions. Identical insertions were found including: 1D17, 2D17, 24D17, and 48D17; 35D17,4D17, and 25D17;

23D17 and 50D17; 40D17 and 46D17. Unlike the A72 and DH82 cell lines, clones from D17 had a cluster that grouped closely with the insertion at site chr5:78,331,557. The *pol* sequences from all cell-lines grouped together showed a similar pattern (Figure 17). There were even identical sequences being expressed in all the cell lines. Similar to the individual phylogenetic trees, a large number of expressed sequences show to group with chr3:21,939,61 and site chrUN:JH,373,247.

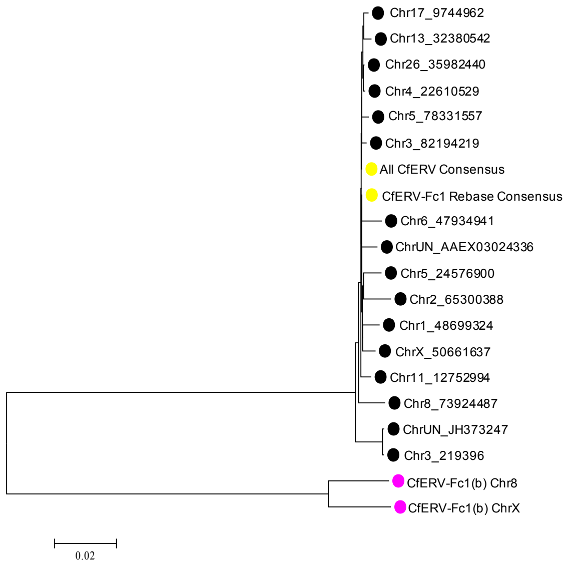


Figure 13. Neighbor Joining tree of full length pol gene in CfERV-Fc1

The alignment of full length pol sequences from annotated CfERV-Fc1(a) insertions along with two CfERV-Fc1(b) sequences were manually aligned using BioEdit v..7.0.9.0. A nieghbor-joining tree was constructed and edited using MEGA7 computed using evolutionary distences determined by Maximum Composite Likelihood method and are in the units of the number of base substitutions per site. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. CfERVFc1CON and Repbase CfERVFc1(a) consesnsus are represented with yellow circles while CfERV-Fc1(b) sequences are shown in pink.

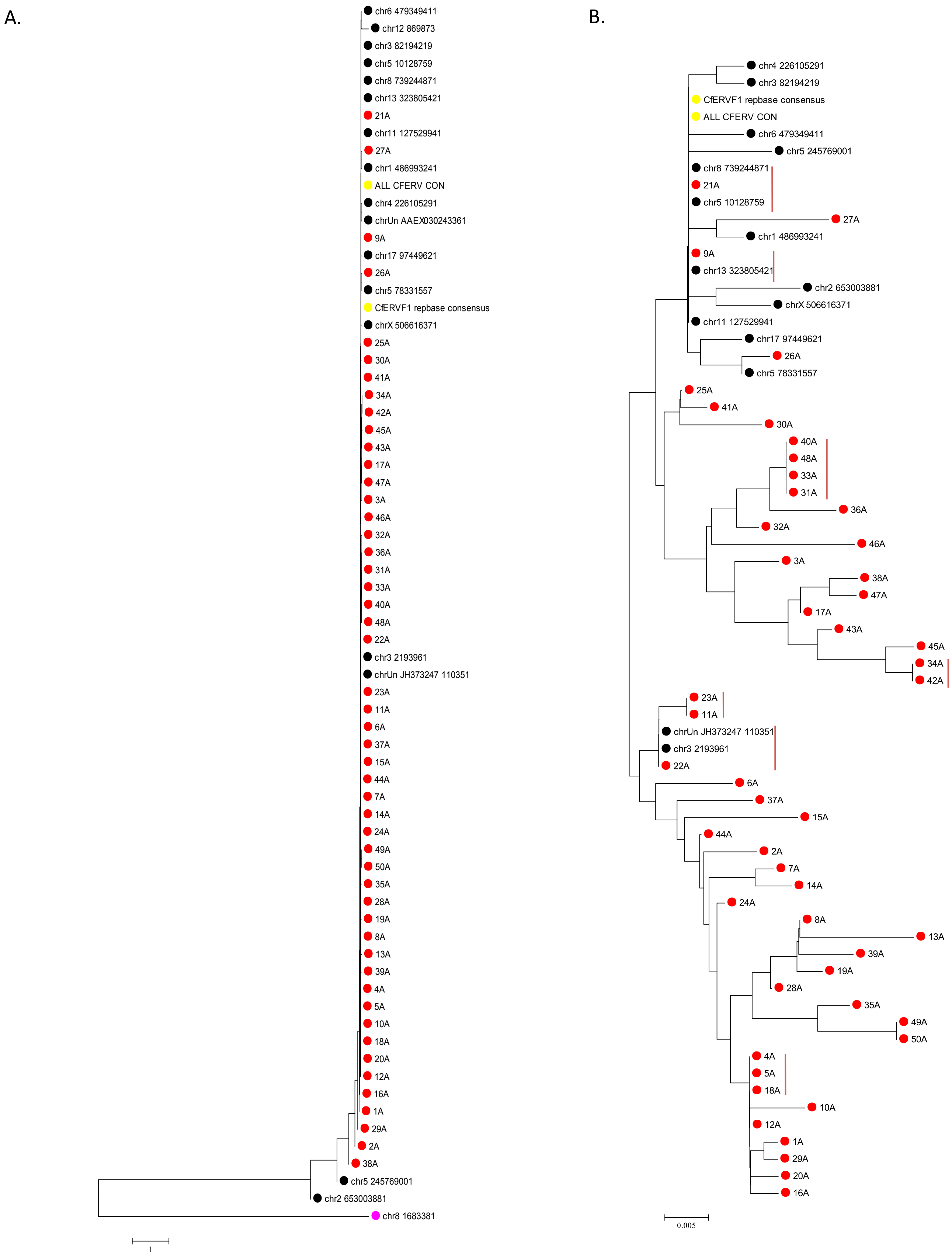

A.
chr6 479349411
chr12 869873
chr3 82194219
chr5 10128759
chr8 739244871
chr13 323805421
21A
chr11 127529941
27A
chr1 486993241
ALL CFERV CON
chr4 226105291
chrUn AAEX030243361
9A
chr17 97449621
26A
chr5 78331557
CfERVF1 repbase consensus
chrX 506616371
25A
30A
41A
34A
42A
45A
43A
17A
47A
3A
46A
32A
36A
31A
33A
40A
48A
22A
chr3 2193961
chrUn JH373247 110351
23A
11A
6A
37A
15A
44A
7A
14A
24A
49A
50A
35A
28A
19A
8A
13A
39A
4A
5A
10A
18A
20A
12A
16A
1A
29A
2A
38A
chr5 245769001
chr2 653003881
chr8 1683381
1
B.
chr4 226105291
chr3 82194219
CfERVF1 repbase consensus
ALL CFERV CON
chr6 479349411
chr5 245769001
chr8 739244871
21A
chr5 10128759
27A
chr1 486993241
9A
chr13 323805421
chr2 653003881
chrX 506616371
chr11 127529941
chr17 97449621
26A
chr5 78331557
25A
41A
30A
40A
48A
33A
31A
36A
32A
46A
3A
38A
47A
17A
43A
45A
34A
42A
23A
11A
chrUn JH373247 110351
chr3 2193961
22A
6A
37A
15A
44A
2A
7A
14A
24A
8A
13A
39A
19A
28A
35A
49A
50A
4A
5A
18A
10A
12A
1A
29A
20A
16A
0.005

Figure 14. Neighbor-Joining tree of A72 *pol* gene

The full length alignment of pol sequences expressed from A72 was manually aligned using BioEdit v..7.0.9.0. A nieghbor-joining tree was constructed and edited using MEGA7 computed using evolutionary distences determined by Maximum Composite Likelihood method and are in the units of the number of base substitutions per site. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. Annotated CfERV-Fc1 insertions are denoted in black, with the consesnus sequences represented in yellow (both repbase consesnus and CfERVFc1CON are represented) and red signifies sequences expressed from A72 cell line. **A.** To exclude the possiblity of sequences coming form another gamma-like lineages, this tree has CfERV-Fc1(b) insertion at chr8:16,833,81, was used as an outgroup and is represented in pink. **B**. CfERV-Fc1(b) insertion at chr8:16,833,81, was excluded leading to a center rooted tree. Identical sequences are identigied with solid red lines.

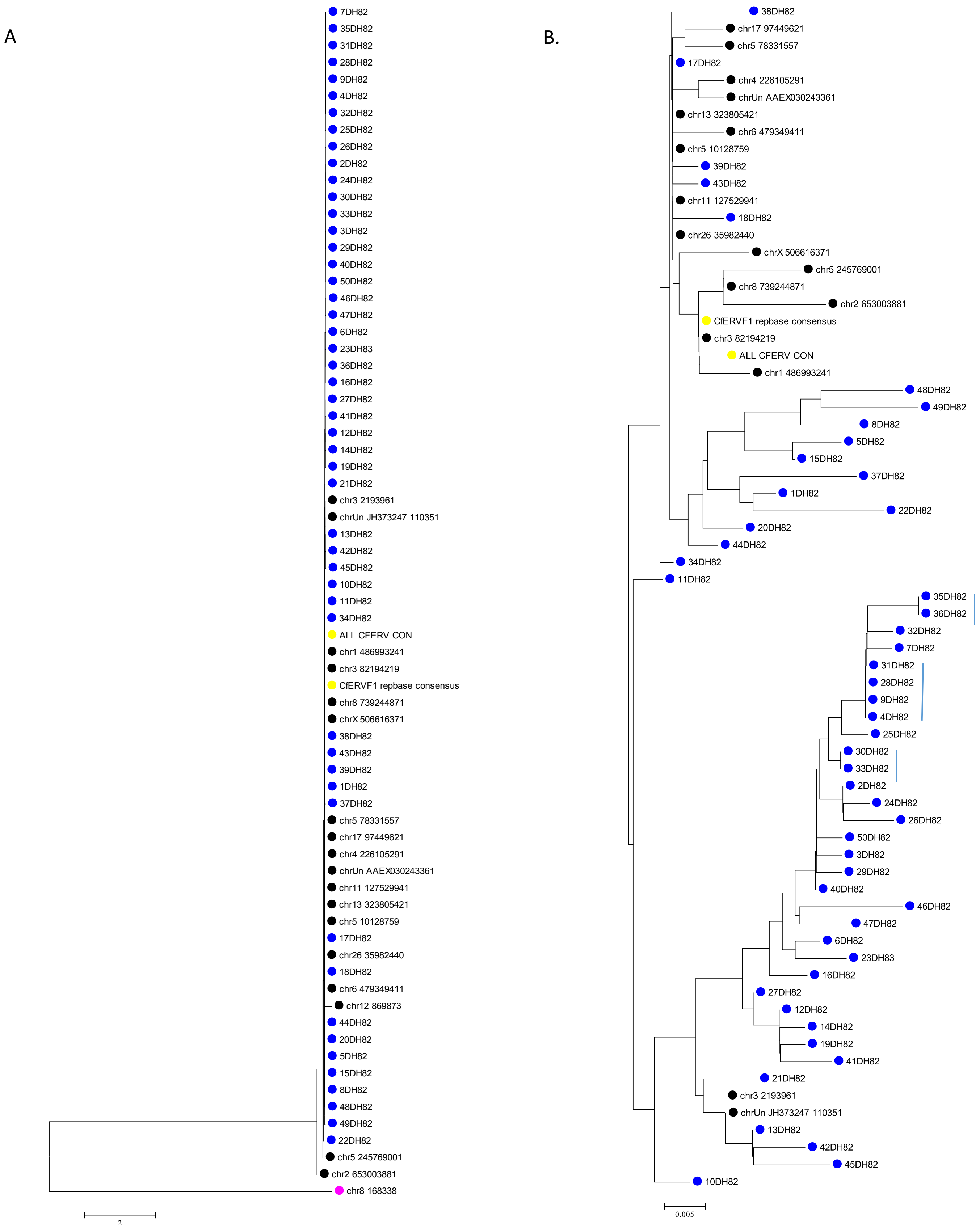

A
7DH82
35DH82
31DH82
28DH82
9DH82
4DH82
32DH82
25DH82
26DH82
2DH82
24DH82
30DH82
33DH82
3DH82
29DH82
40DH82
50DH82
46DH82
47DH82
6DH82
23DH83
36DH82
16DH82
27DH82
41DH82
12DH82
14DH82
19DH82
21DH82
chr3 2193961
chrUn JH373247 110351
13DH82
42DH82
45DH82
10DH82
11DH82
34DH82
ALL CFERV CON
chr1 486993241
chr3 82194219
CfERVF1 repbase consensus
chr8 739244871
chrX 506616371
38DH82
43DH82
39DH82
1DH82
37DH82
chr5 78331557
chr17 97449621
chr4 226105291
chrUn AAEX030243361
chr11 127529941
chr13 323805421
chr5 10128759
17DH82
chr26 35982440
18DH82
chr6 479349411
chr12 869873
44DH82
20DH82
5DH82
15DH82
8DH82
48DH82
49DH82
22DH82
chr5 245769001
chr2 653003881
chr8 168338
2
B.
38DH82
chr17 97449621
chr5 78331557
17DH82
chr4 226105291
chrUn AAEX030243361
chr13 323805421
chr6 479349411
chr5 10128759
39DH82
43DH82
chr11 127529941
18DH82
chr26 35982440
chrX 506616371
chr5 245769001
chr8 739244871
chr2 653003881
CfERVF1 repbase consensus
chr3 82194219
ALL CFERV CON
chr1 486993241
48DH82
49DH82
8DH82
5DH82
15DH82
37DH82
1DH82
22DH82
20DH82
44DH82
34DH82
11DH82
35DH82
36DH82
32DH82
7DH82
31DH82
28DH82
9DH82
4DH82
25DH82
30DH82
33DH82
2DH82
24DH82
26DH82
50DH82
3DH82
29DH82
40DH82
46DH82
47DH82
6DH82
23DH83
16DH82
27DH82
12DH82
14DH82
19DH82
41DH82
21DH82
chr3 2193961
chrUn JH373247 110351
13DH82
42DH82
45DH82
10DH82
0.005

Figure 15. Neighbor-Joining tree of DH82 *pol* gene

The full length alignment of pol sequences expressed from DH82 was manually aligned using BioEdit v..7.0.9.0. A nieghbor-joining tree was constructed and edited using MEGA7 computed using evolutionary distences determined by Maximum Composite Likelihood method and are in the units of the number of base substitutions per site. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. Annotated CfERV-Fc1 insertions are denoted in black, with the consesnus sequences represented in yellow (both repbase consesnus and CfERVFc1CON are represented) and blue signifies sequences expressed from DH82 cell line. **A.** To exclude the possiblity of sequences coming form another gamma-like lineages, this tree has CfERV-Fc1(b) insertion at chr8:16,833,81, was used as an outgroup and is represented in pink. **B**. CfERV-Fc1(b) insertion at chr8:16,833,81, was excluded in order to better examine phylogenic relationships amongst the CfERV-Fc1(a) insertions leading to a center rooted tree. Identical sequences are identigied with solid blue lines.

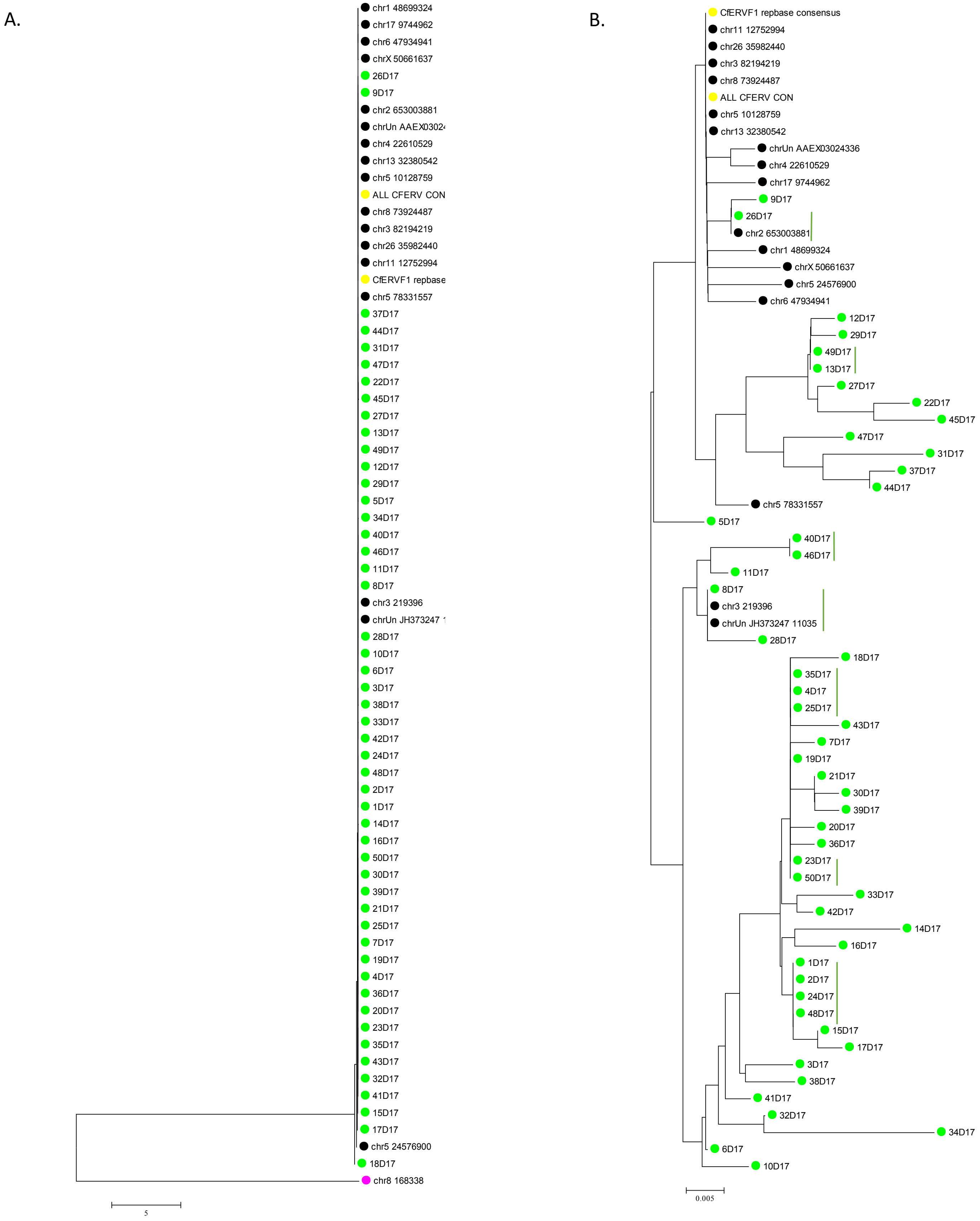

A.
chr1 48699324
chr17 9744962
chr6 47934941
chrX 50661637
26D17
9D17
chr2 653003881
chrUn AAEX0302
chr4 22610529
chr13 32380542
chr5 10128759
ALL CFERV CON
chr8 73924487
chr3 82194219
chr26 35982440
chr11 12752994
CfERVF1 repbase
chr5 78331557
37D17
44D17
31D17
47D17
22D17
45D17
27D17
13D17
49D17
12D17
29D17
5D17
34D17
40D17
46D17
11D17
8D17
chr3 219396
chrUn JH373247
28D17
10D17
6D17
3D17
38D17
33D17
42D17
24D17
48D17
2D17
1D17
14D17
16D17
50D17
30D17
39D17
21D17
25D17
7D17
19D17
4D17
36D17
20D17
23D17
35D17
43D17
32D17
41D17
15D17
17D17
chr5 24576900
18D17
chr8 168338
5
B.
CfERVF1 repbase consensus
chr11 12752994
chr26 35982440
chr3 82194219
chr8 73924487
ALL CFERV CON
chr5 10128759
chr13 32380542
chrUn AAEX03024336
chr4 22610529
chr17 9744962
9D17
26D17
chr2 653003881
chr1 48699324
chrX 50661637
chr5 24576900
chr6 47934941
12D17
29D17
49D17
13D17
27D17
22D17
45D17
47D17
31D17
37D17
44D17
chr5 78331557
5D17
40D17
46D17
11D17
8D17
chr3 219396
chrUn JH373247 11035
28D17
18D17
35D17
4D17
25D17
43D17
7D17
19D17
21D17
30D17
39D17
20D17
36D17
23D17
50D17
33D17
42D17
14D17
16D17
1D17
2D17
24D17
48D17
15D17
17D17
3D17
38D17
41D17
32D17
34D17
6D17
10D17
0.005

Figure 16. Neighbor-Joining tree of D17 *pol* gene

The full length alignment of pol sequences expressed from D17 was manually aligned using BioEdit v..7.0.9.0. A nieghbor-joining tree was constructed and edited using MEGA7 computed using evolutionary distences determined by Maximum Composite Likelihood method and are in the units of the number of base substitutions per site. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. Annotated CfERV-Fc1 insertions are denoted in black, with the consesnus sequences represented in yellow (both repbase consesnus and CfERVFc1CON are represented) and green signifies sequences expressed from D17 cell line. **A.** To exclude the possiblity of sequences coming form another gamma-like lineages, this tree has CfERV-Fc1(b) insertion at chr8:16,833,81, was used as an outgroup and is represented in pink. **B**. CfERV-Fc1(b) insertion at chr8:16,833,81, was excluded leading to a center rooted tree. Identical sequences are identigied with solid green lines.

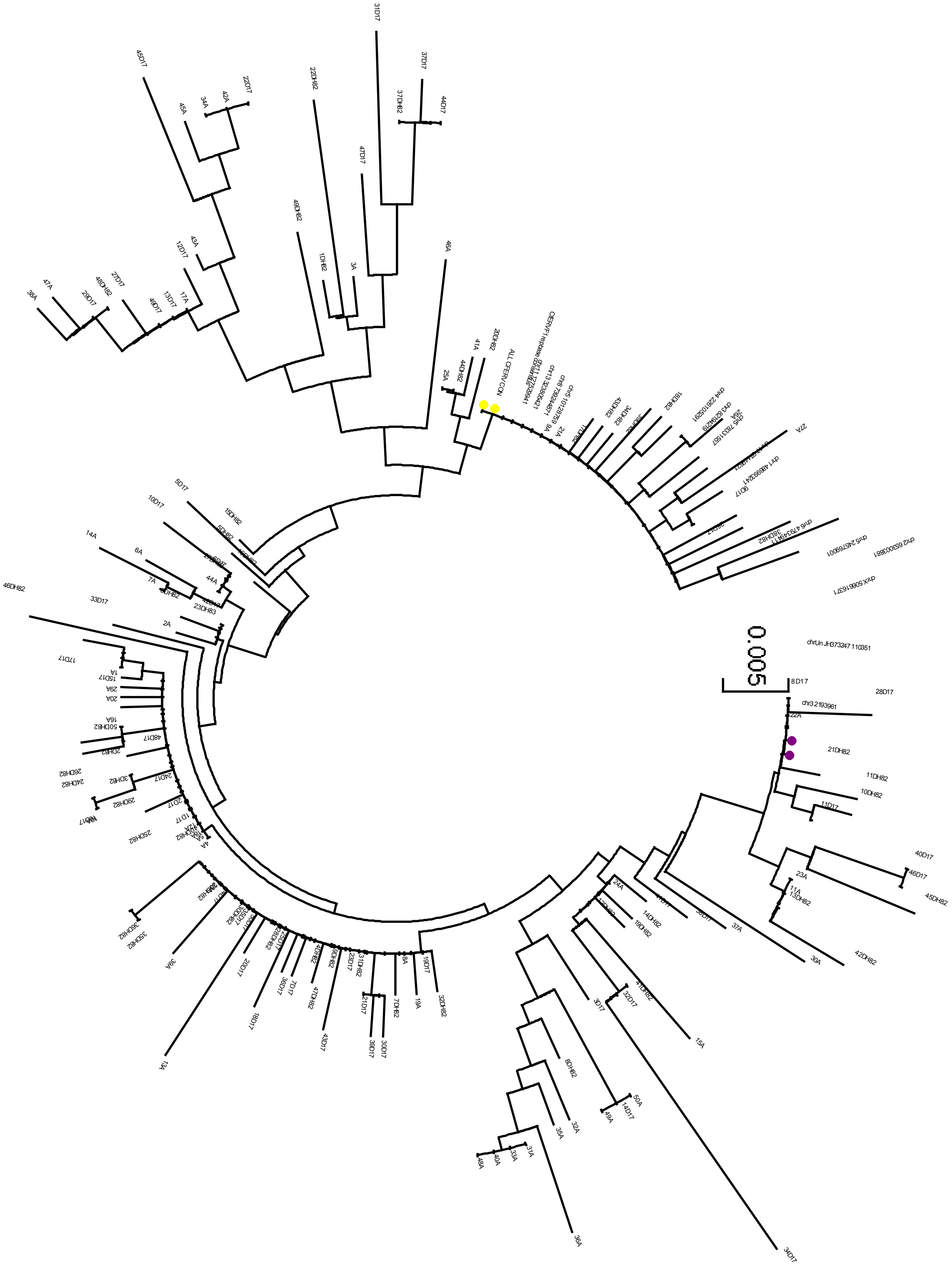
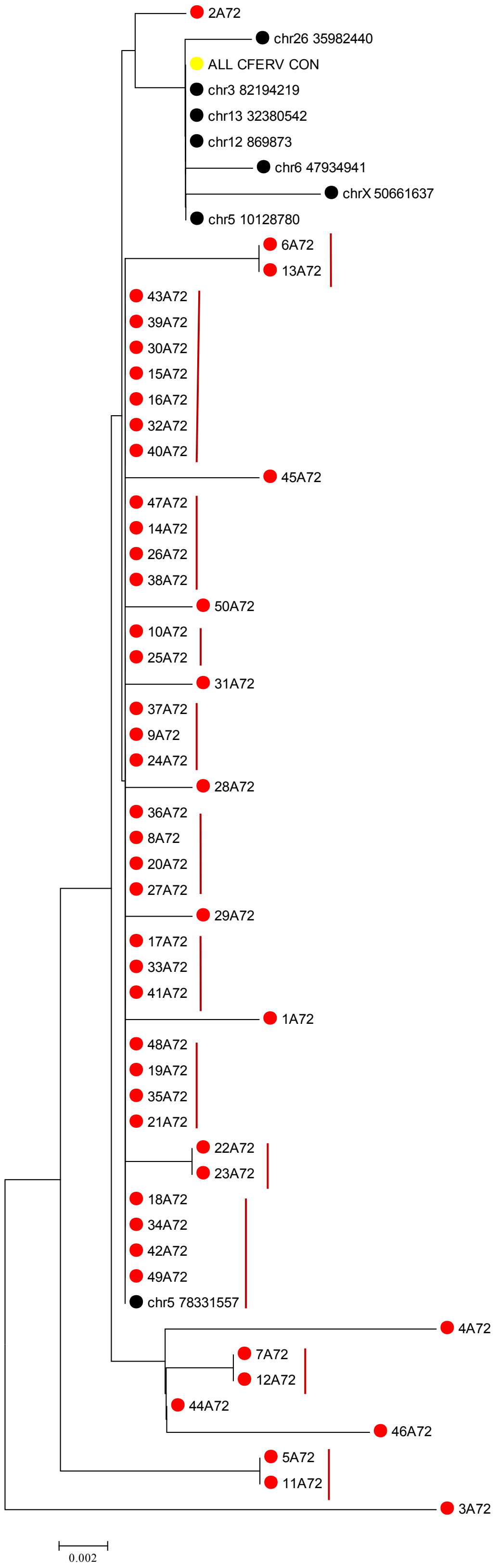

0.005
31D17
37D17
37DH82
44D17
45D17
22D17
42A
34A
45A
22DH82
47D17
49DH82
46A
43A
12D17
1DH82
3A
38A
47A
29D17
48DH82
27D17
49D17
13D17
17A
25A
44DH82
41A
20DH82
ALL CfERV/CON
CfERV1 Integrase
9A
21A
18DH82
chr4 22610291
26A
27A
chr1 48699241
9D17
chr6
chr5 24576001
chr2 65030881
chrX 50661637 1
chrUn JH373247 110351
8D17
28D17
chr3 2193961
22A
21DH82
11DH82
10DH82
11D17
40D17
46D17
45DH82
23A
11A
13DH82
42DH82
30A
37A
38D17
24A
14DH82
19DH82
15A
3D17
32D17
34D17
8DH82
32A
35A
50A
14D17
49A
31A
33A
40A
48A
36A
19D17
32DH82
19A
8A
7DH82
31DH82
30D17
39D17
21D17
43D17
47DH82
7D17
36D17
20D17
18D17
39A
13A
35DH82
36DH82
4A
25DH82
1D17
2D17
29DH82
24D17
3DH82
24DH82
26DH82
2DH82
48D17
50DH82
16A
20A
29A
15D17
1A
17D17
46DH82
33D17
2A
23DH83
7A
6A
44A
14A
10D17
5D17
15DH82
5DH82

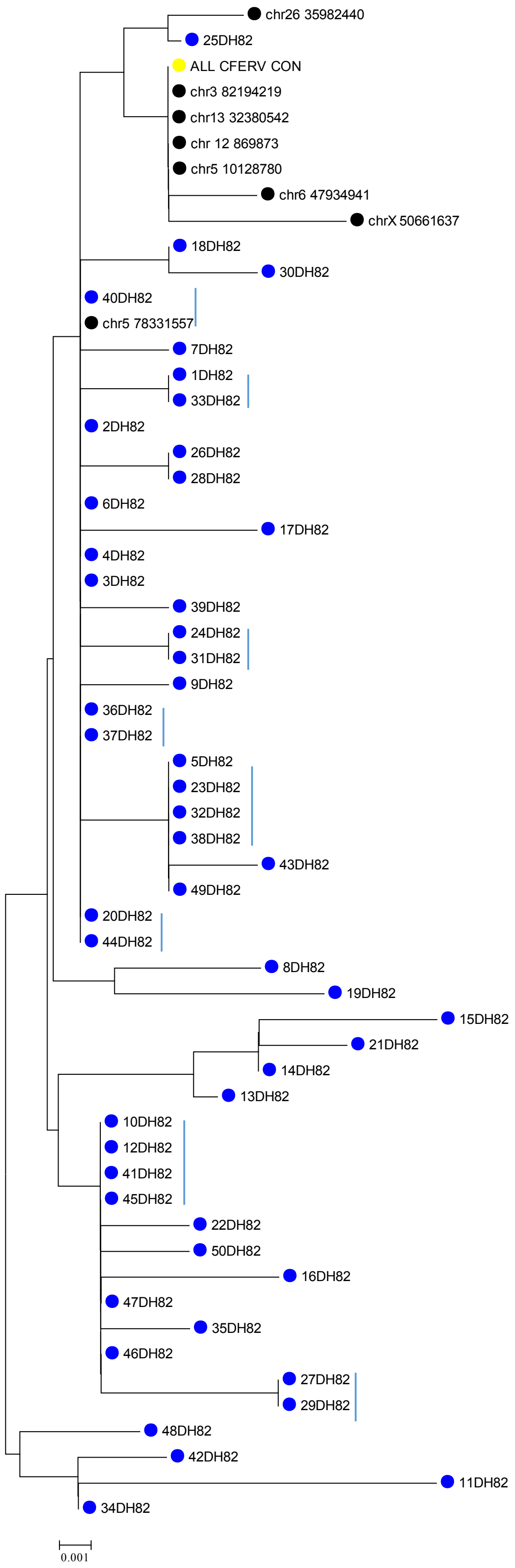
Figure 17. Neighbor Joining tree containing the *pol* sequences expressed in all the cell lines The full length alignment of env sequences expressed from A72, DH82, and D17 was manually aligned using BioEdit v..7.0.9.0. A nieghbor-joining tree was constructed and edited using MEGA7 computed using evolutionary distances determined by Maximum Composite Likelihood method and are in the units of the number of base substitutions per site. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. CfERVFc1CON is highlighted by the yellow circle and purple circles are sequences from proviral insetion chr3: 21,939,61 and chrUN: JH,373,247.

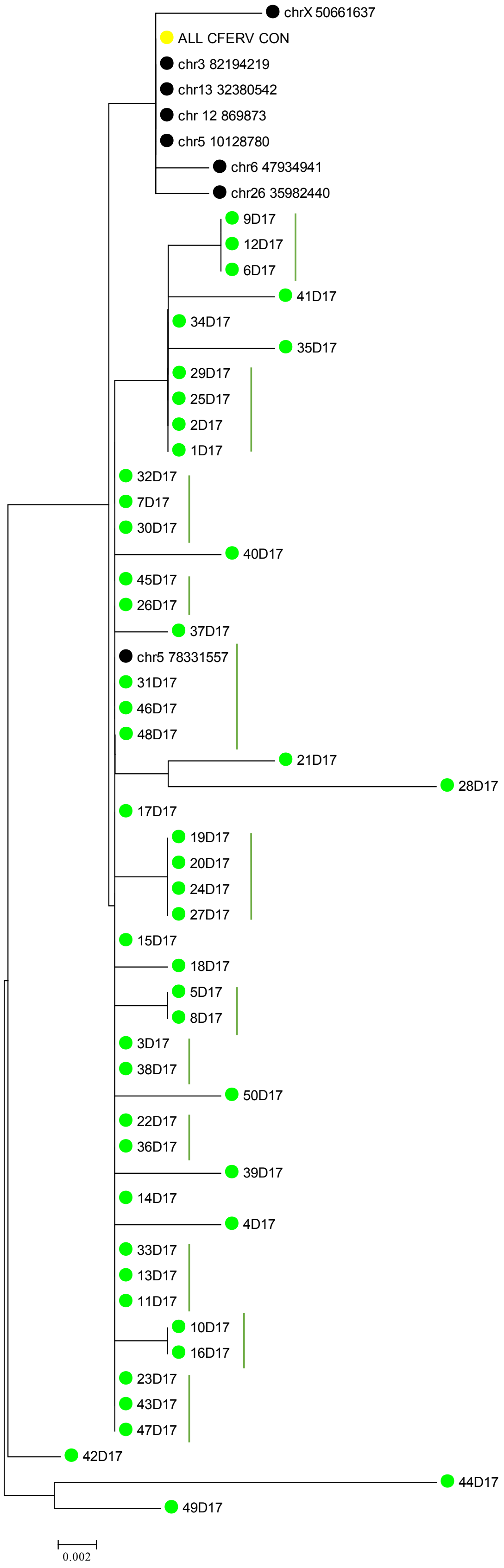
Env gene

The *env* sequences were amplified and cloned from the 3' end within the common 1073 bp deletion; due to this, the Repbase Consensus of CfERV-Fc1 is not represented the neighbor joining trees, since the common deletion is present. Proviruses with the deletion therefore were not amplified, allowing us to only compare sequences from a full length *env* product. In each cell line, the *env* sequences from the annotated proviruses clustered together at the top separate from sequences retrieved from the cell lines with the exception of the *env* sequences from site chr5:78,331,557. In A72 cell line (Figure 18), 31 expressed sequences were identical to the provirus found on chr5:78,331,557, representing the majority of the phylogenetic tree (8A72, 9A72,10A72, 14A72, 15A72, 16A72, 17A72, 18A72, 19A72, 20A72, 21A72, 24A72, 25A72, 26A72, 27A72, 30A72, 32A72, 33A72, 34A72, 35A72, 36A72, 38A72, 39A72, 40A72, 41A72, 42A72, 43A72, 47A72, 48A72, 48A72). Other identical sequences occurred with sequences 6A72 and 13A72; 22A72 and 23A72; 7A72 and 12A72; as well as 5A72 and 11A72.

The DH82 cell-line (Figure 19) had nine expressed sequences identical to chr5:78331557 (2DH82, 3DH82, 4DH82, 6DH82, 20DH82, 36DH82, 37DH82, 40DH82, and 44DH82). One sequence grouped closely with the other annotated provirus chr26:35,982,438, however this insertion was not detected in the DH82 cell line by PCR screens. D17 (Figure 20) also had a large group that clustered with insertion chr5:78,331,557 (3D17, 7D17, 11D17, 13D17, 14D17,15D17, 17D17, 22D17, 23D17, 26D17, 30D17, 31D17, 32D17, 33D17, 36D17, 38D17, 43D17, 45D17, 46D17, 47D17 and 48D17). When all the sequences from the cell lines are mapped together (Figure 21), the overall pattern of clustering with chr5:78,331,557 is mimicked making up the majority of the tree. As seen in the *pol* tree, the cell lines expressed identical sequences of *env*.


2A72
chr26 35982440
ALL CFERV CON
chr3 82194219
chr13 32380542
chr12 869873
chr6 47934941
chrX 50661637
chr5 10128780
6A72
13A72
43A72
39A72
30A72
15A72
16A72
32A72
40A72
45A72
47A72
14A72
26A72
38A72
50A72
10A72
25A72
31A72
37A72
9A72
24A72
28A72
36A72
8A72
20A72
27A72
29A72
17A72
33A72
41A72
1A72
48A72
19A72
35A72
21A72
22A72
23A72
18A72
34A72
42A72
49A72
chr5 78331557
4A72
7A72
12A72
44A72
46A72
5A72
11A72
3A72
0.002


chr26 35982440
25DH82
ALL CFERV CON
chr3 82194219
chr13 32380542
chr 12 869873
chr5 10128780
chr6 47934941
chrX 50661637
18DH82
30DH82
40DH82
chr5 78331557
7DH82
1DH82
33DH82
2DH82
26DH82
28DH82
6DH82
17DH82
4DH82
3DH82
39DH82
24DH82
31DH82
9DH82
36DH82
37DH82
5DH82
23DH82
32DH82
38DH82
43DH82
49DH82
20DH82
44DH82
8DH82
19DH82
15DH82
21DH82
14DH82
13DH82
10DH82
12DH82
41DH82
45DH82
22DH82
50DH82
16DH82
47DH82
35DH82
46DH82
27DH82
29DH82
48DH82
42DH82
11DH82
34DH82
0.001


chrX 50661637
ALL CFERV CON
chr3 82194219
chr13 32380542
chr 12 869873
chr5 10128780
chr6 47934941
chr26 35982440
9D17
12D17
6D17
41D17
34D17
35D17
29D17
25D17
2D17
1D17
32D17
7D17
30D17
40D17
45D17
26D17
37D17
chr5 78331557
31D17
46D17
48D17
21D17
28D17
17D17
19D17
20D17
24D17
27D17
15D17
18D17
5D17
8D17
3D17
38D17
50D17
22D17
36D17
39D17
14D17
4D17
33D17
13D17
11D17
10D17
16D17
23D17
43D17
47D17
42D17
44D17
49D17
0.002

Phylogenetic Trees comparing individual cell line's expressed env sequence

The full length alignment of env sequences expressed from A72, DH82, and D17 was manually aligned using BioEdit v..7.0.9.0. A nieghbor-joining tree was constructed and edited using MEGA7 computed using evolutionary distences determined by Maximum Composite Likelihood method and are in the units of the number of base substitutions per site. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree.

Figure 18. Neighbor-Joining tree of A72 env gene

Annotated CfERV-Fc1 insertions are denoted in black, with the CfERVFc1CON is represented in yellow and red signifies sequences expressed from A72 cell line. Red solid lines highlight identical sequences.

Figure 19. Neighbor-Joining tree of DH82 *env* gene

Annotated CfERV-Fc1 insertions are denoted in black, with CfERVFc1CON represented in yellow and blue signifies sequences expressed from DH82 cell line. Blue solid lines highlight identical sequences.

Figure 20. Neighbor-Joining tree of D17 *env* gene

Annotated CfERV-Fc1 insertions are denoted in black, with CfERVFc1CON represented in yellow and green signifies sequences expressed from D17cell line. Green solid lines highlight identical sequences.

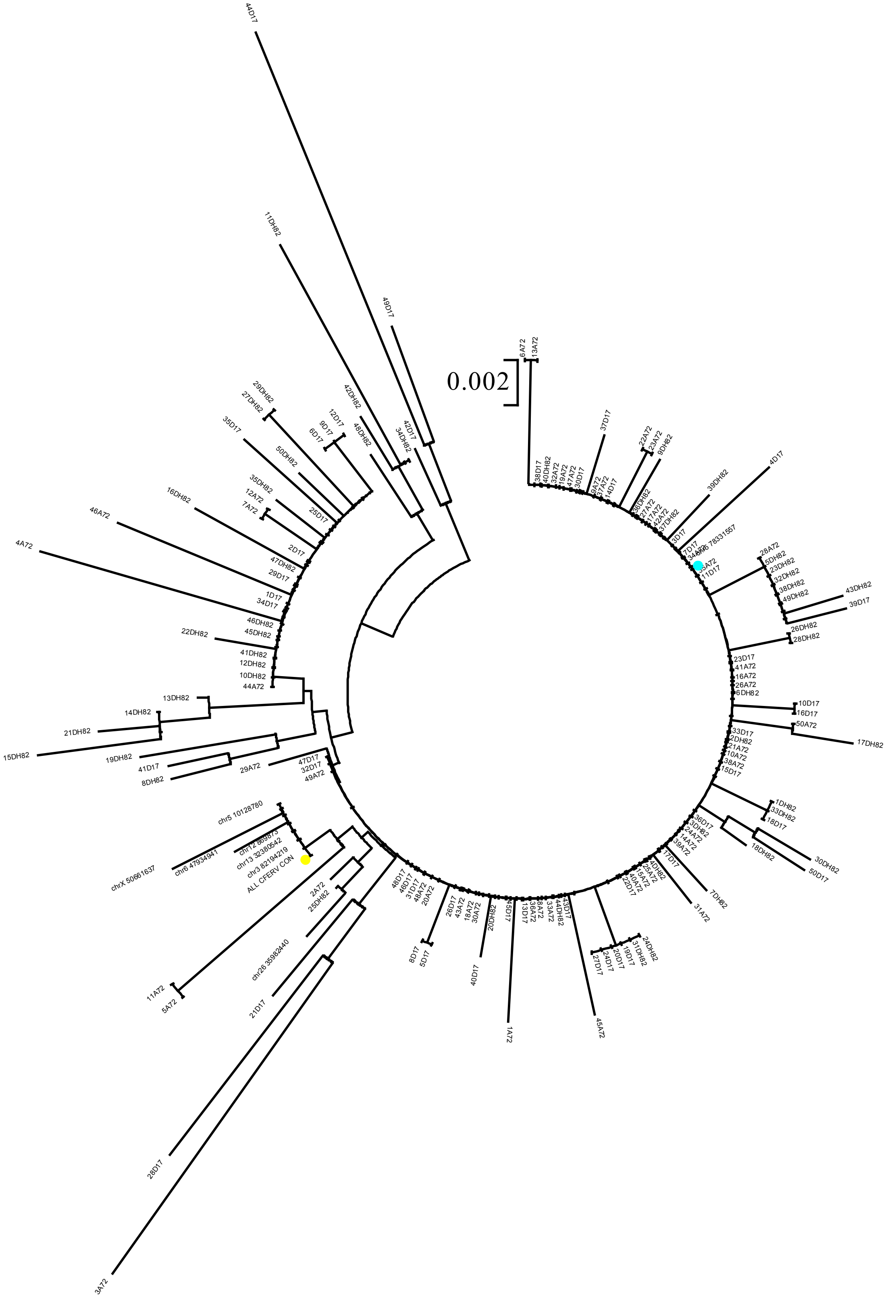
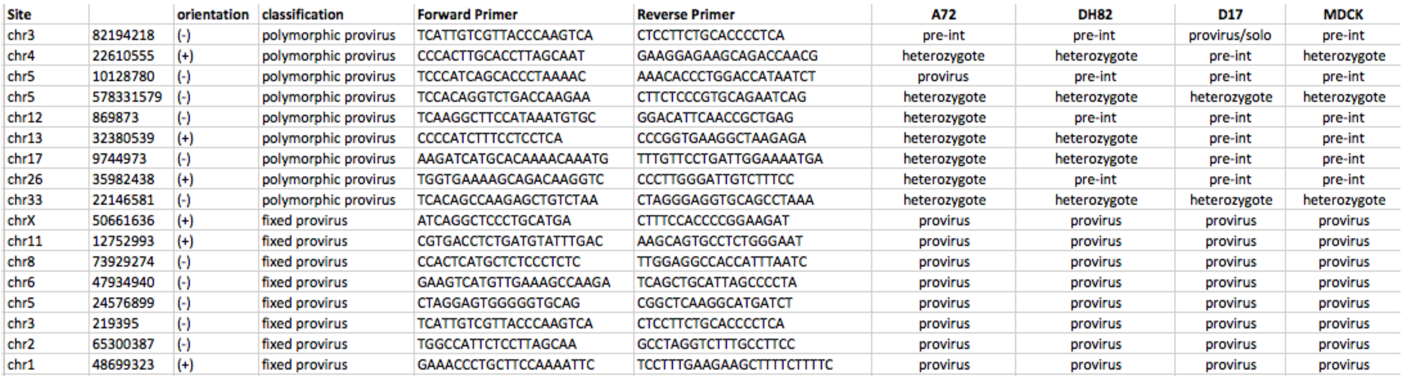

0.002
6A72
13A72
38D17
40DH82
32A72
19A72
47A72
30D17
37D17
9A72
37A72
14D17
22A72
23A72
9DH82
36DH82
27A72
17A72
42A72
37DH82
39DH82
3D17
4D17
7D17
34A72
chr6 78331557
35A72
11D17
28A72
5DH82
23DH82
32DH82
38DH82
49DH82
43DH82
39D17
26DH82
28DH82
23D17
41A72
16A72
26A72
6DH82
10D17
16D17
50A72
17DH82
33D17
2DH82
21A72
10A72
38A72
15D17
1DH82
33DH82
18D17
30DH82
50D17
18DH82
36D17
3DH82
24A72
14A72
39A72
17D17
7DH82
31A72
4DH82
26A72
15A72
40A72
22D17
24DH82
31DH82
19D17
20D17
24D17
27D17
45A72
43D17
44DH82
33A72
8A72
36A72
13D17
45D17
1A72
20DH82
40D17
26D17
43A72
18A72
30A72
48D17
46D17
31D17
48A72
20A72
8D17
5D17
2A72
25DH82
chr26 35982440
21D17
28D17
3A72
11A72
5A72
chr5 10128780
chr12 869673
chr13 32380542
chr3 82194219
ALL CFERV CON
chr6 47934941
chrX 50661637
47D17
32D17
49A72
29A72
41D17
8DH82
19DH82
13DH82
14DH82
21DH82
15DH82
41DH82
12DH82
10DH82
44A72
22DH82
46DH82
45DH82
1D17
34D17
4A72
47DH82
29D17
2D17
46A72
18DH82
35DH82
12A72
7A72
25D17
35D17
50DH82
29DH82
27DH82
12D17
9D17
6D17
42DH82
48DH82
42D17
34DH82
49D17
11DH82
44D17

Figure 21. Neighbor Joining tree containing the *env* sequences expressed in all the cell lines The full length alignment of env sequences expressed from A72, DH82, and D17 was manually aligned using BioEdit v..7.0.9.0. A nieghbor-joining tree was constructed and edited using MEGA7 computed using evolutionary distences determined by Maximum Composite Likelihood method and are in the units of the number of base substitutions per site. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. CfERVFc1CON is highlighted by the yellow circle and blue circle signifies sequences from proviral insetion chr5:78331583.

Distribution of CfERV-Fc1 in Cell Lines

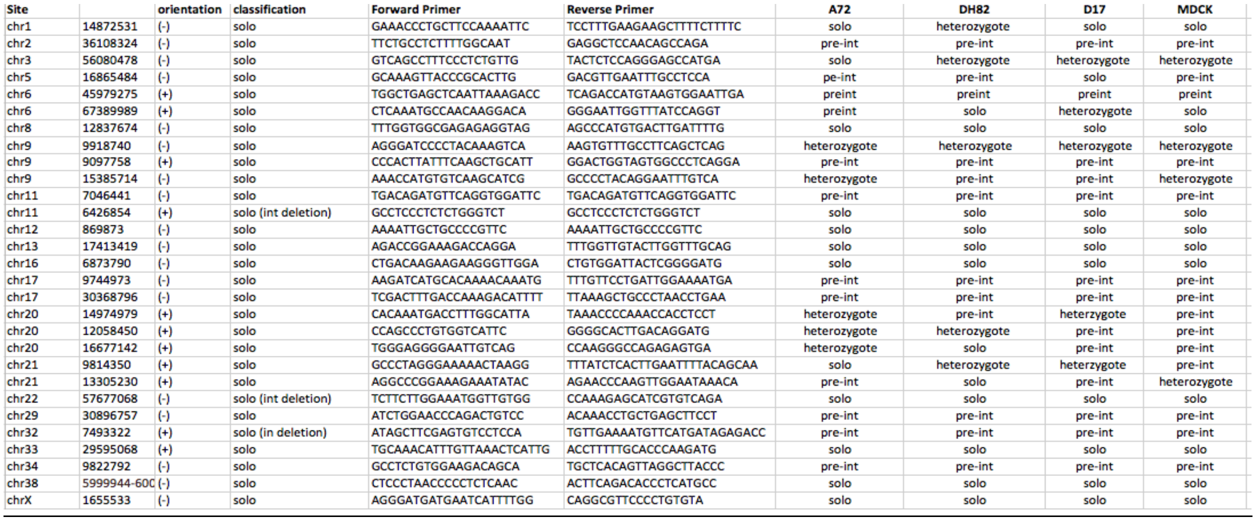
We determined the presence of all fixed insertions in the individual cell lines through PCR genotype screening. A72 showed to have the most polymorphic insertions including sites: chr4:22,610,555, chr5:10,128,780, chr5:78,331,579, chr12:86,987,3, chr13:32,380,539, chr17:97,449,73, chr26:35,982,438, and chr33:22,146,581. DH82 and A72 are both from golden retrievers, and shared the majority of polymorphic sites, however, DH82 lacked provirus insertions: chr5:10,128,780, chr12:86,987,3, and chr26:35,982,438. D17 and MDCK both had even fewer provirus insertions with three insertions found in each. D17 cell line contained polymorphic proviruses at site chr3:38,219,421,8, chr5:78,331,579 and chr33:22,146,581. The MDCK cell line had insertions at chr4:22,610,555, chr5:78,331,579 and chr33:22,146,581. All polymorphic insertions were heterozygous, containing the allele for a pre integration site, with the exception of chr3:82,194,218 in D17 having both a proviral insertion and a solo-LTR present and A72 being a homozygote for provirus at chr5:10,128,780. All proviruses were genotyped using an internal primer at 2210bp internal to provirus amplifying the 5' of the provirus with the exception of the provirus at chr5:78,331,579. This site was screened using an internal primer at 2500bp used to amplify the 3' viral end because there is a transposable element derived highly repetitive sequence found in the beginning of the provirus within its *gag* gene. The insertion presence/absence of all proviruses can be found in Table 3. Presence of solo LTR in the cancer-derived cell lines are represented in Table 4.

Table 3. Presence of proviral insertions in individual cell line



Site		orientation	classification	Forward Primer	Reverse Primer	A72	DH82	D17	MDCK
chr3	82194218	(-)	polymorphic provirus	TCATTGTCGTTACCCAAGTCA	CTCCTTCTGCACCCCTCA	pre-int	pre-int	provirus/solo	pre-int
chr4	22610555	(+)	polymorphic provirus	CCCACTTGCACCTTAGCAAT	GAAGGAGAAGCAGACCAACG	heterozygote	heterozygote	pre-int	heterozygote
chr5	10128780	(-)	polymorphic provirus	TCCCATCAGCACCCTAAAAC	AAACACCCTGGACCATAATCT	provirus	pre-int	pre-int	pre-int
chr5	578331579	(-)	polymorphic provirus	TCCACAGGTCTGACCAAGAA	CTTCTCCCGTGCAGAATCAG	heterozygote	heterozygote	heterozygote	heterozygote
chr12	869873	(-)	polymorphic provirus	TCAAGGCTTCCATAAATGTGC	GGACATTCAACCGCTGAG	heterozygote	pre-int	pre-int	pre-int
chr13	32380539	(+)	polymorphic provirus	CCCCATCTTTCCTCCTCA	CCCGGTGAAGGCTAAGAGA	heterozygote	heterozygote	pre-int	pre-int
chr17	9744973	(-)	polymorphic provirus	AAGATCATGCACAAAACAAATG	TTTGTTCCTGATTGGAAAATGA	heterozygote	heterozygote	pre-int	pre-int
chr26	35982438	(+)	polymorphic provirus	TGGTGAAAAGCAGACAAGGTC	CCCTTGGGATTGTCTTTCC	heterozygote	pre-int	pre-int	pre-int
chr33	22146581	(-)	polymorphic provirus	TCACAGCCAAGAGCTGTCTAA	CTAGGGAGGTGCAGCCTAAA	heterozygote	heterozygote	heterozygote	heterozygote
chrX	50661636	(+)	fixed provirus	ATCAGGCTCCCTGCATGA	CTTTCCACCCCGGAAGAT	provirus	provirus	provirus	provirus
chr11	12752993	(+)	fixed provirus	CGTGACCTCTGATGTATTTGAC	AAGCAGTGCCTCTGGGAAT	provirus	provirus	provirus	provirus
chr8	73929274	(-)	fixed provirus	CCACTCATGCTCTCCCTCTC	TTGGAGGCCACCATTTAATC	provirus	provirus	provirus	provirus
chr6	47934940	(-)	fixed provirus	GAAGTCATGTTGAAAGCCAAGA	TCAGCTGCATTAGCCCCTA	provirus	provirus	provirus	provirus
chr5	24576899	(-)	fixed provirus	CTAGGAGTGGGGGTGCAG	CGGCTCAAGGCATGATCT	provirus	provirus	provirus	provirus
chr3	219395	(-)	fixed provirus	TCATTGTCGTTACCCAAGTCA	CTCCTTCTGCACCCCTCA	provirus	provirus	provirus	provirus
chr2	65300387	(-)	fixed provirus	TGGCCATTCTCCTTAGCAA	GCCTAGGTCTTTGCCTTCC	provirus	provirus	provirus	provirus
chr1	48699323	(+)	fixed provirus	GAAACCCTGCTTCCAAAATTC	TCCTTTGAAGAAGCTTTTCTTTTC	provirus	provirus	provirus	provirus

Table 4. Presence of solo LTR insertions of individual cell lines



Site		orientation	classification	Forward Primer	Reverse Primer	A72	DH82	D17	MDCK
chr1	14872531	(-)	solo	GAAACCCTGCTTCCAAAATTC	TCCTTTGAAGAAGCTTTTCTTTTC	solo	heterozygote	solo	solo
chr2	36108324	(-)	solo	TTCTGCCTCTTTTGGCAAT	GAGGCTCCAACAGCCAGA	pre-int	pre-int	pre-int	pre-int
chr3	56080478	(-)	solo	GTCAGCCTTTCCCTCTGTTG	TACTCTCCAGGGAGCCATGA	solo	heterozygote	heterozygote	heterozygote
chr5	16865484	(-)	solo	GCAAAGTTACCCGCACTTG	GACGTTGAATTTGCCTCCA	pe-int	pre-int	solo	pre-int
chr6	45979275	(+)	solo	TGGCTGAGCTCAATTAAAGACC	TCAGACCATGTAAGTGGAATTGA	preint	preint	preint	preint
chr6	67389989	(+)	solo	CTCAAATGCCAACAAGGACA	GGGAATTGGTTTATCCAGGT	preint	solo	heterozygote	solo
chr8	12837674	(-)	solo	TTTGGTGGCGAGAGAGGTAG	AGCCCATGTGACTTGATTTTG	solo	solo	solo	solo
chr9	9918740	(-)	solo	AGGGATCCCCTACAAAGTCA	AAGTGTTTGCCTTCAGCTCAG	heterozygote	heterozygote	heterozygote	heterozygote
chr9	9097758	(+)	solo	CCCACTTATTTCAAGCTGCATT	GGACTGGTAGTGGCCCTCAGGA	pre-int	pre-int	pre-int	pre-int
chr9	15385714	(-)	solo	AAACCATGTGTCAAGCATCG	GCCCCTACAGGAATTTGTCA	heterozygote	pre-int	pre-int	heterozygote
chr11	7046441	(-)	solo	TGACAGATGTTCAGGTGGATTC	TGACAGATGTTCAGGTGGATTC	pre-int	pre-int	pre-int	pre-int
chr11	6426854	(+)	solo (int deletion)	GCCTCCCTCTCTGGGTCT	GCCTCCCTCTCTGGGTCT	solo	solo	solo	solo
chr12	869873	(-)	solo	AAAATTGCTGCCCCGTTC	AAAATTGCTGCCCCGTTC	solo	solo	solo	solo
chr13	17413419	(-)	solo	AGACCGGAAAGACCAGGA	TTTGGTTGTACTTGGTTTGCAG	solo	solo	solo	solo
chr16	6873790	(-)	solo	CTGACAAGAAGAAGGGTTGGA	CTGTGGATTACTCGGGGATG	solo	solo	solo	solo
chr17	9744973	(-)	solo	AAGATCATGCACAAAACAAATG	TTTGTTCCTGATTGGAAAATGA	pre-int	pre-int	pre-int	pre-int
chr17	30368796	(-)	solo	TCGACTTTGACCAAAGACATTTT	TTAAAGCTGCCCTAACCTGAA	pre-int	pre-int	pre-int	pre-int
chr20	14974979	(+)	solo	CACAAATGACCTTTGGCATTA	TAAACCCCAAACCACCTCCT	heterozygote	pre-int	heterzygote	pre-int
chr20	12058450	(+)	solo	CCAGCCCTGTGGTCATTC	GGGGCACTTGACAGGATG	heterozygote	heterozygote	pre-int	pre-int
chr20	16677142	(+)	solo	TGGGAGGGGAATTGTCAG	CCAAGGGCCAGAGAGTGA	heterozygote	solo	pre-int	pre-int
chr21	9814350	(+)	solo	GCCCTAGGGAAAAACTAAGG	TTTATCTCACTTGAATTTTACAGCAA	solo	heterozygote	heterzygote	pre-int
chr21	13305230	(+)	solo	AGGCCCGGAAAGAAATATAC	AGAACCCAAGTTGGAATAAACA	pre-int	solo	pre-int	heterozygote
chr22	57677068	(-)	solo (int deletion)	TCTTCTTGGAAATGGTTGTGG	CCAAAGAGCATCGTGTCAGA	solo	solo	solo	solo
chr29	30896757	(-)	solo	ATCTGGAACCCAGACTGTCC	ACAAACCTGCTGAGCTTCCT	pre-int	pre-int	pre-int	pre-int
chr32	7493322	(+)	solo (in deletion)	ATAGCTTCGAGTGTCCTCCA	TGTTGAAAATGTTCATGATAGAGACC	pre-int	pre-int	pre-int	pre-int
chr33	29595068	(+)	solo	TGCAAACATTTGTTAAACTCATTG	ACCTTTTTGCACCCAAGATG	solo	solo	solo	solo
chr34	9822792	(-)	solo	GCCTCTGTGGAAGACAGCA	TGCTCACAGTTAGGCTTACCC	pre-int	pre-int	pre-int	pre-int
chr38	5999944-600	(-)	solo	CTCCCTAACCCCCTCTCAAC	ACTTCAGACACCCTCATGCC	solo	solo	solo	solo
chrX	1655533	(-)	solo	AGGGATGATGAATCATTTTGG	CAGGCGTTCCCCTGTGTA	solo	solo	solo	solo

CONCLUSIONS

Modern mammalian genomes are littered with ERVs that are mostly highly mutated and only act as a fossilized remnant of its former infectious self. Although majority of ERVs are fixed amongst a population in decrepit forms and appear to ancient origin, some ERV lineages retain intact loci despite its ancient origin. Some of these intact proviruses even appear to both fixed and unfixed in its host, suggesting that a replicating form of the source insertion may have been active in more recent years. Unfixed insertions with nearly intact retain the potential to be expressed by the host. This expression could be either derived from the provirus genome or the host genome through the exaptation of the insertion's LTR promoting expression of an adjacent gene. Expression of such ERVs has shown to be associated with numerous positive physiological functions as well as to the development of disease [18].

The canine model is understudied model in ERV-host relationship compared to other mammals, due to the previous assumption that ERVs were underrepresented in the genome and the few that were identified appeared to of ancient origin. Not until recently, all insertions appeared to be acquired by its preceding ancestor. There have been previous, but unsubstantiated, reports of reverse transcriptase activity as well as gamma-type C particles in tumor tissues of canines diagnosed with lymphoma [56]. With no known active exogenous retroviruses found in the canine model, and the appearance of only older and deformed insertions, the source of this expression remained puzzling. A number of new insertions in the canine, known as CfERV-Fc1(a), being recently discovered by the Halo lab to have levels of polymorphism exceeding that of the well-researched HERV-K in humans (Figure 22). Insertions of CfERV-Fc1(a) having this high of levels of polymorphism, there is reason to hypothesize there is an association with "younger" insertions and disease as found with HERV-K in humans [37, 39, 47].

The analysis of individual CfERV-Fc1 insertions suggest the lineage had evolved as a group of circulating viruses which infected a canid ancestor. The phylogenies of these individual elements can be explained through spurts of extracellular replication, leading to clusters of what appear to be identical insertions (unpublished data; manuscript in prep). All the viral genes in the CfERV-Fc1 consensus remain intact, suggesting that only a few mutations would be necessary to generate a putatively replication competent virus. Despite having a complete open reading frame present in the consensus, the *gag* regions found in individual insertions were interrupted by mutations leading to an early stop codon. Further analysis of *gag* insertions with nearly complete open reading frames contained the conserved motifs required for *gag* function when inactivated mutations were corrected. *Pol* and *env* open reading frames were found in several of the individual insertions. In fact, four insertions had both intact *pol* and *env* open reading frames. All but two of the individual insertions maintained the predicted motifs in *env* sequence that are necessary for infection. Due to the presence of these possible functional *env* products suggest the spread by infectious particles that lead to the formation of a recombinant as opposed to acquiring it through retrotransposition which does not require a functional *env* gene. These common mutations could be the result of co-packaging of the mutated viral genomes. Along with the common deletion found in *env* and the common mutations leading to premature stops in *gag* and *pol* can support the idea of proliferation in a mechanism predominantly involving complementation.

In this study, we have shown three cell lines that were derived from canine cancer expressed both *pol* and *env* gene found in CfERV-Fc1 while the cell line derived from ‘normal’ canine tissue showed no expression. Majority of the sequences expressed involving *pol* group with insertions chr3:21,939,61 and site chrUN:JH,373,247, leading us to speculate, these insertions become transcriptionally active in some cancers while remaining silenced in normal tissue. These insertions are derived from a single exogenous retrovirus which had a duplicated as a consequence of a segmental duplication event. Due to this segmental duplication, it is possible this insertion has a higher copy number than originally assumed. Without knowing the karyotype, that these cell lines are aneuploidy, with a higher number of chromosome 3 than other chromosomes, accounting for the majority of the sequences being transcribed from this site. This appears to be a fixed insertion amongst canines, that infected the genome 5.83-10.94 million years ago, and is therefore one of the older of the CfERV-Fc1 insertions. Neither of the insertions contain an open reading frame in pol, both interrupted by a premature stop codon. Both also contain the common deletion found in the envelope gene, inhibiting their viral function.

The *env* primers were designed to amplify the portion of the *env* gene where the common deletion is found. This design would amplify the envelope genes where open reading frames are found and are from a presumably full length gene. Directly flanking the deletion appears to be highly conserved across the proviral insertions, with four being identical to the CfERV-Fc1CON including chr3:82,194,218, chr5:10,128,780, chr12:86,987,3 and chr13:32,380,539. Combined with the conservation and the short sequence length, there may not be enough of a phylogenetic signal to create an accurate representation of sequence relatedness. The majority of the sequences expressed involving *env* group with insertion found on chr5:78,331,579. Interestingly, this site could not be completely annotated due to an internal transposable element insertion with highly repetitive sequence in the *gag* gene.

Despite having multiple polymorphic CfERV-Fc1 present, MDCK showed no expression of the *pol* and *env* gene from this particular lineage. Lack of expression suggests there is a mechanism acting to prevent the expression of the proviral sequences that may be deregulated in the other cell lines derived from cancerous tissues, such as methylation. In this regard, demethylation of these CfERV-Fc1 elements would seem a likely cause of expression in these canine cancer derived cell lines. The majority of ERV insertions are thought to be silenced in otherwise ‘normal’ tissues through methylation in mammalian genomes, and a scenario whereby suppression of retroelements expression is the reason for the origin of methylation has even been raised [22]. It is also possible that there is a defect in protein KAP1 (or known as TRIM28), which regulates chromatin structure and works harmoniously with methylation to silence these elements [58]. Alternatively, there could also be an alternate factor that is promoting the expression of these proviruses in A72, DH82 and D17 that is not found in MDCK, or a correlated effect. In order to confirm the lack of expression in MDCK, a quantitative RT PCR will have to be performed.

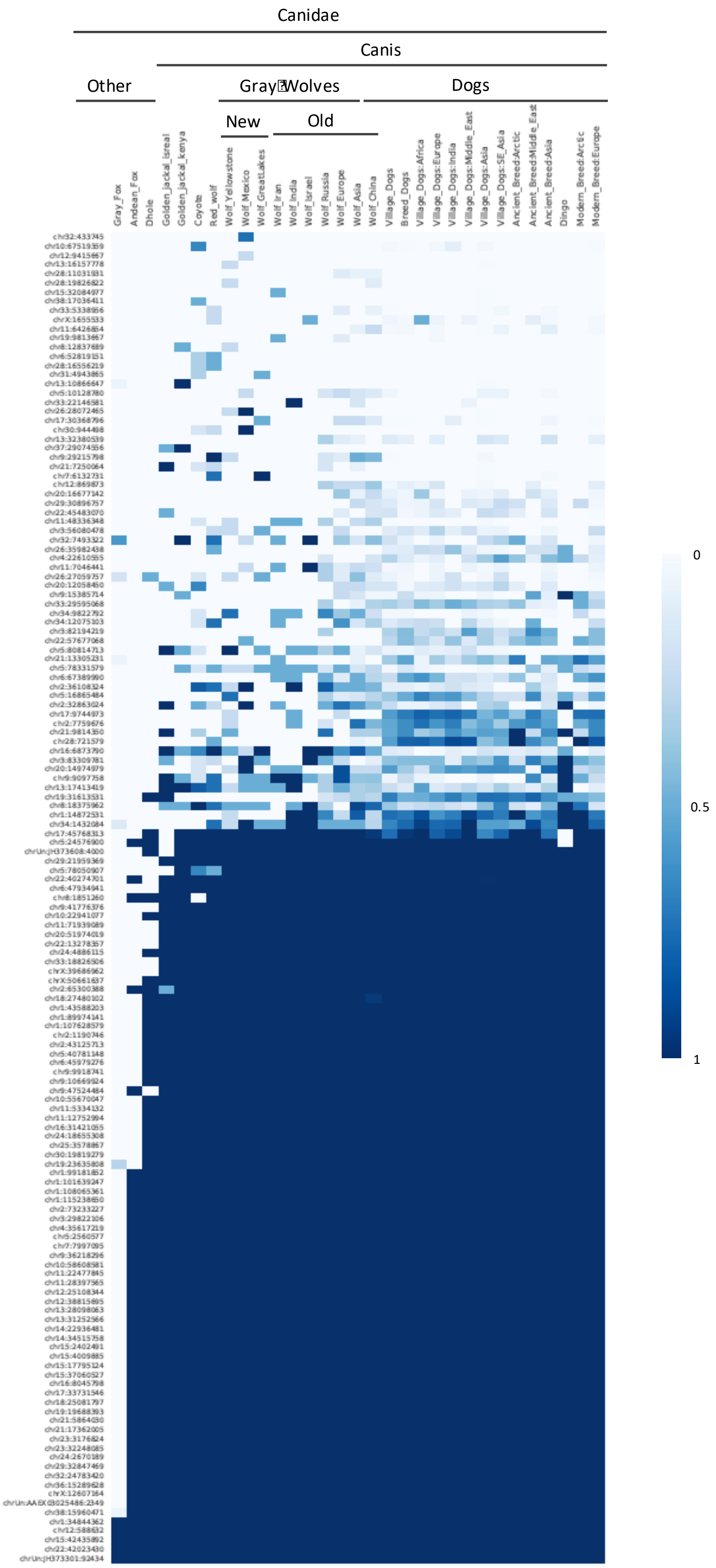

Canidae
Canis
Other
Gray Wolves
Dogs
New
Old
Gray_Fox
Andean_Fox
Dhole
Golden_jackal_isreal
Golden_jackal_kenya
Coyote
Red_wolf
Wolf_Yellowstone
Wolf_Mexico
Wolf_GreatLakes
Wolf_Iran
Wolf_India
Wolf_Israel
Wolf_Russia
Wolf_Europe
Wolf_Asia
Wolf_China
Village_Dogs
Breed_Dogs
Village_Dogs:Africa
Village_Dogs:Europe
Village_Dogs:India
Village_Dogs:Middle_East
Village_Dogs:Asia
Village_Dogs:SE_Asia
Ancient_Breed:Arctic
Ancient_Breed:Middle_East
Ancient_Breed:Asia
Dingo
Modern_Breed:Arctic
Modern_Breed:Europe
chr32:433745
chr10:67519359
chr12:9415667
chr13:16157778
chr28:11031931
chr28:19826822
chr15:32084977
chr38:17036411
chr33:5338956
chrX:1655533
chr11:6426854
chr19:9813667
chr8:12837689
chr6:52819151
chr28:16556219
chr31:4943865
chr13:10866647
chr5:10128780
chr33:22146581
chr26:28072465
chr17:30368796
chr30:944498
chr13:32380539
chr37:29074556
chr9:29215798
chr21:7250064
chr7:6132731
chr12:869873
chr20:16677142
chr29:30896757
chr22:45483070
chr11:48336348
chr3:56080478
chr32:7493322
chr26:35982438
chr4:22610555
chr11:7046441
chr26:27059757
chr20:12058450
chr9:15385714
chr33:29595068
chr34:9822792
chr34:12075103
chr3:82194219
chr22:57677068
chr5:80814713
chr21:13305231
chr5:78331579
chr6:67389990
chr2:36108324
chr5:16865484
chr2:32863024
chr17:9744973
chr2:7759676
chr21:9814350
chr28:721579
chr16:6873790
chr3:83309781
chr20:14974979
chr9:9097758
chr13:17413419
chr19:31613531
chr8:18375962
chr1:14872531
chr34:1432084
chr17:45768313
chr5:24576900
chrUn:JH373608:4000
chr29:21959369
chr5:78050907
chr22:40274701
chr6:47934941
chr8:1851260
chr9:41776376
chr10:22941077
chr11:71939089
chr20:51974019
chr22:13278357
chr24:4886115
chr33:18826506
chrX:39686962
chrX:50661637
chr2:65300388
chr18:27480102
chr1:43588203
chr1:89974141
chr1:107628579
chr2:1190746
chr2:43125713
chr5:40781148
chr6:45979276
chr9:9918741
chr9:10669924
chr9:47524484
chr10:55670047
chr11:5334132
chr11:12752994
chr16:31421055
chr24:18655308
chr25:3578867
chr30:19819279
chr19:23635808
chr1:99181852
chr1:101639247
chr1:108065361
chr1:115238650
chr2:73233227
chr3:29822106
chr4:35617219
chr5:2560577
chr7:7997095
chr9:36218296
chr10:58608581
chr11:22477845
chr11:28397565
chr12:25108344
chr12:38815695
chr13:28098063
chr13:31252566
chr14:22936481
chr14:34515758
chr15:2402491
chr15:4009885
chr15:17795124
chr15:37060527
chr16:8045798
chr17:33731546
chr18:25081797
chr19:19688393
chr21:5864030
chr21:17362005
chr23:3176824
chr23:32248085
chr24:2670189
chr29:32847469
chr32:24783420
chr36:15289628
chrX:12607164
chrUn:AAEX03025486:2349
chr38:15960471
chr1:34844362
chr12:588632
chr15:42435892
chr22:42023430
chrUn:JH373301:92434
0
0.5
1

Figure 22. Heat map of CfERV insertions across dogs and other Canidae

Estimated insertion of allele frequency of unfixed CfERV-Fc1across different Candidae members. Allele frequencies are depicted as a heat map according to the color legend to the right.

Picture from Halo, 2018 (unpublished).

REFERENCES

1. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.

2. Wessler, S.R., *Transposable elements and the evolution of eukaryotic genomes.* Proc Natl Acad Sci U S A, 2006. **103**(47): p. 17600-1.

3. Skipper, K.A., et al., *DNA transposon-based gene vehicles - scenes from an evolutionary drive.* J Biomed Sci, 2013. **20**: p. 92.

4. Sverdlov, E.D., *Retroviruses and primate genome evolution*. Molecular biology intelligence unit. 2005, Georgetown, Tex.: Landes Bioscience. 250 p.

5. Martin, S.L. and F.D. Bushman, *Nucleic acid chaperone activity of the ORF1 protein from the mouse LINE-1 retrotransposon.* Mol Cell Biol, 2001. **21**(2): p. 467-75.

6. Sun, F.J., et al., *Common evolutionary trends for SINE RNA structures.* Trends Genet, 2007. **23**(1): p. 26-33.

7. Tropp, B.E., *Molecular biology : genes to proteins*. 4th ed. 2012, Sudbury, Mass.: Jones & Bartlett Learning. xxxviii, 1097 p.

8. Tropp, B.E., *Biochemistry : concepts and applications*. 1997, Pacific Grove: Brooks/Cole Pub. Co. xxiii, 840 p.

9. Jern, P., *Genomic Variation and Evolution of HERV-H and other Endogenous Retroviruses (ERVs)*, in *Medical Sciences*. 2005, Uppsala Universitet: Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine p. 77.

10. Wildschutte, J.H., *A Study of Polymorphic Endogenous Retroviruses in Humans: Implications in Host Health and Genome Evolution* in *Biomedical Sciences*. 2011, Tufts University p. 239.

11. Coffin, J.M., S.H. Hughes, and H. Varmus, *Retroviruses*. 1997, Plainview, N.Y.: Cold Spring Harbor Laboratory Press. xv, 843 p.

12. Nisole, S. and A. Saib, *Early steps of retrovirus replicative cycle.* Retrovirology, 2004. **1**: p. 9.

13. Guibinga, G.H., et al., *Cell surface heparan sulfate is a receptor for attachment of envelope protein-free retrovirus-like particles and VSV-G pseudotyped MLV-derived retrovirus vectors to target cells.* Mol Ther, 2002. **5**(5 Pt 1): p. 538-46.

14. Stake, M.S., et al., *Nuclear trafficking of retroviral RNAs and Gag proteins during late steps of replication.* Viruses, 2013. **5**(11): p. 2767-95.

15. Basu, V.P., et al., *Strand transfer events during HIV-1 reverse transcription.* Virus Res, 2008. **134**(1-2): p. 19-38.

16. Bush, D.L. and V.M. Vogt, *In Vitro Assembly of Retroviruses.* Annu Rev Virol, 2014. **1**(1): p. 561-80.

17. Swanstrom, R. and J.W. Wills, *Synthesis, Assembly, and Processing of Viral Proteins*, in *Retroviruses*, J.M. Coffin, S.H. Hughes, and H.E. Varmus, Editors. 1997: Cold Spring Harbor (NY).

18. Jern, P. and J.M. Coffin, *Effects of retroviruses on host genome function.* Annu Rev Genet, 2008. **42**: p. 709-32.

19. Sveda, M.M. and R. Soeiro, *Host restriction of Friend leukemia virus: synthesis and integration of the provirus.* Proc Natl Acad Sci U S A, 1976. **73**(7): p. 2356-60.

20. Mura, M., et al., *Late viral interference induced by transdominant Gag of an endogenous retrovirus.* Proc Natl Acad Sci U S A, 2004. **101**(30): p. 11117-22.

21. Perez-Caballero, D., et al., *Tetherin inhibits HIV-1 release by directly tethering virions to cells.* Cell, 2009. **139**(3): p. 499-511.

22. Yoder, J.A., C.P. Walsh, and T.H. Bestor, *Cytosine methylation and the ecology of intragenomic parasites.* Trends Genet, 1997. **13**(8): p. 335-40.

23. Diehl, W.E., et al., *Tracking interspecies transmission and long-term evolution of an ancient retrovirus using the genomes of modern mammals.* Elife, 2016. **5**: p. e12704.

24. Magiorkinis, G., et al., *Env-less endogenous retroviruses are genomic superspreaders.* Proc Natl Acad Sci U S A, 2012. **109**(19): p. 7385-90.

25. Grandgenett, D.P., *Symmetrical recognition of cellular DNA target sequences during retroviral integration.* Proc Natl Acad Sci U S A, 2005. **102**(17): p. 5903-4.

26. Holman, A.G. and J.M. Coffin, *Symmetrical base preferences surrounding HIV-1, avian sarcoma/leukosis virus, and murine leukemia virus integration sites.* Proc Natl Acad Sci U S A, 2005. **102**(17): p. 6103-7.

27. Kassiotis, G., *Endogenous retroviruses and the development of cancer.* J Immunol, 2014. **192**(4): p. 1343-9.

28. Wildschutte, J.H., et al., *Discovery of unfixed endogenous retrovirus insertions in diverse human populations.* Proc Natl Acad Sci U S A, 2016. **113**(16): p. E2326-34.

29. Zhang, Y., J. Shi, and S. Liu, *Recent advances in the study of active endogenous retrovirus envelope glycoproteins in the mammalian placenta.* Virol Sin, 2015. **30**(4): p. 239-48.

30. Lokossou, A.G., C. Toudic, and B. Barbeau, *Implication of human endogenous retrovirus envelope proteins in placental functions.* Viruses, 2014. **6**(11): p. 4609-27.

31. Stoye, J.P., *Endogenous retroviruses: still active after all these years?* Curr Biol, 2001. **11**(22): p. R914-6.

32. Benachenhou, F., et al., *Evolutionary conservation of orthoretroviral long terminal repeats (LTRs) and ab initio detection of single LTRs in genomic data.* PLoS One, 2009. **4**(4): p. e5179.

33. Mager, D.L. and J.P. Stoye, *Mammalian Endogenous Retroviruses.* Microbiol Spectr, 2015. **3**(1): p. MDNA3-0009-2014.

34. Rous, P., *A Sarcoma of the Fowl Transmissible by an Agent Separable from the Tumor Cells.* J Exp Med, 1911. **13**(4): p. 397-411.

35. Weiss, R.A., *The discovery of endogenous retroviruses.* Retrovirology, 2006. **3**: p. 67.

36. Dewannieux, M., et al., *Identification of an infectious progenitor for the multiple-copy HERV-K human endogenous retroelements.* Genome Res, 2006. **16**(12): p. 1548-56.

37. Boller, K., et al., *Human endogenous retrovirus HERV-K113 is capable of producing intact viral particles.* J Gen Virol, 2008. **89**(Pt 2): p. 567-72.

38. Brosius, J., *Genomes were forged by massive bombardments with retroelements and retrosequences.* Genetica, 1999. **107**(1-3): p. 209-38.

39. Boller, K., et al., *Evidence that HERV-K is the endogenous retrovirus sequence that codes for the human teratocarcinoma-derived retrovirus HTDV.* Virology, 1993. **196**(1): p. 349-53.

40. Gimenez, J., et al., *Custom human endogenous retroviruses dedicated microarray identifies self-induced HERV-W family elements reactivated in testicular cancer upon methylation control.* Nucleic Acids Res, 2010. **38**(7): p. 2229-46.

41. Stratton, W.T., *Physician's obligation to cooperate with third-party payors.* Kans Med, 1988. **89**(2): p. 34.

42. Wildschutte, J.H., et al., *The distribution of insertionally polymorphic endogenous retroviruses in breast cancer patients and cancer-free controls.* Retrovirology, 2014. **11**: p. 62.

43. Flockerzi, A., et al., *Expression patterns of transcribed human endogenous retrovirus HERV-K(HML-2) loci in human tissues and the need for a HERV Transcriptome Project.* BMC Genomics, 2008. **9**: p. 354.

44. Cherkasova, E., et al., *Detection of an Immunogenic HERV-E Envelope with Selective Expression in Clear Cell Kidney Cancer.* Cancer Res, 2016. **76**(8): p. 2177-85.

45. Lin, D.Y., et al., *Analysis of the interaction between Zinc finger protein 179 (Znf179) and promyelocytic leukemia zinc finger (Plzf).* J Biomed Sci, 2013. **20**: p. 98.

46. Depil, S., et al., *Expression of a human endogenous retrovirus, HERV-K, in the blood cells of leukemia patients.* Leukemia, 2002. **16**(2): p. 254-9.

47. Bergallo, M., et al., *Expression of the pol gene of human endogenous retroviruses HERV-K and -W in leukemia patients.* Arch Virol, 2017.

48. Turner, G., et al., *Insertional polymorphisms of full-length endogenous retroviruses in humans.* Curr Biol, 2001. **11**(19): p. 1531-5.

49. Mangeney, M., et al., *The full-length envelope of an HERV-H human endogenous retrovirus has immunosuppressive properties.* J Gen Virol, 2001. **82**(Pt 10): p. 2515-8.

50. Ryan, F.P., *Human endogenous retroviruses in health and disease: a symbiotic perspective.* J R Soc Med, 2004. **97**(12): p. 560-5.

51. Barrio, A.M., et al., *The first sequenced carnivore genome shows complex host-endogenous retrovirus relationships.* PLoS One, 2011. **6**(5): p. e19832.

52. Keane, T.M., K. Wong, and D.J. Adams, *RetroSeq: transposable element discovery from next-generation sequencing data.* Bioinformatics, 2013. **29**(3): p. 389-90.

53. Wildschutte, J.H., et al., *Discovery and characterization of Alu repeat sequences via precise local read assembly.* Nucleic Acids Res, 2015. **43**(21): p. 10292-307.

54. Hartley, J.W., et al., *A new class of murine leukemia virus associated with development of spontaneous lymphomas.* Proc Natl Acad Sci U S A, 1977. **74**(2): p. 789-92.

55. Sinha, A. and W.E. Johnson, *Retroviruses of the RDR superinfection interference group: ancient origins and broad host distribution of a promiscuous Env gene.* Curr Opin Virol, 2017. **25**: p. 105-112.

56. Tarlinton, R.E., et al., *Characterisation of a group of endogenous gammaretroviruses in the canine genome.* Vet J, 2013. **196**(1): p. 28-33.

57. Hytonen, M.K. and H. Lohi, *Canine models of human rare disorders.* Rare Dis, 2016. **4**(1): p. e1241362.

58. Rowe, H.M., et al., *KAP1 controls endogenous retroviruses in embryonic stem cells.* Nature, 2010. **463**(7278): p. 237-40.