Swedish University of Agricultural Sciences
Faculty of Veterinary Medicine and Animal Science

Molecular analysis of dog and wolf genomic DNA to explore integration polymorphisms of Canine Endogenous Retroviruses, *CfERV*

MohammadReza Mirzazadeh

Examensarbete / Swedish University of Agricultural Sciences, Department of Animal Breeding and Genetics,

392

Uppsala 2012

Master's Thesis, 30 HEC

Erasmus Mundus Programme – European Master in Animal Breeding and Genetics

Swedish University of Agricultural Sciences
Faculty of Veterinary Medicine and Animal Science
Department of Animal Breeding and Genetics

Molecular analysis of dog and wolf genomic DNA to explore integration polymorphisms of Canine Endogenous Retroviruses, *CfERV*

MohammadReza Mirzazadeh

Supervisors:
Göran Andersson, SLU, Department of Animal Breeding and Genetics
Paatric Jern, UU, Department of Medical Biochemistry and Microbiology
Alvaro Martinez Barrio, UU, Department of Medical Biochemistry and Microbiology
Examiner:
Leif Andersson, SLU, Department of Animal Breeding and Genetics

Credits: 30 HEC
Course title: Degree project in Animal Science
Course code: EX0556
Programme: Erasmus Mundus Programme - European Master in Animal Breeding and Genetics
Level: Advanced, A2E

Place of publication: Uppsala
Year of publication: 2012
Name of series: Examensarbete / Swedish University of Agricultural Sciences, Department of Animal Breeding and Genetics, 392
On-line publication: http://epsilon.slu.se

Key words: Endogenous Retrovirus (ERV), Canine Endogenous Retrovirus (CfERV), germ line, Vertical transmission, Polymorphism, Copy Number Variation (CNV)

1 Table of Contents

2 ABBREVIATIONS **2**

3 ABSTRACT **3**

4 INTRODUCTION **3**

4.1 TRANSPOSABLE ELEMENTS 4

4.2 RETROVIRAL REPLICATION CYCLE 5

4.3 ENDOGENOUS RETROVIRUSES (ERVS) 7

4.4 BIOLOGICAL RELEVANCE OF ERVS 7

4.5 USING DOG MODEL TO STUDY INTEGRATION POLYMORPHISM OF ERVS 8

4.6 CANINE ENDOGENOUS RETROVIRUSES (CFERVS) 9

5 MATERIALS AND METHODS **11**

5.1 PROVIRAL DATA COLLECTION 11

5.2 PRIMER DESIGN 11

5.3 DNA SAMPLES 13

5.4 COPY NUMBER VARIATION (CNV) DATA 14

5.5 BIOINFORMATICS TOOLS 14

6 Results **14**

6.1 INTEGRATION POLYMORPHISM OF CFERVS 14

6.2 CNV ANALYSIS 20

7 DISCUSSION AND FUTURE PROSPECTS **21**

7.1 LACK OF INTEGRATIONAL POLYMORPHISM 21

7.2 COPY NUMBER VARIATION 22

8 ACKNOWLEDGMENTS **22**

9 REFERENCES: **22**

2 ABBREVIATIONS

HERV	human endogenous retrovirus
ERV	endogenous retrovirus
XRV	exogenous retrovirus
MA	matrix protein
CA	capsid protein
NC	nucleocapsid protein
CfERV	*Canis familiaris* endogenous retrovirus
gag	gag gene
pro	protease gene
pol	polymerase gene
env	envelope gene
bp	base pairs (nucleotides)
Kbp	kilo base pairs
LTR	long terminal repeat
R	repeat sequence (in the LTR)
PBS	primer binding site
RT	reverse transcriptase enzyme
IN	integrase enzyme
SU	(envelope) surface unit
TM	(envelope) transmembrane protein
U3	unique 3'-sequence
U5	unique 5'-sequence
tRNA	transfer ribonucleic acid
ss	single stranded
ds	double stranded

3 ABSTRACT

Endogenous retroviruses (ERVs) are found in all examined vertebrate genomes. Different mammals have been reported to contain different amounts of ERVs. For example, in the dog genome 0.15% of sequences are derived from retroviruses. Genome rearrangements driven by retroviral transposition likely have had effects on plasticity of mammalian genomes. During evolution, occasionally, exogenous retrovirus (XRVs) infected germ line cells and the acquired provirus might have been transmitted vertically from generation to generation as a normal Mendelian trait. These rare events of germline infections will result in the generation of ERVs. To gain further insights into the nature of Canine Endogenous Retroviruses (CfERVs) we have performed a PCR-based survey of insertional polymorphism of those CfERVs that were estimated to have integrated recently due to the low degree of divergence of their respective 5' and 3' long terminal repeats (LTRs). The presence of potential integration polymorphism was analysed in genomic DNA prepared from different dog breeds and several wolves by using locus-specific primers for CfERV-chromosomal junctions. We did not find any evidence for integration polymorphism for the CfERV-Fc4 group, which may indicate that integration of this group of CfERVs occurred prior to domestication of dogs from *Canis lupus*. Furthermore, using sequence annotation tools the implication of CfERVs in canine copy-number variation (CNVs) was estimated and we have found evidences for overlap between CfERVs and CNVs.

Keywords: Endogenous Retrovirus (ERV), Canine Endogenous Retrovirus (CfERV), germ line, Vertical transmission, Polymorphism, Copy Number Variation (CNV).

4 INTRODUCTION

Retroviruses are enveloped RNA viruses that contain two molecules of positive-sense single-stranded (ss) RNA ranging approximately 6-11 Kbp. The retroviral RNA is packaged within a capsid structure consisting of *gag*-encoded proteins [1]. They have a unique morphology and means of replication provided by the enzyme RNA polymerase (reverse transcriptase, RT) encoded by the *pol* gene [2]. Retroviruses are classified in seven genera (Table 1) and their gene content varies from simple to complex. The genome structure of simple retroviruses (*e.g*. alpha, gamma) (Figure 1) consists of only four essential genes; *gag*, *pro*, *pol* and *env*, whereas complex retroviruses (*e.g*. delta, epsilon, lenti, and spumaviruses) contain additional genes. The order of the four main coding genes (*i.e.*: *gag*, *pro*, *pol* and *env*) is invariant for all retroviruses and they are flanked by regulatory sequences to control proviral transcription and retroviral RNA processing [reviewed in 1].

Table1. The retrovirus genera [reviewed in 1].

Genus	Type	Representing Virus	Organization
Alpha	Avian type-C	ALV Avian Leukemia Virus	Simple
Beta	Mammalian type-B, D	MMTV Mouse Mammary Tumour Virus	Simple/Complex
Gamma	Type-C	MLV Mouse Leukemia Virus	Simple
Delta	Type-C like	BLV Bovine Leukemia Virus	Complex
Epsilon	Type-C	WDSV Walleye Dermal Sarcoma Virus	Complex
Lenti	-	HIV Human Immunodeficiency Virus	Complex
Spuma	Type-C like	HSRV Human Spuma RetroVirus	Complex

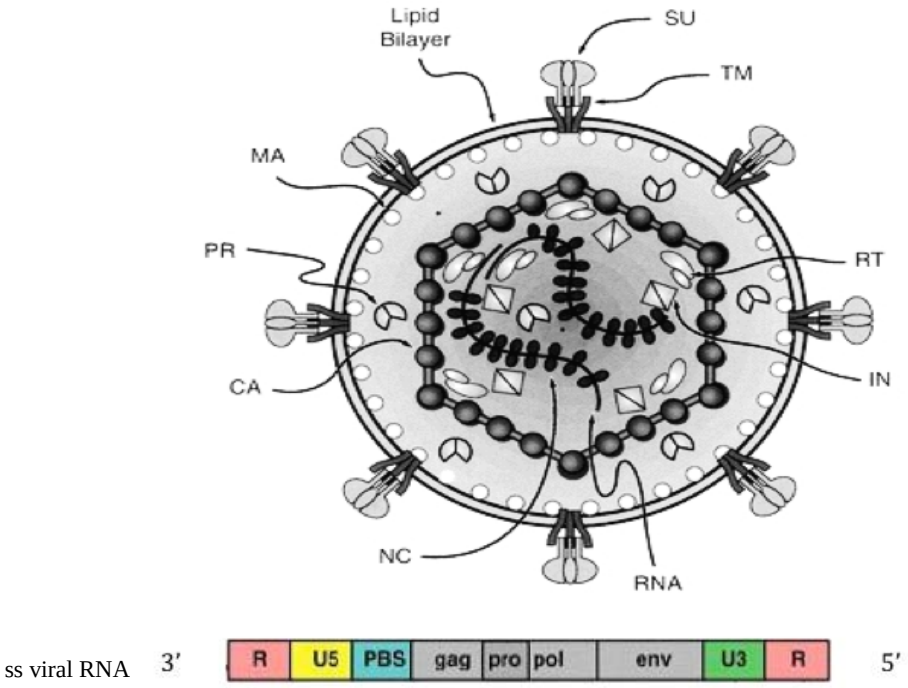

Figure 1. Retrovirus structure and genome organization: R region (repeat sequence), U5 (unique 5'-sequence), PBS (primer-binding site), *gag* (MA, matrix; CA, capsid; NC, nucleocapsid); *pro* (PR, protease); *pol* (RT, reverse transcriptase; IN, integrase); *env* (SU, surface protein; TM, transmembrane protein); U3 (unique 3'-sequence). Picture adapted from : http://www.genetherapyreview.com/education/gene-transfer-vectors/viral-vectors/retrovirus

4.1 Transposable elements

Mobile genetic elements that have the capacity to be mobilized by different mechanisms within the genome where they are present are known as transposable elements (TEs) [3]. TEs are divided in two classes [3]; (*i*) DNA transposons that are mobilized in the genome by a "cut and paste" mechanism, and (*ii*) retrotransposons, which in contrast are mobilized by a "copy and paste"

mechanism using an RNA-intermediate by the enzyme RT into DNA that subsequently may become integrated into another genomic location [4]. The retrotransposons are also further subdivided into:(*i*) non-LTR retrotransposons (*i.e.* long interspersed nuclear elements, LINEs, and short interspersed nuclear elements, SINEs) and (*ii*) LTR retrotransposons (*i.e.* retroviral-like elements) (Figure. 2). LINEs are able to retrotranspose autonomously ("autonomous") but SINEs borrow the enzymes that catalyse transposition (*e.g.* reverse transcriptase) from other elements, usually from LINEs and therefore are "non-autonomous" [4, 5]. The study performed in this thesis, has focused on the LTR retrotransposons.

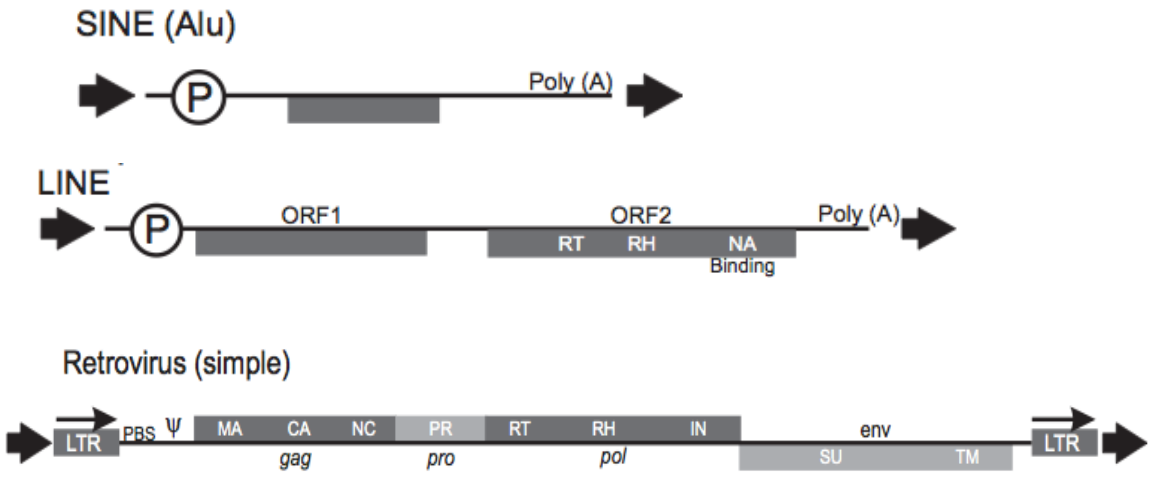

Figure 2. Non-LTR (LINEs and SINEs) and LTR retrotranspososns. [Picture adapted from 1].

4.2 Retroviral replication cycle

The affinity of the exogenous retrovirus (XRV) surface protein (SU) localizes and determines the cellular tropism (*i.e*. which cell type will be infected). Upon attachment of SU to its specific host receptor together with the conformational changes in the transmembrane protein (TM) the fusion of the retrovirus into the cell membrane occurs [reviewed in1].

Once the uncoated virion core is released into the cytoplasm, reverse transcription is started when the host 3' tRNA binds to the primer binding site (PBS) to prime the 5' end of the viral RNA into the minus DNA strand [6]. Classification of human endogenous retroviruses (HERVs) is based on the PBS, which is complementary to the 3' end of the host tRNAs (*e.g.* HERV-E, HERV-H, ASLVs-W), although many ERVs contain alternative tRNA affinities (*e.g.* HERV-H/F and ERV9/HERV-W) to their PBS rendering this classification unsuitable for unambiguous classification of ERVs [7]. After the 5' end of minus DNA strand is copied, the hybrid of DNA-tRNA then attaches to the complementary R region on the 3'end of viral RNA and RT completes the synthesis of DNA minus strand. For the synthesis of positive DNA strand, the poly-purine tract (PPT) is used as a primer to initiate the synthesis (Figure 3)[8]. The two identical long terminal repeats (LTRs) at both ends of the retrovirus are formed during the reverse transcription process. This is one of the differences between retroviral RNA and the proviral DNA where the former contains two unique sequences at the different ends, 5' end (R-U5) and 3' end (U3-R) while after reverse transcription, the integrated provirus comprises identical U3-R-U5 at both the

5’ and 3’ ends [1]. Unlike the DNA polymerase, RT lacks the ability of proofreading during reverse transcription, making it more error prone and thereby creates more genetic variation [9].

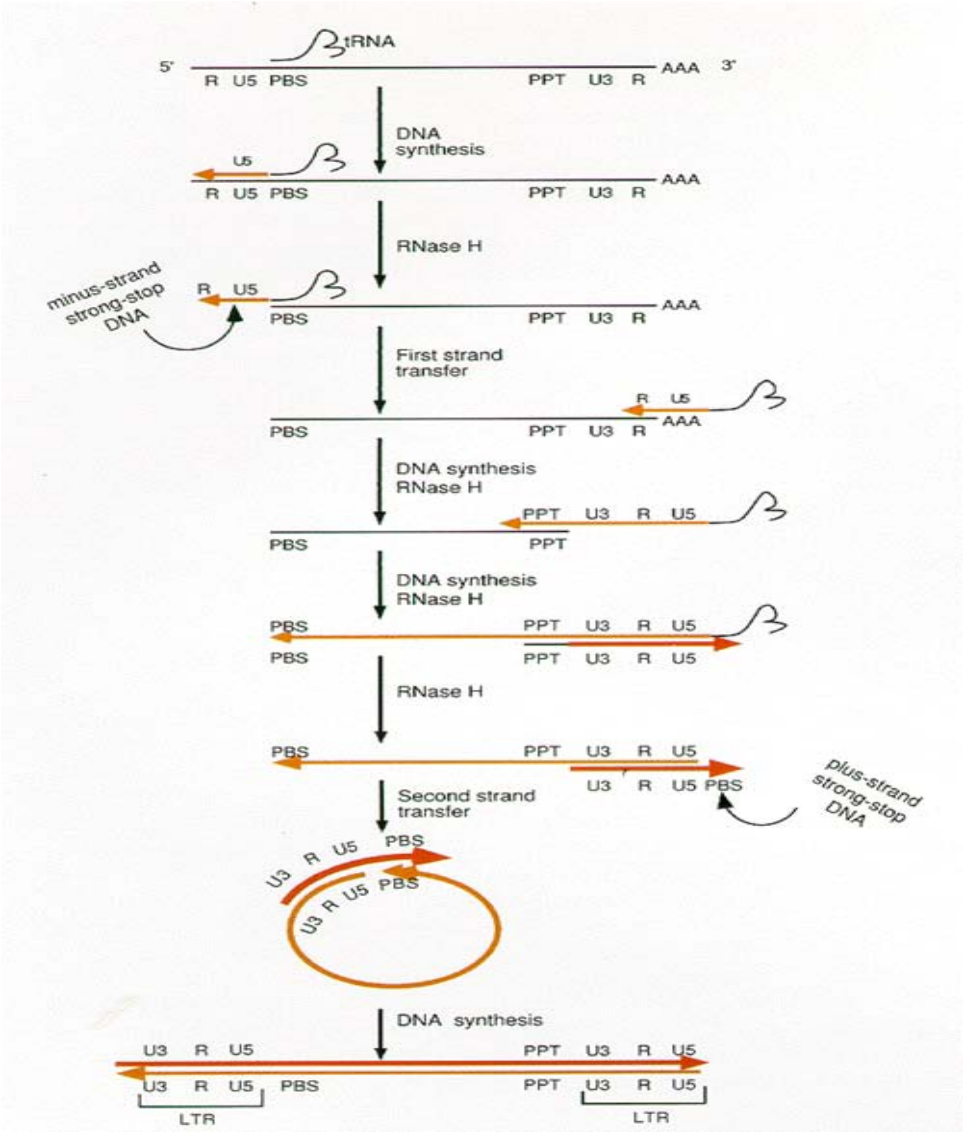

Figure 3. Reverse transcription process in retroviruses. (Black line) RNA; (light color) minus-strand DNAs; (dark color) plus-strand DNA (see the text for details). (Picture adapted from Coffin, JM 1997, *Retroviruses*. Cold Spring Harbor (NY)).

The 5’ LTR acts as potent transcriptional regulatory sequence for proviral gene expression [10]. The LTRs contain *cis*-regulatory elements for specific transcription factors expressed in the cells for which the retroviruses have tropism. The U3 region of the LTRs contains transcription factor binding sites such as a TATA-box and a GC/GT-box for initiation of transcription that usually starts at the 5’ U3-R boundary. The R region also contains a poly-adenylation signal (AAUAAA) which is used for 3’ processing of the retroviral mRNA [1]. The formed viral dsDNA needs to be integrated into the host chromosomal DNA in order to continue its replication cycle. The enzyme integrase (IN) encoded by the viral *pol* gene integrates the proviral DNA into the host chromosome. Thereafter, the newly integrated viral genome is called a provirus [11, 12] that is flanked by the two retroviral LTRs (5’ and 3’). Also, the two ends of the proviral LTRs always contain the same nucleotide sequence (5' - TG...CA 3') [13] and flanked by a duplication of

chromosomal nucleotides at the target site due to integration mechanism (Target Site Duplication, TSD) [14]. After the integration, the viral LTR promoter directs the transcription of viral genes in order to produce the essential proteins needed to make viral particles and after encapsidation and packaging these particles are budded from the infected cell. The polyproteins are cleaved into functional subunits in a process called "maturation" that the virus is ready to infect other host cells. (Figure 4) [11].

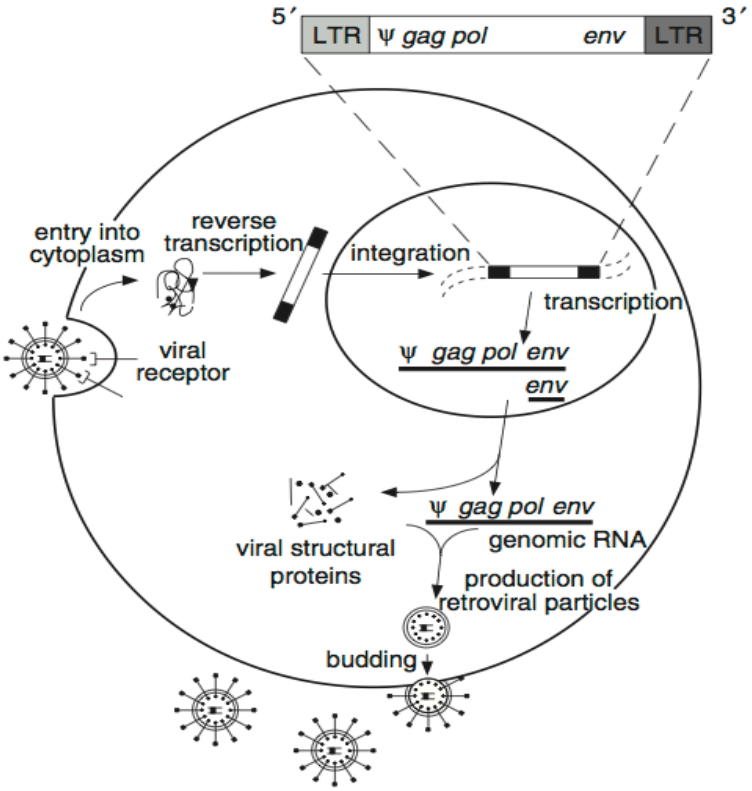

Figure 4. The retroviral life cycle (see the text for details). (Picture adapted from [11]).

4.3 Endogenous retroviruses (ERVs)

Occasionally, XRVs may infect the germ line cells and the acquired provirus can be passed to the offspring as an endogenous retrovirus according to a normal Mendelian inheritance [15].
The first discovered ERVs, avian leukemia virus (ALV) and murine mammary tumor virus (MMTV) were described in the late 1960s and early 1970s (For a review see [16]). The LTRs of ERVs are demonstrated to be identical upon integration into the host chromosomal DNA and at or soon after their integration they accumulate mutations to avoid homologous recombination [17, 18]. The most frequent ERV inactivation is when homologous recombination between two identical LTRs occurs and the coding regions are lost, leaving a solitary LTR (solo LTR) [17, 19]. Therefore, most of the ERVs lost their ability to replicate although there are examples that some family of HERVs are supposed to remained active (*e.g.* HERV-K (HML-2.HOM)) [20].

4.4 Biological relevance of ERVs

It has been reported that at least 7-8 % of human genome is constituted by sequences derived from retroviruses [16] and genome rearrangements driven by retroviral transposition likely have had

effects on plasticity of mammalian genomes. [21]. The potential TF binding sites of retroviral LTRs may affect the expression of surrounding cellular genes as normal promoters or enhancers (*e.g*. HERV-E integration on the antisense strand near the amylase gene, promotes its expression in the human parotid glands) [22] or in some cases act as the primary promoters (Table 2) [23]. Solo LTRs also have shown to act as bidirectional promoters for transcription of nearby genes [24]. They are also able to *cis*-enhance the transcription level of genes and their core promoters (*e.g.* HERV-E enhances the expression of Apolipoprotein C-I gene by influencing on its core promoter) [25]. Another impact of proviral sequences that have been reported is alteration of splicing patterns, which in turn may lead to production of different protein isoforms or premature transcription termination [23].

Table 2. Examples of LTRs functioning as primary promoter [23].

Gene name (full name)	HERV type and location	Function (disease)	LTR expression
ADH1C (alcohol dehydrogenase 1C)	LTR12C/HERV-9 chr4:100493718–100494457	Alcohol metabolism	Liver
GBP5 (guanylate-binding protein 5)	LTR12C/HERV-9 chr1:89510724–89512161	Immune response to intracellular pathogens	Endothelial cells, lymphocytes
HSD17B1 hydroxysteroid 17-beta dehydrogenase 1)	MER21A/ERV1 chr17:37957561–37958104	Estrogen synthesis (breast cancer)	Ovary, placenta
INSL4 (insulin-like 4)	LTR22B/HML-5 chr9:5220548–5221041	Placental morphogenesis	Placenta
PAPPA2 (pappalysin 2)	MER41E/ERV1 chr1:174698554–174699129	Pregnancy (pre-eclampsia)	Placenta
BAAT (bile acid coA; Amino acid N-acyltransferase)	MER11A/HML-8 chr9:103186962–103188103	Bile metabolism (familial hypercholanemia)	Liver
MSLN (mesothelin)	MER54B/ERV-L chr16:750975–751307	Glycoprotein (cancer)	Mesothelium

Important beneficial functions have been linked with the endogenization of retroviruses. Two envelope genes (*syncytin-1* and *syncytin-2* in human and *syncytin-A* and *syncytin-B* in mouse) of retroviral origin were identified recently. They are thought to have been co-opted by mammals from retroviruses during the evolution and have been demonstrated to be essential for trophoblast cell fusion and placental development [26, 27].

Despite of some advantageous roles that ERVs may play in their host, any mutational insertion or homologous recombination in the genome may cause severe damages [16]. Furthermore, several cases of gene duplication were caused by retroelements, especially by LINE1 as well as ERVs [3, 28]. Therefore, a high density of these elements with high degree of sequence similarity would produce ample chances for unequal crossing-over during meiosis.

4.5 Using dog model to study integration polymorphism of ERVs

Genome-wide screening of the publicly available human genome sequence has revealed insertional polymorphism of HERV-K (HML2) family, which suggests the integration of this

family of HERVs after the human divergence [18]. Artificial selection in dogs after their domestication from wolves has created around 400 different breeds of domestic dogs. The domestication process was initiated several thousands of years ago. This resulted in the first genetic bottleneck that limited the genetic variation among the early-domesticated dog types. The modern dog breeds were created around 200 years ago and second genetic bottleneck occurred that in turn further limited the genetic variation and this has been shown by large haplotypes and extensive degree of linkage disequilibrium within breeds [29]. Together, this has resulted in an enormous variation of phenotypic traits among different breeds. Furthermore, a high degree of breed-specific genetic diseases caused by identical-by-descent mutations exist. Dogs and humans live close together and the advantage of sharing the same environment with their owners makes the dog a good model for genetic studies [30]. Moreover, the availability of the genome of a sequenced female boxer dog (canFam2) [31] provides a good quality source to analyze for the potential that canine ERVs exhibit integration polymorphism.

4.6 Canine Endogenous Retroviruses (CfERVs)

Recently an "*in silico*" analysis of the dog genome (canFam2) by our group found 407 proviral CfERVs with average size of 9,187 Kbp. The dog genome (2.5 Gbp) had 3.7 Mbp of CfERVs integrated, which is about 0.15% of its genome size. This amount is considerably lower in comparison with other mammalian species and is almost similar to the red Jungle Fowl (*Gallus gallus*) (Table 3) [32].

Table 3. Number of proviral chains identified by RetroTector and genome percentage for different sequenced species [32].

Species	Chains present	Genome percentage
Dog (*Canis familiaris*)	407	0.15%
Red jungle fowl (*Gallus gallus*)	260	0.20%
Zebra fish (*Danio rerio*)	2048	0.8%
Rhesus macaque (*Macaca mulatta*)	2690	<0.80%
Chimpanzee (*Pan troglodytes*)	2919	<0.80%
Human (*Homo sapiens*)	3149	0.80%
Opossum (*Monodelphis domestica*)	7456	~2%
Mouse (*Mus musculus*)	7582	~2%

Among the most interesting CfERVs identified, a group of HERV-Fc-like proviruses were found to have integrated recently and were further divided into CfERV-Fc like subgroups. There were 33 of these CfERV-Fc proviruses that were estimated to have integrated recently due to the low degree of divergence of their respective 5' and 3' LTRs.

These 33 proviruses were clustered in groups according to the sequence similarity of the Pol protein, which is the most conserved one, and the expected age of integration. These clusters were named CfERV–Fc1 to Fc4-like elements. The Fc4 group with LTR divergence less than 5%, corresponding to an estimated integration time of less than 12 million years ago (mya), and the Fc1 group with more than 10% (over 25 mya) LTR divergence are the youngest and oldest groups, respectively. We focused our analysis on the CfERV-Fc4 group owing to that they possibly could retain potential for active retrotransposition because of less deleterious mutations, which is in agreement with their lower expected time since integration. (Figure 5) [32].

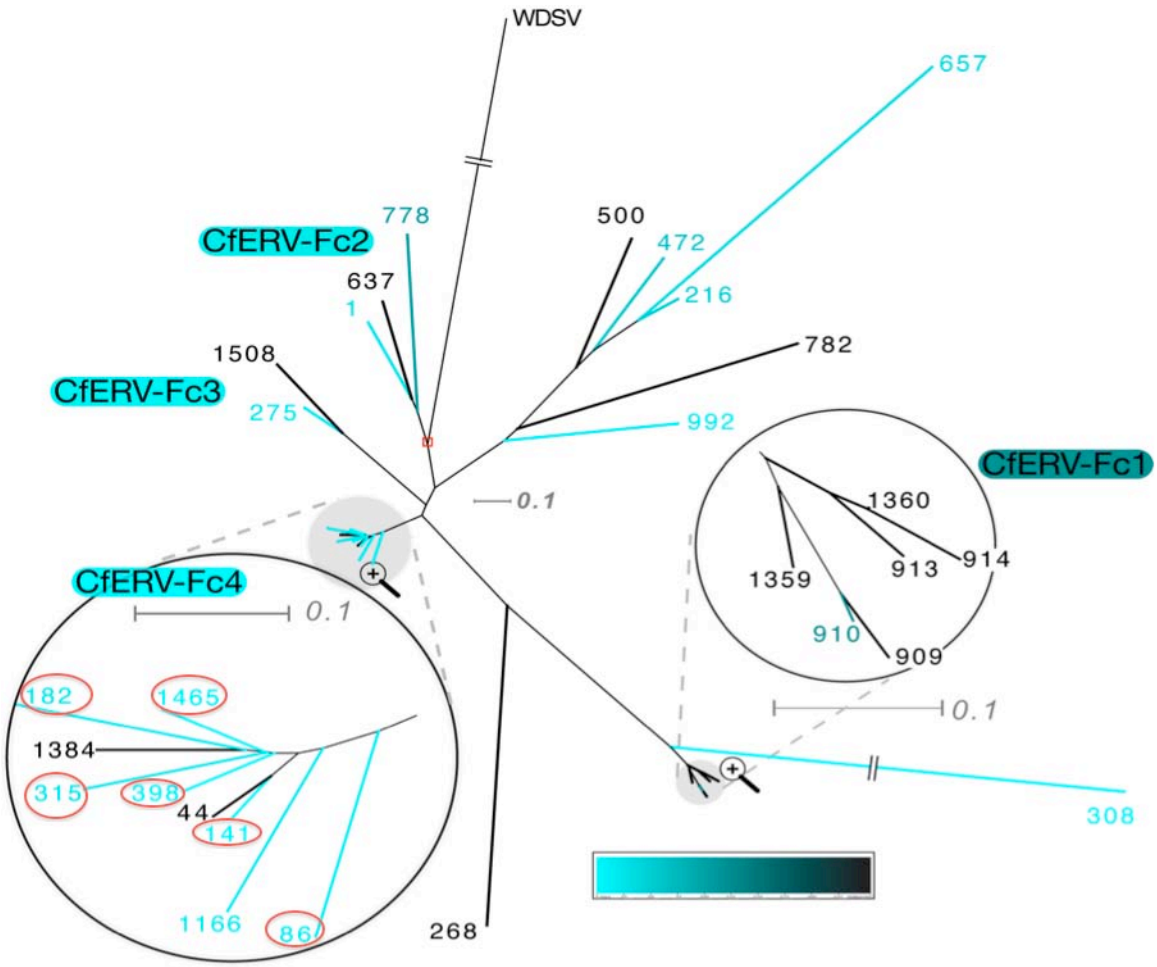

Figure 5. Phylogram showing different CfERV-Fc groups according to their age. Light blue colour indicates the youngest and darker colours represent ancient ones. Black is assigned to undated CfERVs due to lack of any of the LTRs [32].

The CfERVs are mostly integrated in intergenic regions in the autosomes and X chromosomes analysed [32]. The integration landscape of CfERVs showed an indication of selection against their integration in the same transcriptional direction as genes located in their vicinity (<100kb) [32].

The aim of this thesis is to investigate whether integration polymorphism within and between different dog breeds and their ancestor wolf (*Canis lupus*) could be identified. Furthermore, by

taking advantage of some of the most commonly used annotation tools we analyse the possible implication of CfERVs in some discovered canine copy-number variations (CNV).

5 MATERIALS AND METHODS

5.1 Proviral data collection

In order to validate the previous annotations of the Fc4 group LTRs; we used RetroTector (ReTe) online, a program to help identifying and annotating proviral chains in vertebrate sequences [33]. To increase the accuracy of our analysis and improve existing annotations by ReTe, we extracted the positions of both LTRs (provided by the previous study) in each CfERV-Fc4 and some flanking sequence in order to look for an extension that could have been miss-annotated. Then, we aligned both extended LTRs with the program BLASTn (*www.ncbi.nlm.nih.gov/BLAST/)* in order to confirm the real LTRs within each of the sequences extracted (see Results).

5.2 Primer design

The design of the primers was done with the program Primer 3 (http://frodo.wi.mit.edu/primer3). Primers were designed to amplify the identified 5' and 3' CfERV chromosomal integration junctions (Figure 6) of each Fc4 proviral chain. The primers obtained were further analyzed at the IDT website (http://eu.idtdna.com/analyzer/Applications/OligoAnalyzer/) for their melting temperatures (Tm), possible formation of homodimer, hairpin structures and heterodimer with the threshold under -10 for Gibbs free energy (ΔG) to increase the specificity and efficiency of the primers. Primer specificity was also confirmed with the BLAT program (http://genome.ucsc.edu/cgi-bin/hgBlat?command=start) against the dog (canFam2) genome.

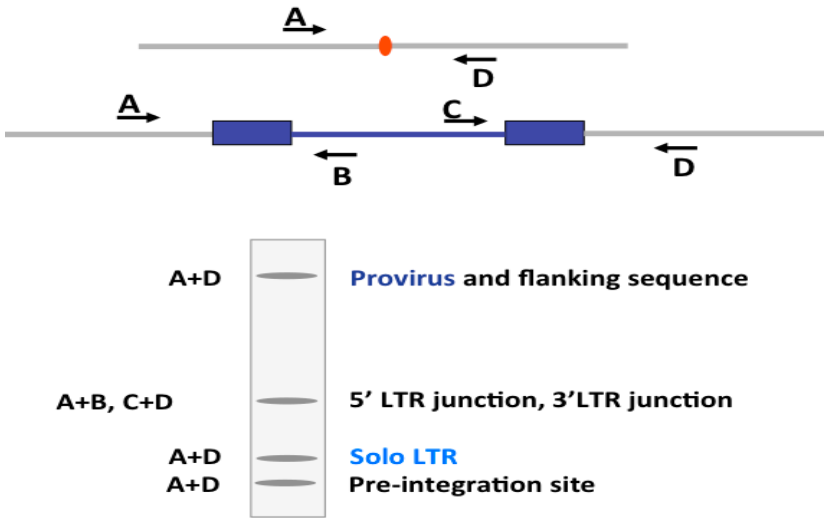

Figure 6. Primer location on the proviral sequence and their amplicon expected product sizes. Labels used: A (forward primer for 5'LTR), B (reverse primer for 5'LTR), C (forward primer for 3'LTR), D (reverse primer for 3'LTR). The thick blue boxes represent LTRs and the blue line in between the coding parts of the ERV. The designed primer sequences are shown in Table 4.

Table 4. The designed primers for CfERV–Fc 4 group according to the Figure 6.

CfERV ID[1]	Chromosomes	Forward	Reverse	Product size[2]
86	Chr2-5'LTR (A-B)	5' CTTCAGGTGGCCATTGTGAT 3'	5'AAGCAAACCAATTTGCGAAG 3'	914 bp
	Chr2-3'LTR (C-D)	5'AAGAAGCGTGGCAACAAAAC3'	5'CATACTCTCGTCTGCCCACA 3'	968 bp
141	Chr3-5'LTR (A-B)	5'GGTGACCCCAGCAGAGAGAG 3'	5'CCTTCTGCACCCCTCATCAC 3'	816 bp
	Chr3-3'LTR (C-D)	5'TTCTTGCAGCCCTTTTCTGA3'	5'TTGGAATCCCTCTCCCAACT 3'	1677 bp
182	Chr5-5'LTR (A-B)	5'CGCTGCCATGTACAGTTACG3'	5'AGCCTGGATACGTTCCCTCT 3'	4648 bp
	Chr5-3'LTR (C-D)	5'GCATGGATGCACATCAACAA3'	5'CACCAGGGAAATGCAATGAC 3'	1061 bp
315	Chr8-5'LTR (A-B)	5'ATCCATCCAACCCTCTTTGG3'	5'CAGCAACCTGGAAGAAATGG 3'	1050 bp
	Chr8-3'LTR (C-D)	5'CTTGCCCACATCCGACTACC3'	5'GCAGAAGGGGAGGTGTATGG 3'	1293 bp
398	Chr11-5'LTR (A-B)	5'GGAATTTGGGGAAGAACAGC3'	5'GGAACCAGGGAACACTCACC 3'	1015 bp
	Chr11-3'LTR (C-D)	5'AGCCATTCCTCCACCTCTTT3'	5'CCCAGATGAGGGACAGTGAT 3'	1312 bp
1465	ChrX-5'LTR (A-B)	5'CATGGACACTTGTGCTGCTT3'	5'AGGGTCAGAAACCAGGGAAC 3'	1068 bp
	ChrX-3'LTR (C-D)	5'CCCCTTCCTGATAGGAGCAG3'	5'GGGAACTAGGCGCACTCTCT 3'	1429 bp
Positive control [3]		5' GATCCCCCCGTCCCCACAG 3'	5' CGCCCGCTGCGCTCA 3'	400 bp

[1] Valid identifiers for Canfam2 genome, analyzed with RetroTector 1.01 on the 2nd December 2007.

[2] Expected product size from the amplification of our genomic target.

[3] Positive control primers were obtained from our colleagues from a previous dog study [34].

The PCR protocol established and the AmpliTaq Gold polymerase reaction setup is described in Table 5.

Table 5. Reaction concentrations used for the experiments.

Reaction component	Final concentration
AmpliTaq Gold Pol (hotstart, 5U)	1U
Taq Pol buffer (10x)	1x
$MgCl_2$ (25 mM)	2mM
dNTP (10mM)	300□□□
FORWARD PRIMER (10□M)	400 nM
REVERSE PRIMER (10□M)	400 nM
H_2O (MQ water)	-
Template DNA (50 ng/µL)	50 ng

We used a touch down PCR (TD-PCR) strategy [35] in order to amplify the genomic sequences for the selected proviral LTRs (Table 6 and Figure 7).

Table 6. TD-PCR protocol

Cycling step	Temperature & time	Cycles
Initial denaturation	10-15 min at 95 °C	
Denaturation	30 sec at 95 °C	
Annealing	45 sec at 65 °C	× 15 and reduce 1 °C every successive cycle
Extension	5 min at 72 °C	
Denaturation	30 sec at 95 °C	
Annealing	45 sec at 50 °C	× 35
Extension	5 min at 72 °C	
Final extension	7 min at 72 °C	

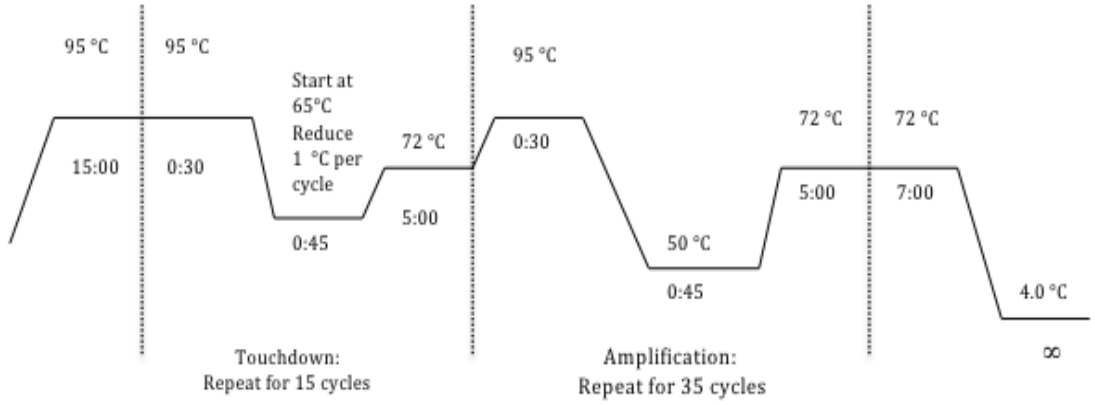

Figure 7. Parameters (temperature, time and cycles) used for the PCR experiments.

5.3 DNA samples

Genomic DNA samples from 7 boxers, 5 poodles and 13 wolves were provided by collaborators and their DNA concentrations were determined using a NanoDrop device (http://www.nanodropdevice.com) (Table 7).

Table 7 a. Boxer DNA samples.

Boxer DNA samples	Individual 1 Female	Individual2 Female	Individual3 Male	Individual4 Female	Individual5 Female	Individual6 Male	Individual7 Male
Initial conc.	93.6 ng/□l	80.3 ng/□l	63.6 ng/□l	57.9 ng/□l	50 ng/□l	50 ng/□l	50 ng/□l

Table 7 b. Poodle DNA samples. *

Poodle DNA samples	Individual 1	Individual2	Individual3	Individual4	Individual5

Initial conc.	83.5 ng/□l	49.5 ng/□l	41.1 ng/□l	36.4 ng/□l	31.3 ng/□l

*The sex of the poodle DNA samples is uncharacterized.

Table 7 c. Wolf DNA samples.

Wolf DNA samples	1 M[1]	2 M	3 M	4 M	5 F[2]	6 F	7 M	8 F	9 F	10 M Alive	11 M Alive	12 M	13 M
Initial conc.	500 ng /□l	400 ng /□l	300 ng /□l	300 ng /□l	400 ng /□l	400 ng /□l	300 ng /□l	300 ng /□l	25 ng /□l	25 ng /□l	25 ng /□	25 ng/ □	25 ng/ □

[1] M - Male

[2] F – Female

5.4 Copy Number Variation (CNV) data

A dataset of CNVs in 17 dog breeds and wolf produced by our collaborators [Jonas Berglund and Matthew T. Webster; Department of Medical Biochemistry and Microbiology, Uppsala University, unpublished] was used to show a possible correlation between our identified CfERVs and the CNVs by overlapping these two sets' genome loci.

5.5 Bioinformatics tools

In order to find overlaps between the CNV data and CfERVs, we used a program (intersectBed) from the bioinformatics package BEDTools [36]. BEDTools are fast, flexible and reliable tools to compare large sets of genomic features. Another program we used during this study is AWK and it is used to search for certain patterns within lines or other units of text in files (http://www.gnu.org/software/gawk/manual/gawk.html).

6 Results

6.1 Integration Polymorphism of CfERVs

To experimentally validate whether integration polymorphism of CfERVs exist in different dog breeds and their ancestor, wolf, we used the sequences and genomic annotations from our previous datasets [32]. In order to validate the actual LTR positions and avoid that primers would overlap repetitive LTRs, we re-annotated some of our previous data (Table 8a, 8b) and our experimental primers were designed (see primer locations in the Materials and Methods section) for CfERV-Fc4 regarding these new annotations.

Table 8 a. 5'LTR positions obtained from previous study [32] and re-annotated each LTR locus.

		Previous annotation		Reannotation			
Chr	**CfERV-Fc4 identifier**	**Start**	**End**	**Start**	**End**	**Retrovirus class**	**Strand**
Chr2	86	68192127	68191653	-[2]	-	Unclassified	(-)
Chr3	141	85023462	85023328	85023639	85023194	Gamma-like	(-)
Chr5[1]	182	27587508	27587754	-	-	Gamma-like	(+)
Chr8	315	76899653	76899853	76899574	76900048	Gamma-like	(+)
Chr11	398	15757292	15757702	-	-	Gamma-like	(+)
ChrX	1465	53586021	53586128	53585819	53586275	Unclassified	(+)

[1] Re-annotation showed that CfERV-Fc4 on chr 5 needs to be excluded from the CfERV-Fc4 group due to higher LTR divergence.
[2] Dashes mean that there is no re-annotation because the previous annotation in [32] is correct.

Table 8 b. 3'LTR positions obtained from previous studies [32] and re-annotated LTR positions.

		Previous annotation		Re-annotation			
Chr	**CfERV-Fc4 code**	**Start**	**End**	**Start**	**End**	**Retrovirus class**	**Strand**
Chr2	86	68186084	68185605	-[2]	-	Unclassified	(-)
Chr3	141	85016067	85015933	85016244	85015845	Gamma-like	(-)
Chr5[1]	182	27593915	27594165	27593594	27594093	Unclassified	(+)
Chr8	315	76914785	76914985	76914717	76915183	Gamma-like	(+)
Chr11	398	15763594	15764004	-	-	Unclassified	(+)
ChrX	1465	53593428	53593535	53593230	53593682	Unclassified	(+)

[1] Re-annotation showed that CfERV-Fc4 on chr5 needs to be excluded from the youngest CfERV group due to higher LTR divergence.
[2] Dashes mean that there is no re-annotation because the previous annotation in [32] is correct.

Locus-specific primers for CfERV LTRs chromosomal junction sequences were able to amplify the genomic DNA of boxer, poodle and wolf. PCR products from different individual DNA samples (Table 7) were obtained with a TD-PCR strategy to increase the primer binding affinity without time-consuming optimizations [35]. TD-PCRs were run multiple times and almost all generated the same pattern of bands with the expected product sizes after running in 1% agarose gel at 80 V cm^{-1} for 45 minutes. It is apparent from the results (Figure 8-11) that the banding patterns obtained are similar in all the individuals analyzed for these two breeds and the wolves.

Figure 8.

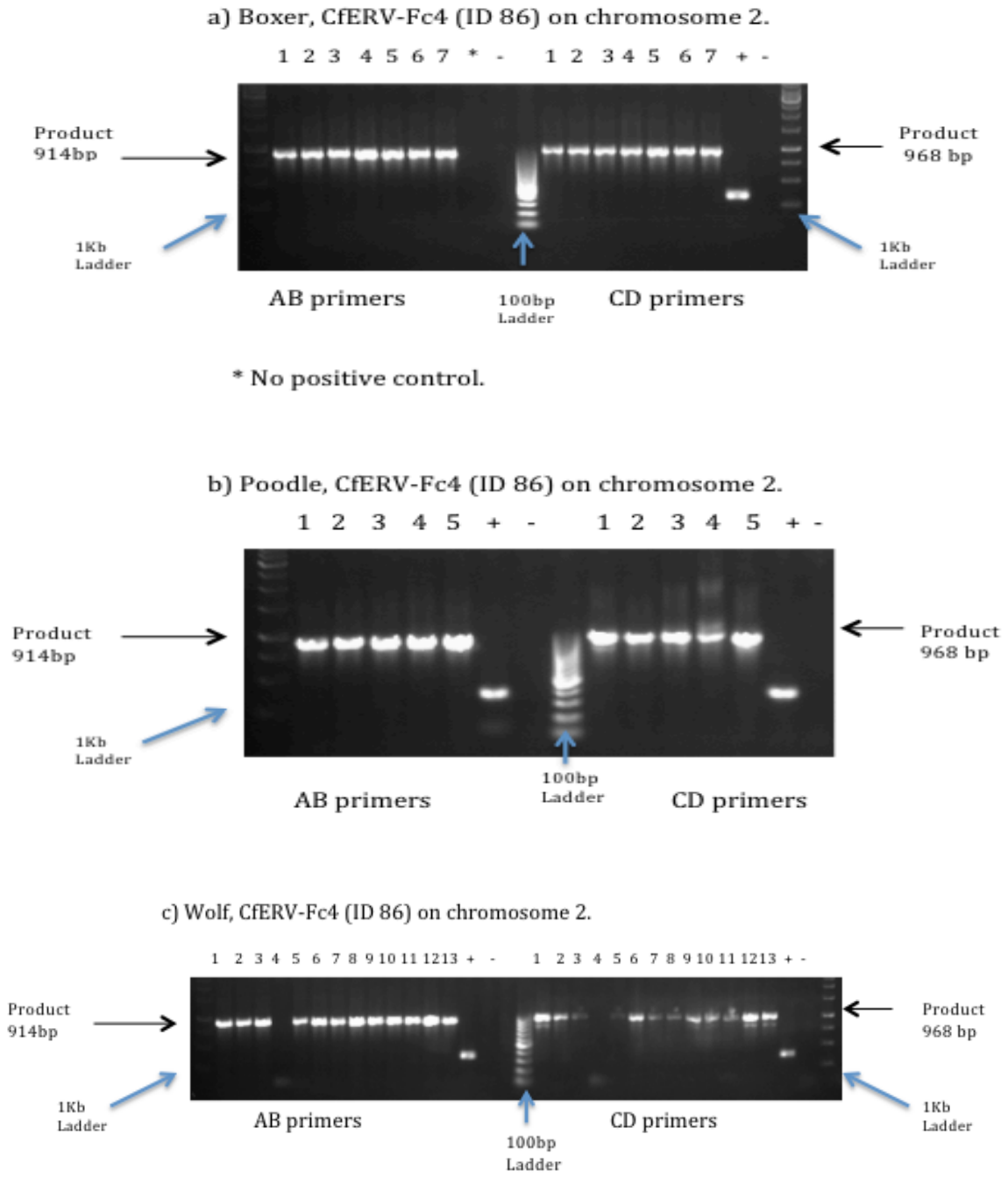

Figure 8. Agarose gel of PCR amplification products of CfERV-Fc4 on **chromosome 2** from boxer, poodle and wolf DNA samples using primers described in Table 4.

a) Left: AB primers for 5'LTR. Lanes 1-7 indicate boxer DNA of different individuals. Right: CD primers for 3' LTR. Lanes 1-7 indicate boxer DNA of different individuals and the positive control (see Table 4) for individual No. 2 DNA with 400 bp expected product size.

b) Left: AB primers for 5'LTR. Lanes 1-5 poodle DNA of different individuals and the positive control (see Table 4) for individual's 1 DNA with 400 bp expected size. Right: CD primers for 3' LTR. Lanes 1-7 indicate poodle DNA of different individuals and the positive control (see Table 4) for individual No. 2 DNA with 400 bp expected product size.

c) Left: AB primers for 5'LTR. Lanes 1-13 wolf DNA of different individuals and the positive control (see Table 4) for individual's 1 DNA with 400 bp expected size. Right: CD primers for 3' LTR. Lanes

1-13 indicate wolf DNA of different individuals and the positive control (see Table 4) for individual No. 2 DNA with 400 bp expected product size.

Figure 9.

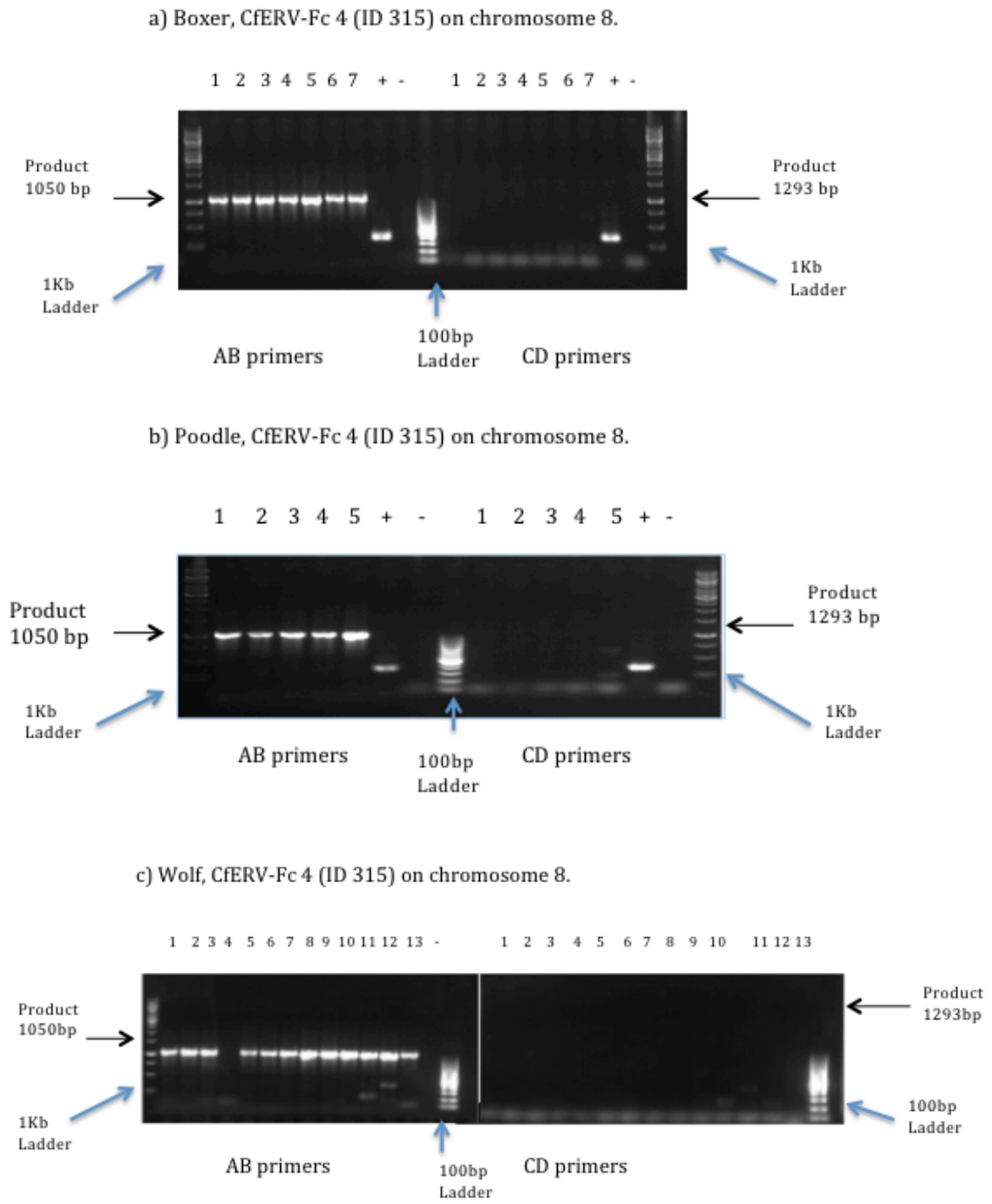

Figure 9. Agarose gel of PCR amplification products of CfERVs-Fc4 on **chromosome 8** from boxer, poodle and wolf DNA samples using primers described in the table 4.

a) Left: AB primers for 5'LTR. Lanes 1-7 indicate boxer DNA of different individuals and the positive control (see Table 4) for individual No. 3 DNA with 400 bp expected product size. Right: CD primers

for 3' LTR. Lanes 1-7 indicate boxer DNA of different individuals and the positive control (see Table 4) for individual No. 4 DNA with 400 bp expected size.
b) Left: AB primers for 5'LTR. Lanes 1-5 poodle DNA of different individuals and the positive control (see Table 4) for individual No. 3 DNA with 400 bp expected size. Right: CD primers for 3' LTR. Lanes 1-7 indicate poodle DNA of different individuals and the positive control (see Table 4) for individual No. 2 DNA with 400 bp expected product size.
c) Left: AB primers for 5'LTR. Lanes 1-13 wolf DNA of different individuals with no positive control. Right: CD primers for 3' LTR. Lanes 1-13 indicate wolf DNA of different individuals.

Figure 10.

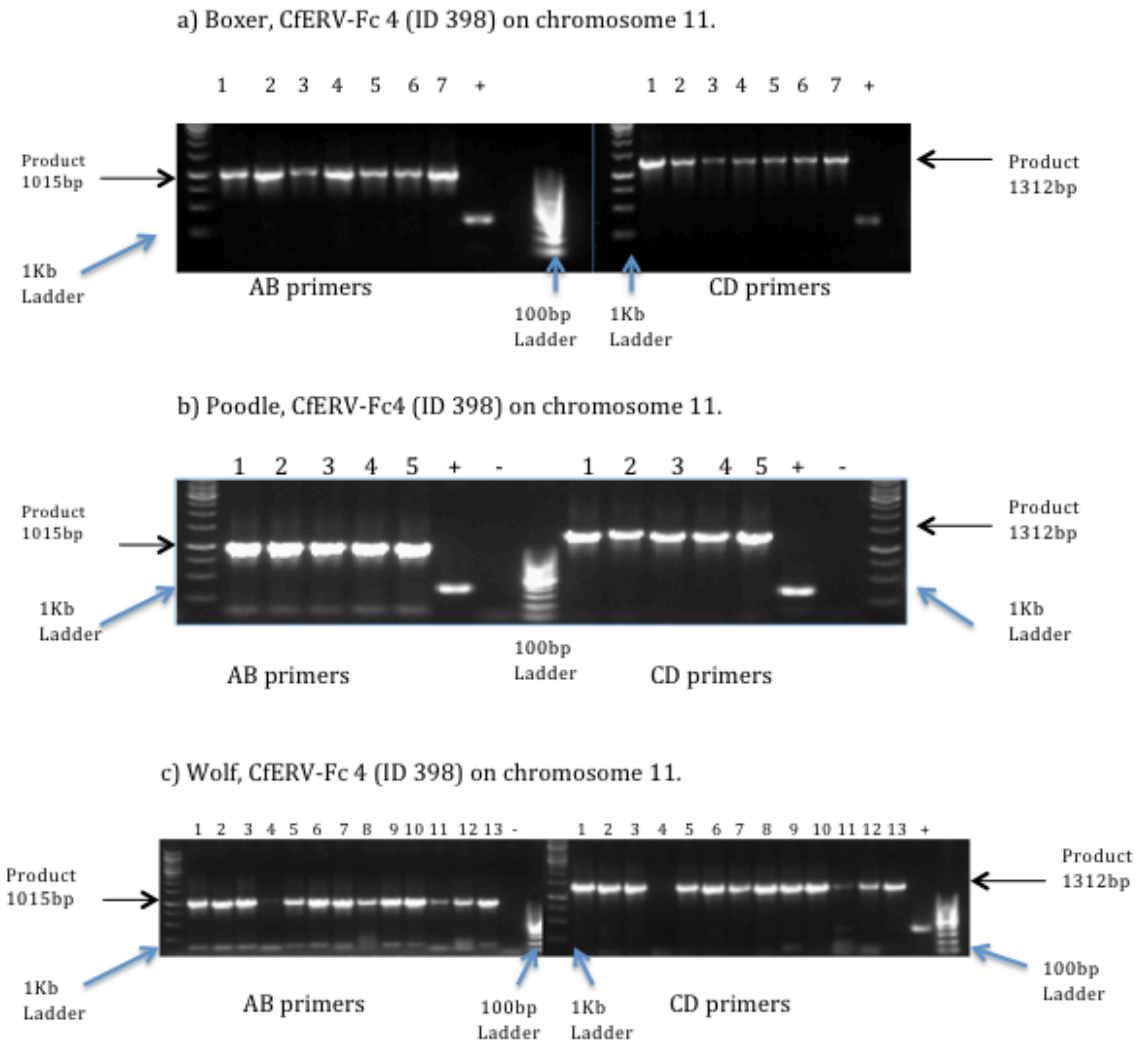

Figure 10. Agarose gel of PCR amplification products of CfERVs-Fc4 on **chromosome 11** from boxer, poodle and wolf DNA samples using primers described in Table 4.
a) Left: AB primers for 5'LTR. Lanes 1-7 indicate PCR products obtained with boxer DNA of different individuals and the positive control (see Table 4) for individual No. 5 DNA with 400 bp expected product size. Right: CD primers for 3' LTR. Lanes 1-7 indicate boxer DNA of different individuals and the positive control (see Table 4) for individual No. 6 DNA with 400 bp expected size.
b) Left: AB primers for 5'LTR. Lanes 1-5 poodle DNA of different individuals and the positive control (see Table 4) for individual No. 5 DNA with 400 bp expected product size. Right: CD primers

for 3' LTR. Lanes 1-7 indicate PCR products obtained with poodle DNA of different individuals and the positive control (see Table 4) for individual No. 1 DNA with 400 bp expected product size.
c) Left: AB primers for 5'LTR. Lanes 1-13 wolf DNA of different individuals with no positive control. Right: CD primers for 3' LTR. Lanes 1-13 indicate wolf DNA of different individuals with no positive control. Right: CD primers for 3' LTR. Lanes 1-13 indicate wolf DNA of different individuals and the positive control (see Table 4) for individual No.3 DNA with 400 bp expected product size.

Figure 11.

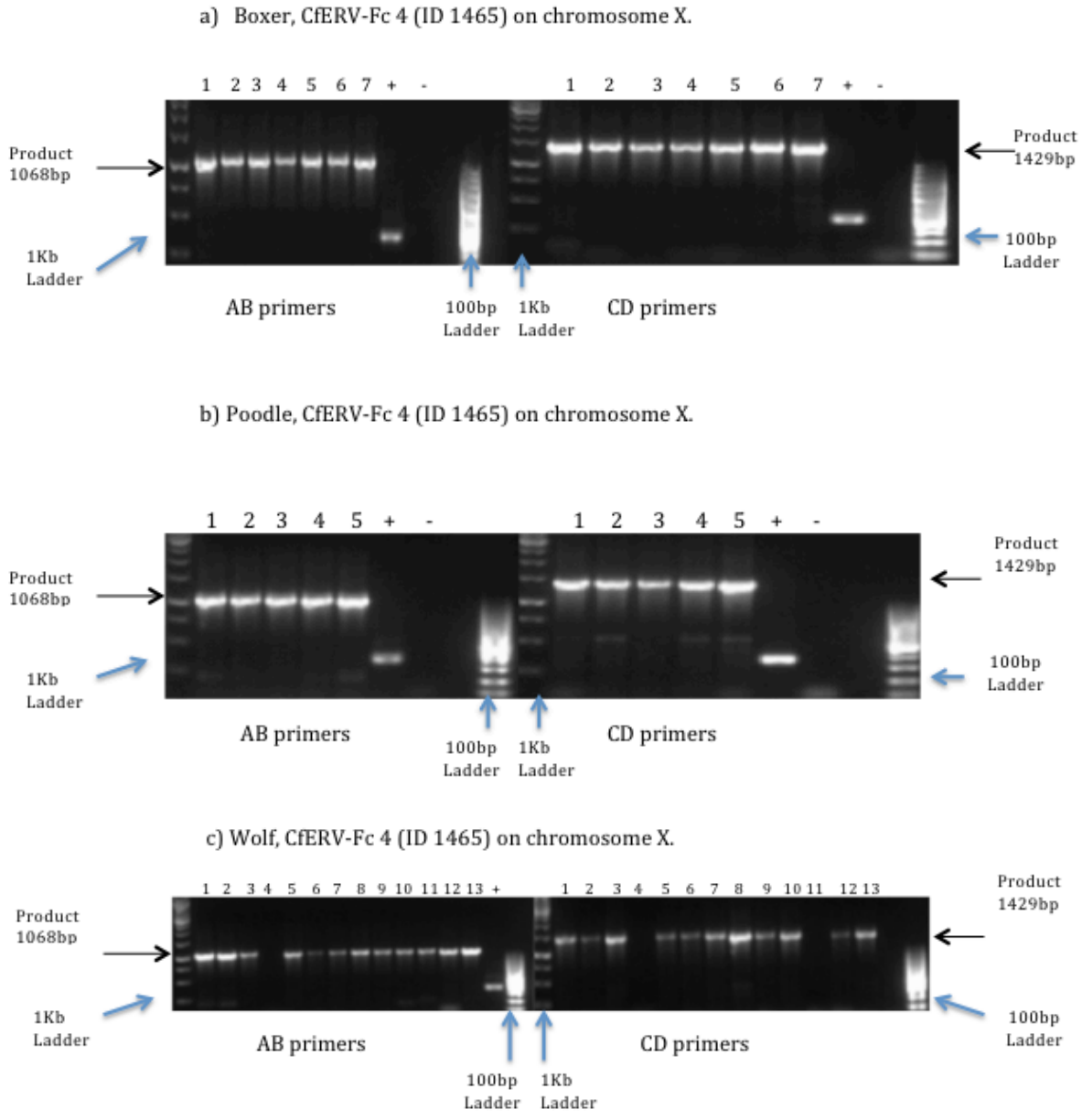

Figure 11. Agarose gel of PCR amplification products of CfERVs-Fc4 on **chromosome X** from boxer, poodle and wolf DNA samples using primers described in the Table 4.

a) Left: AB primers for 5'LTR. Lanes 1-7 indicate boxer DNA of different individuals and the positive control (see Table 4) for individual No.7 DNA with 400 bp expected size. Right: CD primers for 3' LTR. Lanes 1-7 indicate boxer DNA of different individuals and the positive control (see Table 4) for individual No.1 DNA with the 400 bp expected product size.

b) Left: AB primers for 5'LTR. Lanes 1-5 poodle DNA of different individuals and the positive control (see Table 4) for individual No.1 DNA with 400 bp expected size. Right: CD primers for 3' LTR. Lanes 1-7 indicate poodle DNA of different individuals and the positive control (see Table 4) for individual No.5 DNA with 400 bp expected product size.
c) Left: AB primers for 5'LTR. Lanes 1-13 wolf DNA of different individuals with the positive control (see Table 4) for individual No.3 DNA with 400 bp expected product size. Right: CD primers for 3' LTR. Lanes 1-13 indicate wolf DNA of different individuals with no positive control.
* Due to lacking of sufficient amount of DNA, some samples do not have any positive control.

All amplification of 3'LTRs on chromosome 8 failed. The CfERV on chromosome 3 also did not generate PCR products. As mentioned above, re-annotation of LTRs indicated that CfERV on chromosome 5 is older than previously estimated based on the LTR divergence and from that point it was excluded from the primer design stage. The long range PCR was run for A-D primers failed for all loci with yet unknown reasons.

6.2 CNV analysis

To investigate any implication regarding CfERV and copy number variations at certain genomic loci in dogs, especially since CfERVs could possibly be a good mean of providing non-allelic homologous recombination (NAHR), a comparative analysis of both available data sets (*i.e.* CNV and CfERV data) was performed with the help of BEDTools-intersectBed [36]. We found two CfERV-Fc loci that overlapped with CNV (Table 9). Further analysis of the integration on chromosome 26 showed no obvious impact of the CfERV annotated there, whereas the CNV overlapped a region on chromosome 8 which showed a duplication of parts of the *Gpm6b* gene and our CfERV is right at the middle (Figure 12). This gene encodes for a membrane glycoprotein, which belongs to a proteolipid protein family. According to Ensembl, NCBI and UCSC genome databases, the *Gpm6b* gene is present at chromosome X in all mammals. This proteolipid family of proteins is expressed in most part of brain regions and thought to be involved in cellular housekeeping functions such as membrane trafficking and cell-cell communication. This gene is also conserved between dog, cow, mouse, rat, chicken and zebrafish [37].

Table 9. CfERVs LTRs overlapped CNVs.

Chromosome	Start	End
Chr8	76786198	77097282
Chr26	31904960	31978529

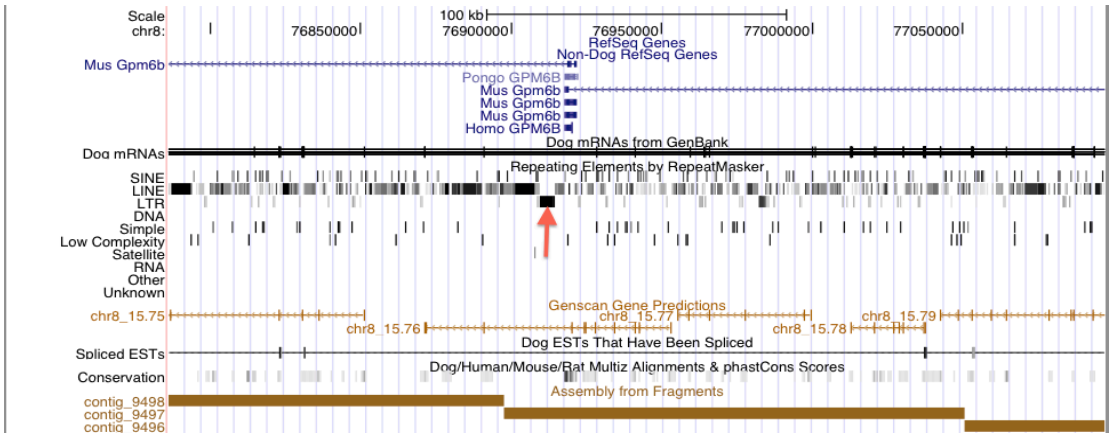

Figure 12. CfERV at chromosome 8 and duplicated parts of the Gpm6b gene (Picture from UCSC genome browser).

7 DISCUSSION AND FUTURE PROSPECTS

Retroviral-related sequences have evolved diverse functional mechanisms such as retrotransposition, insertion, deletion and substitution during evolution, and occupied 0.15% of the dog genome [32]. This considerably low amount of retroviral related sequences in dog could be due to different restriction mechanisms or suggests that canids have been confronted with fewer retroviral infections. Previous studies demonstrated that APOBEC ("apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like") and TRIM ("Tripartite motif-containing") genes in human and in most other mammals are involved in modulation of retrovirus infectivity, replication, and assembly [38, 39]. Therefore, the existence of such restriction mechanisms needs to be addressed in dogs.

7.1 Lack of Integrational polymorphism

By using PCR-based strategy we were unable to find evidence that the CfERV–Fc4 group of canine endogenous retroviruses exhibit integrational polymorphism among those few investigated individuals. The confirmation that integration of this group of CfERVs is non-polymorphic would require analysis of DNA samples from more individuals.

Given that the dogs and wolves diverged approximately 15,000-31000 years ago [31, 32] and they both have endogenous retroviral sequences from the “youngest family” integrated at identical positions without any evidence for integrational polymorphism, suggests that their integration preceded their divergence. Therefore, most CfERVs were incorporated into the canid genome prior to the dog domestication. This hypothesis needs to be confirmed by analysing other canidae species like fox, coyote etc. Furthermore, since these endogenous retroviral sequences have not been retrotransposed after wolf and dog divergence, it is likely that these CfERV-Fc4 group members have been present in the wolves and dogs for a fairly long period of time as previously predicted by their LTRs divergence rate. An important matter to consider is, as we have explained before, at the time of integration, LTRs are identical and there is a higher probability that recombination between LTRs occurs resulting in the formation of solo LTRs, which leads to deletion of the internal sequence. Previous investigations showed that proviruses obtaining mutations in their LTRs at the integration time have the less likely chance of recombination and

remain in their full-length [17]. Therefore, our methods for dating ERVs based on the estimation of LTRs divergence likely cause false approximations of ERVs integration time. The conservation of these sequences for such a long time confirms that the majority of ERVs integrations are selectively neutral, whereby these retroviral sequences could have been remained conserved during evolution.

7.2 Copy Number Variation

Comparative analysis of CNV and CfERV-Fc data sets showed an overlap on chromosome 8, where the *Gpm6b* gene is located. Since the *Gpm6b* gene is present at chromosome X in all mammals, the hypothesis that a CfERV could recombine material from the chromosome X into chromosome 8 remains to be assessed to see what parts of *Gpm6b* from the chromosome X has been inserted into chromosome 8. Since the CfERV is present in the middle of the duplicated loci, it could indicate that the CfERV has played a mediator role during such recombination events.

The results obtained in the current study provide evidences of conservation of CfERV- Fc4 group. Therefore, gene expression studies are required to support that certain candidate CfERV could affect the level of transcription of the nearby genes. Regarding the investigation of integration polymorphism, a key experiment is to carefully mine a wolf genome with RetroTector and also analyse more dogs, wolves and other *Canidae* species like fox and coyote to confirm the absence of integration polymorphism. Another important question that needs to be addressed is how to obtain more knowledge about the plausible restriction or purging mechanism of XRVs in canids in order to determine how dogs have been infected in comparison with humans despite their intimate interaction for thousands of years.

8 ACKNOWLEDGMENTS

This research project would not have been possible without the kindly supports of my supervisors, Prof. Göran Andersson, Dr. Patric Jern and Dr. Alvaro Martinez Barrio. I would like to express my heartfelt gratitude to them for all helpful, valuable suggestions and guidance. Special thanks also to all the lab members whose knowledge assisted me to succeed. I also wish to express my love and gratitude to my wife and beloved families; for their understanding and endless love, through the duration of my studies.

9 REFERENCES:

1. Jern P (2005) Genomic Variation and Evolution of HERV-H and other Endogenous Retroviruses (ERVs) Acta Universitatis Upsaliensis. *Digital Comprehensive summaries of Uppsala Dissertations from the Faculty of Medicine.* Vol (62): 77.
2. Baltimore D (1970) RNA-dependent DNA polymerase in virions of RNA tumour viruses. *Nature.* Vol (226): 1209-11.
3. Han JS (2010) Non-long terminal repeat (non-LTR) retrotransposons: mechanisms, recent developments, and unanswered questions. *Mobile DNA.* Vol (1): 15.
4. Belancio VP, Hedges DJ, Deininger P (2008) Mammalian non-LTR retrotransposons: For

better or worse, in sickness and in health. *Genome Res.* Vol (18): 343-358.
5. Deininge PL, Batzer MA (2002) Mammalian Retroelements. *Genome Res.* Vol (12): 1455-1465.
6. Coffin JM (1979) Structure, Replication, and Replication of Retrovirus Genomes: Some Unifying Hypotheses. *J. Gen. Virol.* Vol (42): 1-26.
7. Jern P, Sperber G.O, and Blomberg J (2005) Use of Endogenous Retroviral Sequences (ERVs) and structural markers for retroviral phylogenetic inference and taxonomy. *Retrovirology*. Vol (2): 50.
8. Temin HM (1985) Reverse Transcription in the Eukaryotic Genome: Retroviruses, Pararetroviruses, Retrotransposons, and Retrotranscripts. *Mol. Biol. Evol*. Vol (6): 455-468.
9. Steinhauer DA, Domingo E, Holland JJ (1992) Lack of evidence for proofreading mechanisms associated with an RNA virus polymerase. *Gene*. Vol (122): 281-288.
10. Jern P, Sperber G.O, Ahlsén G, and Blomberg J (2005) Sequence Variability, Gene Structure and Expression of Full- Length Human Endogenous Retrovirus H. *J. Virol*. Vol 79(10): 6325-6337.
11. Cepko C, Pear W (1996) Transduction Of Genes Using Retrovirus Vectors, Overview of the Retrovirus Transduction System. *Current Protocols in Molecular Biology*. Vol 9.9.1-9.9.16.
12. Nisole S, Saïb A (2004) Early steps of retrovirus replicative cycle. *Retrovirology*. Vol (1): 9.
13. Lee N, Harshey RM (2003) Patterns of sequence conservation at termini of long terminal repeat (LTR) retrotransposons and DNA transposons in the human genome: lessons from phage Mu. *Nucleic Acids Res*. Vol (31): 4531-4540.
14. Taganov K et al (2001) Characterization of Retrovirus-Host DNA Junctions in Cells Deficient in Nonhomologous-End Joining. *J. Virol.* Vol (75): 9549–9552.
15. Jern P, Coffin JM (2008) Effects of retroviruses on host genome function. *Annu. Rev. Genet*. Vol *(*42): 709–732.
16. Weiss RA (2006) The discovery of endogenous retroviruses. *Retrovirology*. Vol (3): 67.
17. Belshaw R et al (2007) Rate of Recombinational Deletion among Human Endogenous Retroviruses. *J. Virol.* Vol (81): 9437–9442.
18. Belshaw R et al (2005) Genomewide Screening Reveals High Levels of Insertional Polymorphism in the Human Endogenous Retrovirus Family HERV-K (HML2): Implications for Present-Day Activity. *J. Virol*. Vol (79): 12507–12514.
19. Bock M, Stoye JP (2000) Endogenous retroviruses and the human germline. *Current Opinion in Genetics & Development*. Vol (10): 651–655.
20. Belshaw R et al (2005) High Copy Number in Human Endogenous Retrovirus Families is Associated with Copying Mechanisms in Addition to Reinfection. *Mol. Biol. Evol.* Vol 22(4): 814–817.
21. Eiden M.V (2008) Endogenous retroviruses – Aiding and abetting genomic plasticity. *Cell. Mol. Life Sci.* Vol (65): 3327-3328.
22. Ting CN, Rosenberg MP, Snow CM, Samuelson LC, Meisler MH (1992) Endogenous retroviral sequences are required for tissue- specific expression of a human salivary amylase gene. *Genes Dev.* Vol (6): 1457-1465.
23. Cohen CJ, Lock WM, Mager DL (2009) Endogenous retroviral LTRs as promoters for human genes: A critical assessment. *Gene*. Vol (448): 105–114.
24. Friesen PD, Rice WC, Miller DW, Miller LK (1986) Bidirectional Transcription from a Solo Long Terminal Repeat of the Retrotransposon TED: Symmetrical RNA Start Sites. *Mol. Cell. Bio.* Vol (6): 1599-1607.
25. Medstrand P, Landry JR, Mager Dl (2001) Long Terminal Repeats Are Used as Alternative Promoters for the Endothelin B Receptor and Apolipoprotein C-I Genes in Humans. *J. Biol. Chem*. Vol (276): 1896-1903.

26. Dupressoir A et al (2011) A pair of co-opted retroviral envelope syncytin genes is required for formation of the two-layered murine placental syncytiotrophoblast. *Proc. Natl. Acad. Sci. U. S. A.* Vol (10): 1073.
27. Pötgens A.J.G (2004) Syncytin: the major regulator of trophoblast fusion? Recent developments and hypotheses on its action. *Human Reproduction Update.* Vol (10): 487-496.
28. Balakrishnan CN et al (2010) Gene duplication and fragmentation in the zebra finch major histocompatibility complex. *BMC Biology*. Vol (8): 29.
29. Karlsson EK et al (2007) Efficient mapping of mendelian traits in dogs through genome-wide association. *Nat Genet. Vol (39): 1321-1328.*
30. Cruz F, Vila C, Webster M.T (2008) The Legacy of Domestication: Accumulation of Deleterious Mutations in the Dog Genome. *Mol. Biol. Evol*. Vol 25 (11): 2331–2336.
31. Lindblad-Toh K et al (2005) Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature.* Vol (438|8): 803-819.
32. Barrio AM et al (2011) The First Sequenced Carnivore Genome Shows Complex Host-Endogenous Retrovirus Relationships. *PLoS ONE*. Vol 6 (5): e19832.
33. Sperber G et al (2009) RetroTector online, a rational tool for analysis of retroviral elements in small and medium size vertebrate genomic sequences. *BMC Bioinformatics*. Vol 10 (Suppl 6): S4.
34. Wilbe M et al (2010) DLA Class II Alleles Are Associated with Risk for Canine Symmetrical Lupoid Onychodystropy (SLO). *PLoS ONE.* Vol 5(8): e12332.
35. Korbie DJ, Mattick JS (2008) Touchdown PCR for increased specificity and sensitivity in PCR amplification. *Nature Protocols* .Vol (3): 1452-1456.
36. Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. Vol (26): 841–842.
37. Drabek K et al (2011) GPM6B Regulates Osteoblast Function and Induction of Mineralization by Controlling Cytoskeleton and Matrix Vesicle Release. *Journal of Bone and Mineral Research*. Vol (26): 2045–2051.
38. Towers GJ (2007) The control of viral infection by tripartite motif proteins and cyclophilin A. *Retrovirology.* Vol *(*4): 40.
39. Münk C et al (2008) Functions, structure, and read-through alternative splicing of feline APOBEC3 genes. *Genome Biol*. Vol (9): R48.