Institutionen för husdjursgenetik

Molecular Systematics: Data mining of canine endogenous retroviruses, *CFERV*

by

Marie Ekerljung

Handledare:
Göran Andersson
Göran Sperber
Jonas Blomberg

Examensarbete 295
2007

Examensarbete ingår som en obligatorisk del i utbildningen och syftar till att under handledning ge de studerande träning i att självständigt och på ett vetenskapligt sätt lösa en uppgift. Föreliggande uppsats är således ett elevarbete och dess innehåll, resultat och slutsatser bör bedömas mot denna bakgrund. Examensarbete på D-nivå i ämnet husdjursgenetik, 20 p (30 ECTS).

Institutionen för husdjursgenetik

Molecular Systematics: Data mining of canine endogenous retroviruses, *CFERV*

by

Marie Ekerljung

Agrovoc: DNA, germ cell, transposable elements
Övrigt: Canine endogenous retroviruses (CFERV), horisontal contagion

Handledare:
Göran Andersson *Dept. Animal Breeding and Genetics , SLU*
Göran Sperber *Dept. Neuroscience, Uppsala University*
Jonas Blomberg *Dept. Medical Science, Uppsala University*

Examensarbete 295
2007

Table of contents

Abbreviations 2
Abstract 2
Background 3
Materials and methods 3
Endogenous retroviruses (Retroviridae) 4
The structural organization of a complete endogenous retrovirus 5
Results 5
Classification of *CFERV* 5
Score 6
Genus 7
Completeness for the CFERVs 8
Average score within the groups, defined based on different numbers of genes in the chains 10
Chromosomal distribution 11
Rep Base Finds 12
Comparison between human *HERVF* and *HERVF*-like *CFERV* 13
Class I gammaretroviruses (type C) 15
Score proportions in Gammaretroviruses 19
Completeness in CF gammaretroviruses 19
Proportions of genes within the CF gammaretroviruses 20
Class II betaretroviruses, β and alpharetroviruses, α 21
Proportions of genes and completeness within the *CF* Betaretroviruses 21
Class III Spuma ς retroviruses 22
CF Delta δ –like retroviruses 23
Comparison of *CFERV* and endogenous retroviruses from other species 24
LTR divergence and ORF 27
Discussion 28
RetroTector© 29
Conclusions 29
Acknowledgements 31
Svensk sammanfattning 31
References 32

Abbreviations

AVL	avian leukosis virus
BaEv	baboon endogenous retrovirus
BLV	bovine leukemia virus
bp	base pairs (nucleotides)
cap	capsid
CF	canis familiaris
CFERV	C F endogenous retroviral virus
DNA	deoxyribonucleic acid
Env	envelope
ERV	endogenous retrovirus
Gag	group specific antigen
GaLV	gibbon ape leukemia virus
HERV	human endogenous retrovirus
HFV	human foamy virus
HIV	human immunodeficiency virus
JSRV	jaagsiekte (sheep) retrovirus
Kb	kilo basepairs
L1, LINE	long interspersed nucleotide element
LTR	long terminal repeat
MLV	murine leukemia virus
MMTV	mouse mammary tumour virus
ORF	open reading frame
PBS	primer binding site
Putein	putative protein
Pol	polymeras
Pro	protease
RV	retrovirus
TE	tranposable elements
tRNA	transfer ribonucleic acid
XRV	exogenous retrovirus

Abstract

Endogenous retroviruses (ERVs) were discovered in the 1960s and in their complete form they consist of two *LTRs* (long terminal repeat sequences) that are each located on either side of the chain/element. ERVs contain the following genes: *gag, pro, pol* and *env* .*C*complete

ERVs are 8-10 kb long and have their best fitness in the enzyme that use reverse transciptase polymerase from single stranded RNA as template to produce DNA and in the possibility the host is not damaged lethally and no selection against the *ERV* element come into force. When a retrovirus infects a germ cell, the retrovirus inserts its genome into the infected cell and become a part of the entire genome. These *ERVs* are inherited according to Mendelian expectations in the same way as all other genes in the genome. The phenomenon of *ERV* inheritance is called vertical contagion. The first exogenous retroviruses that infected a germ cell and became endogenous could have appeared at any time over an extended evolutionary time-scale between 2 to 70 million years ago. Even a horisontal contagion is possible, from one species to another through exogenous retrovirus infection. A possible way of horisontal contagion is when animals of different species are forced to live in a limited area. On average it has been estimated that it takes one million year until retroviral integration is fixed within a population.
All veterbrates have endogenous retroviruses in their genomes, the *ERV* elements are mostly old, truncated and non-functional. Still some of them have open reading frames in RV gene are low in *LTR* divergence and in frequency of nonsense mutation causing premature stop of translation and frame shift mutations. The canine genome has been effective in protection from extensive retroviruses integration, the amount in dog is expected to be a fifth of the real *ERV* amount in species like human and chimpanzees.

Background

The aim of this project was to perform *in silico* analysis to define the abundance and complexity of endogenous retroviruses in the dog genome. The latest available version of the dog genome CanFam2crt10v070313 (CF2c was mined). For collecting the retrovirus chains in this study, RetroTector© was the main tool and a limit was set for elements at > 300 score. Copies and *L1*-like elements were sorted out and excluded from further analysis. All retrieved sequences contained a *pol* (polymerase) gene. The *pol* gene and its protein/putein sequence are necessary, for alignment within the *Canis Familiaris* and between *Canis* and other species. Phylogenetic studies have been done with the following species: canine (*Canis Familiaris*), human *(H. s. Sapiens)*, mouse (*Mus musculus)*, reptile, (*Python molurus* AF500296 and *Crocodylus niloticus* AJ438130), fish *(Walleye dermal sarcoma virus pol NC 001867* and *Snakehead RV pol NC 001724).*
The details regarding the complexity of the Canine Endogenous retroviruses are yet unpublished and the scientific literature do not often mention endogenous retroviruses in dog. The amount of endogenous retroviruses in human (inclusive retrotransposons and single *LTR*) is 7 % (Bock and Stoye, 2000), 7-8 % (Jern, 2005). The canine and the chicken genomes have both being effective in protecting themselves from large amounts (< 0, 2 %) of retroviruses (Blomberg unpublished).

Materials and methods

RetroTector©, RetroTector utilities, Corel DRAW, CanFam2crt10v070313 (CF2crt), Clustal W, science literature and reference sequences (1) supplied by Jonas Blomberg. RetroTector©, is invented by Göran O. Sperber and Jonas Blomberg, at the Section of Virology, Department of Medical Sciences, Uppsala University, Uppsala, Sweden, Department of Neuroscience, Uppsala University, Uppsala, Sweden. RetroTector© is the software that among other things, retrieve complete or fragmented endogenous retroviruses (*ERVs*) chains/elements, the figures for chain details, the putein sequences (with ID and genus) for

classification, score, gene ID, the Open Reading Frames, ORF (*LTR* divergence, the stop codons and shifts) and Rep Base Finds (Sperber et al., 2007) RetroTector© is a program package written in Java. It is in use under Windows, MacOS X and Linux and designed to identify ERVs in genomic DNA sequences. It relies on a database of retroviral motifs and alignments of retroviral proteins. RetroTector© recognizes consensus motif and constructs ERV proteins from the different reading frames (Jern et al, 2005). RetroTector© shell is the software for data mining, and alignment of sequences. Together with Corel DRAW, this data was used to create the phylogenetic trees. The genomic material: CanFam, is the complete genome for *Canis Familiaris* (CF2crt) and the genomic source for this study. Data collections from CLUSTAL W (1.83) are used for multiple sequence alignment. *ERV* elements and the pol putein sequences belonging to the respective *ERV* elements were retrieved by RetroTector©. The limit for element to be further analysed was set to > 300 score (RetroTector©). All lower scored copies (two or more elements) were sorted out and excluded from further analysis, first in case of identical chains position for chains and second in case of identical estimated first or last pol putein position for pol putein sequences. The *L1*-like elements (see Fig 1) were sorted out as all these elements did not include any *pol* gene. Retrovirus elements were selected based on pol putein basis in all the following data mining of canine endogenous retroviruses described here.

Endogenous retroviruses (*Retroviridae)*

The whole animal kingdom and even plants have *ERVs* in their genomes (Jern et al., 2005). Transposable elements (TE) are parts of the entire genome and a part of TE is *ERV* (see Fig. 1.) *ERVs* in completely different species have the same basic genomic organisation. The first identification of *ERVs* was made in chicken, mice and cats. In mouse, ERVs were connected to diseases caused by retroviruses such as the Mouse Mammary Tumor virus (*MMTV*, Beta retrovirus) and Murine Leukemia Virus (*MLV*, Gamma retrovirus) (Reviewed by Blomberg et al., 2004). For more than 35 million years ago the *ERV*s entered the primate genome (Hughes et al., 2005) These exogenous retroviruses (*XRV*) and *ERV* were spread as an infection to other animals and certain *XRV*s became endogenized and transmitted vertically according to Mendelian expectations (Weiss et al., 2006).

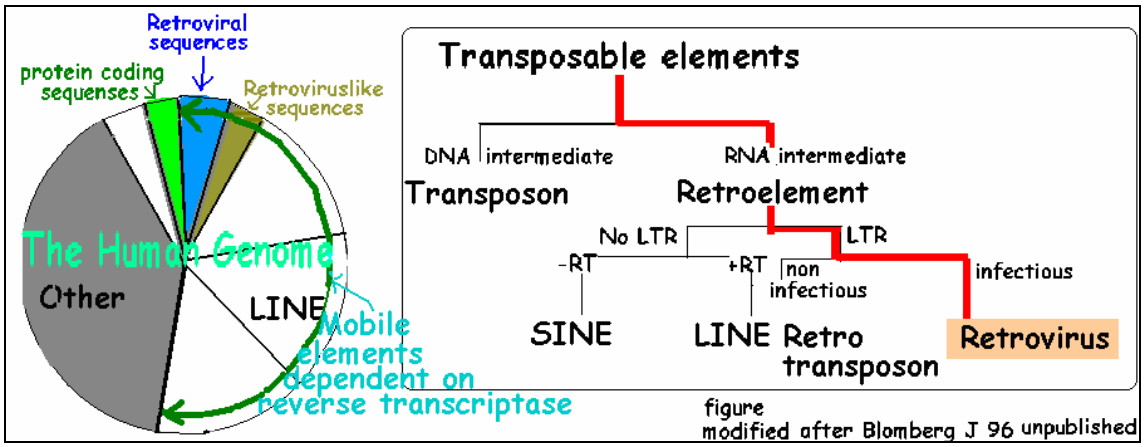


Fig 1. Transposable elements are part of the entire genome. Here the content in the human genome is shown and what split LINE (L1) and retrovirus almong all transposable elements. Modified figure from orginal made by Jonas Blomberg 1996.

The structural organization of a complete endogenous retrovirus
The first nucleotide at the 5' end of a retroviral mRNA is denoted the *cap* position. Transcription of retroviral mRNA initiates in the R region of the 5' *LTR* (long terminal repeat), a repeated regulatory sequence located at each end of the proviral genome. *LTRs* could be identical or at various degree of divergence, according to the evolutionary age of integration in the host genome. *PBS* (Primer-binding Site) is in the 5' end of the RV mRNA located between *cap 0* (*i.e*. the start site of transcription), and *gag* gene. *PBS* bind tRNA as the primer that start position translation. The four next coding genes are basic for all *ERV* elements (Jern et al., 2004). First is a *gag* gene encoding the structural proteins: matrix, capsid and nucleocapsid. The second gene is the *pro* gene encoding protease. The third is a *pol* gene that encodes reverse transcriptase and integrase. The *pol* gene taking a unique position in all studies of endogenous retroviruses because it encodes the most conserved protein sequence and the possibility for alignment of *pol* genes from completely different classes and or species is amenable. The last gene for building up complete endogenous retroviruses is an *env* gene that encodes the structural proteins for surface and transmembrane proteins of the retroviral envelope. Occurrence of an *env* gene in *ERV* elements is required for the ability of a retrovirus to infect other cells of the host or cells from another individual. This was shown as an absolute requirement for infection of *XERV* in a recently published paper (Oja et al., 2007).

Fig 2.A schematic representation of a complete endogenous retrovirus genome and with its components indicated.

Results

RetroTector© denotes all retrieved chains and putein sequences unique ID numbers but some of them are copies of each other. The pol putein sequence similarities are the basis for all alignments that have been performed in the current project. This study includes only one unique sequence (chain) at every given location, first at the chain location and then in the pol putein location. Several pol putein sequences were retrieved for most of the chains, at the same location, but only one was accepted. The *pol* ID with the highest ID number was also often the same *pol* ID with the highest average score and those sequences were chosen. A total of 254 *CFERVs* were identified and every calculation and analysis were based on these 254 selected endogenous retrovirus sequences from RetroTector©, used Canfam 2crt10v070313 by RetroTector©. The average *CFERV* is 7 kb long and the amount of 254 *CFERV* (1793076 bp) is 0, 075 % (based on the 254 chains real lengths) in the dog genome that consist of 2 384 996 543 bp.

Classification of *CFERV*

The classification for *ERV* is based on *Pol* nucleotide sequence similarity and Pol protein conservation (Jern et al., 2005). As indicated above, the Pol protein is the most well-conserved retrovirus protein. The sequence of *pol* encodes a polypeptide of 800–1100 aa (Jern et al., 2005).

Score

Depending on the degree of fulfilment of the chains, the hit is assigned a score from RetroTector©. Selection criteria based on scores > 300. There were 254 unique (based on *pol*. position) Canis Familiaris endogenous retrovirus elements with scores > 300 detected by RetroTector©. The chains were ordered in two groups, one group with scores > 1000 and in the other group the chains with scores ranging between 300 and 1000. The number of *CFERV* elements with scores over 1000 (chains form 1008 to 1771) is 49 (19 %). The group with chains less than 1000 score consists of chains from 301 to 992 score and the number of chains here is 205.

Fig3. Almost 4/5 of CF endogenous retrovirus consist of element with scores between 300 – 1000 (301-992, RetroTector©) and approximately a fifth of the elements are > 1000 (1008-1777) score.

Chain ID 1205 C (RetroTector©) on chomosome X, is the highest scored (1771), *CFERV* of all chains detected. This gammaretrovirus chain meets the necessary requirements for a complete *ERV* (see Fig. 2), as it contains two *LTR*s and all four (*gag, pro, pol* and *env*) important genes. Rep Base Find is: HERVR 32889510029174, the length is 8863 bp and the corresponding *pol* gene is ID 2589 c with stop codons and shifts: 5 and 12. The chain's *LTR* divergence is 7.73 %.

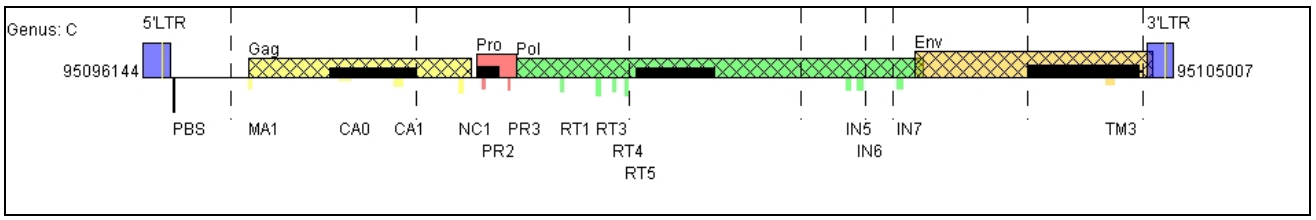


Fig 4. ID number 1205C shown in details. (RetroTector©). The picture of ID 1205 shows a high similarity to the schematic ideal figure of an endogenous retrovirus (Fig 2).

The lowest allowed (305) scored *CFERV* is chain ID 397 B, (see Fig. 3), located on chromosome 10 (RetroTector©), is a betaretrovirus (length 9698 bp) with two *LTRs*, a *pol* gene and an *env* gene. Rep Base Find is: HERVFH21 44 42534399 50, the *pol* gene in this chain is ID 796 c with stop codons and shifts: 11 and 9. The chains *LTR* divergence is 21.38%, the divergence is almost three times higher than ID 1205 (7.73%).

Fig 5. ID number 397B shown in details. (RetroTector©)

One of the lowest scored chains id 136 with a score of 256 that fell under the limit but the interesting part is that the length of this chain, it is 12 kb long and it is only consisting of an *env* gene flanked by two LTRs (see Fig. 6). Id 136 was not selected because lack of the *pol* gene and a low score. As shown in Fig. 4 and Fig. 5, the estimated length of this *CFERV* is similar to high scoring *CFERV*s. This suggests that this proviral genome may contain as yet undefined *gag* and *pol* genes. Further bioinformatic analyzes of these proviral sequences may provide such information.

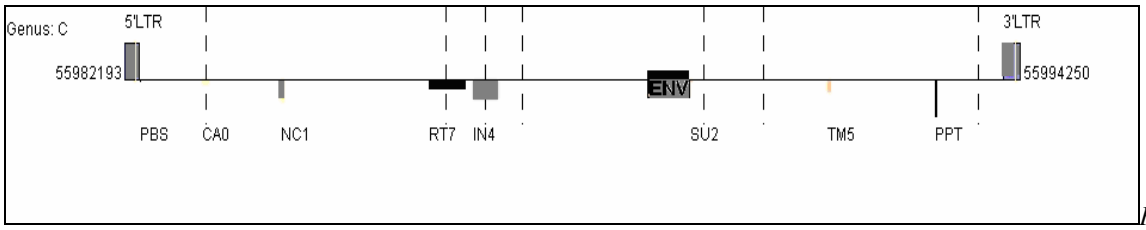


Fig 6. A low scored (256, under the limit) chain is id 136C contain only an env gene, still there is space for all components because the chain is 12 kb long. (RetroTector©)

Genus

Retroviruses are classified in three classes and in nine genus *A, B, C, D, E, L, S, G,* and *O*. The 254 dog endogenous retrovirus chains detected by RetroTector©, were assigned to their different retroviral genera. *CFERV* elements appear in three genera i.e. *Beta*, *Gamma*, and *Spuma* as follows 14 (5, 5 %) *Beta* (*B*), 237 (93.3 %) *Gamma* (*C*), and 1 (0, 38 %) *Spuma* (*S*) genus. *Gamma* and *Beta* are the two main groups as in human (hg16) (Jern et al., 2005). In addition, two chains were classified as *Delta*-like. However, further Phylogenetic analyses are required to corroborate this classification (see Figure 7).

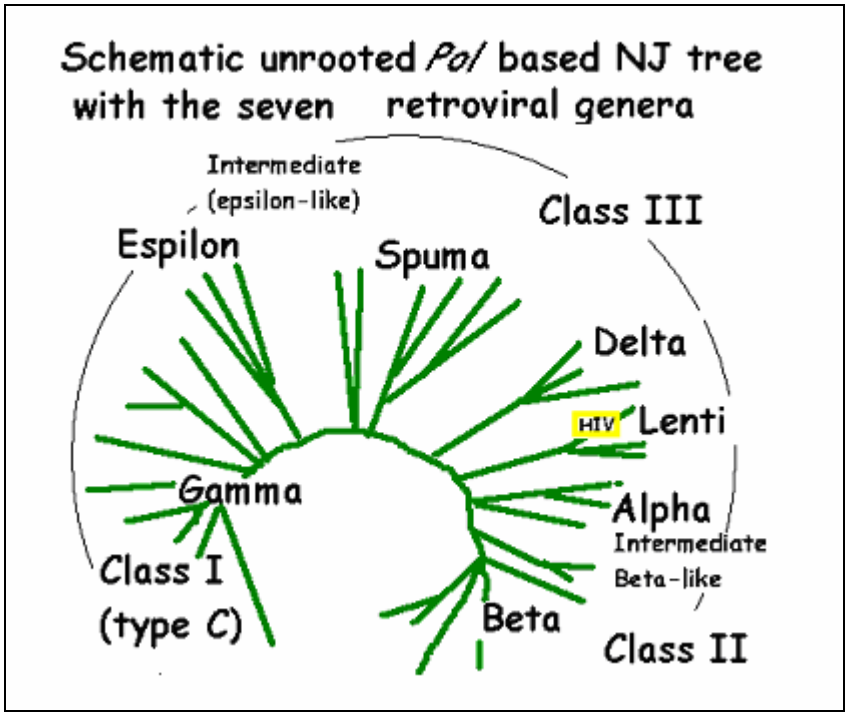


Fig 7. A schematic picture of the most important clusters and classes for ERV

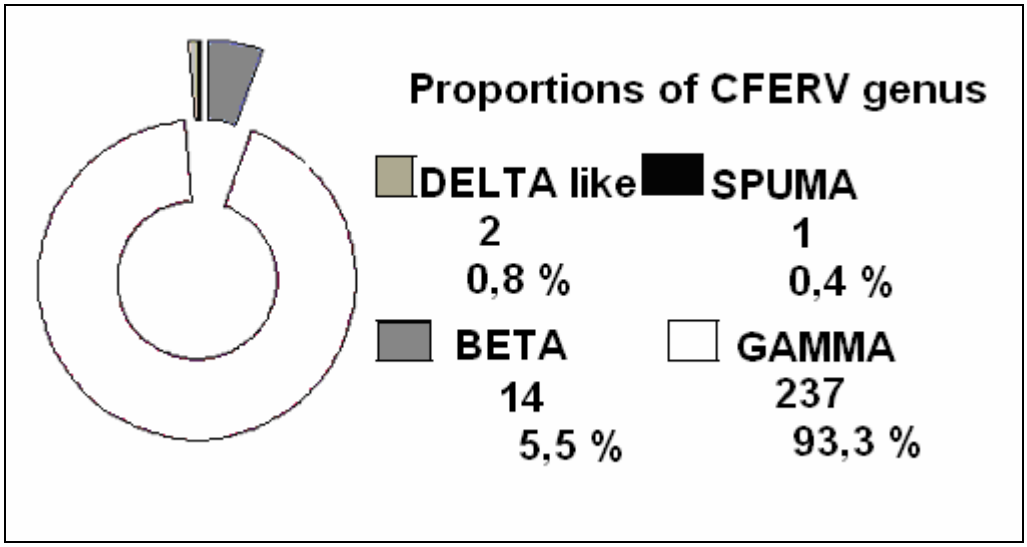


Fig 8. Proportion of chain genus' shows the main group of C element in the CF retroviruses.

Completeness for the CFERVs

For classification as a complete endogenous retrovirus, the *CFERV* elements need all four genes, *gag, pro*, *pol* and *env*.The number of elements that have just one (*pol*) gene is 48 (18.9%), 65 (25.6 %) elements have two genes, 96 (37.8 %) elements have three genes and 45 (17.8 %) elements have all four genes (see Fig 9).

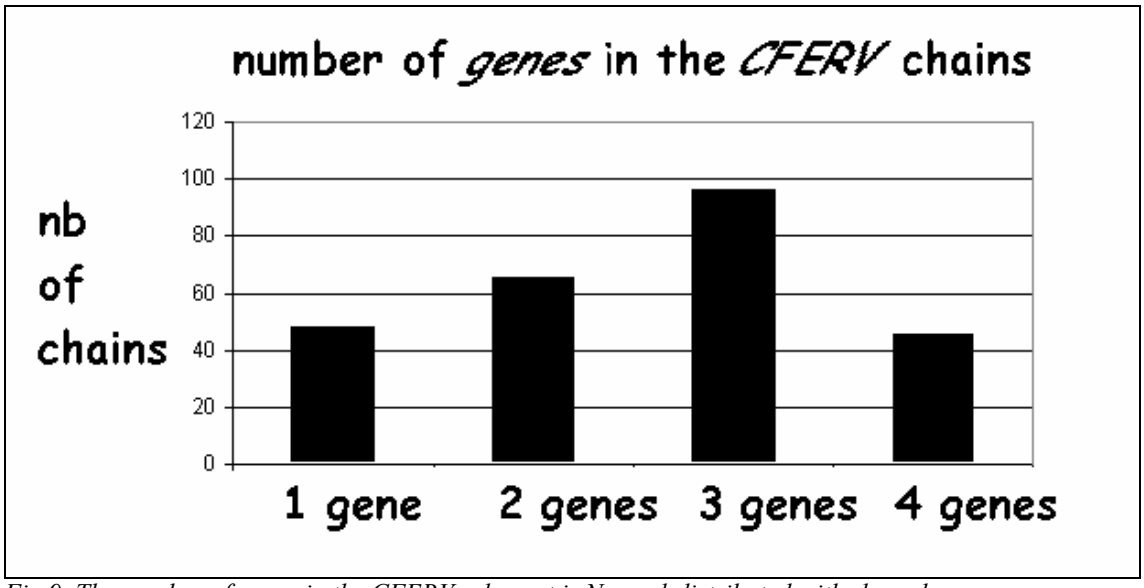


Fig 9. The number of genes in the CFERVs element is Normal-distributed with skewedness.

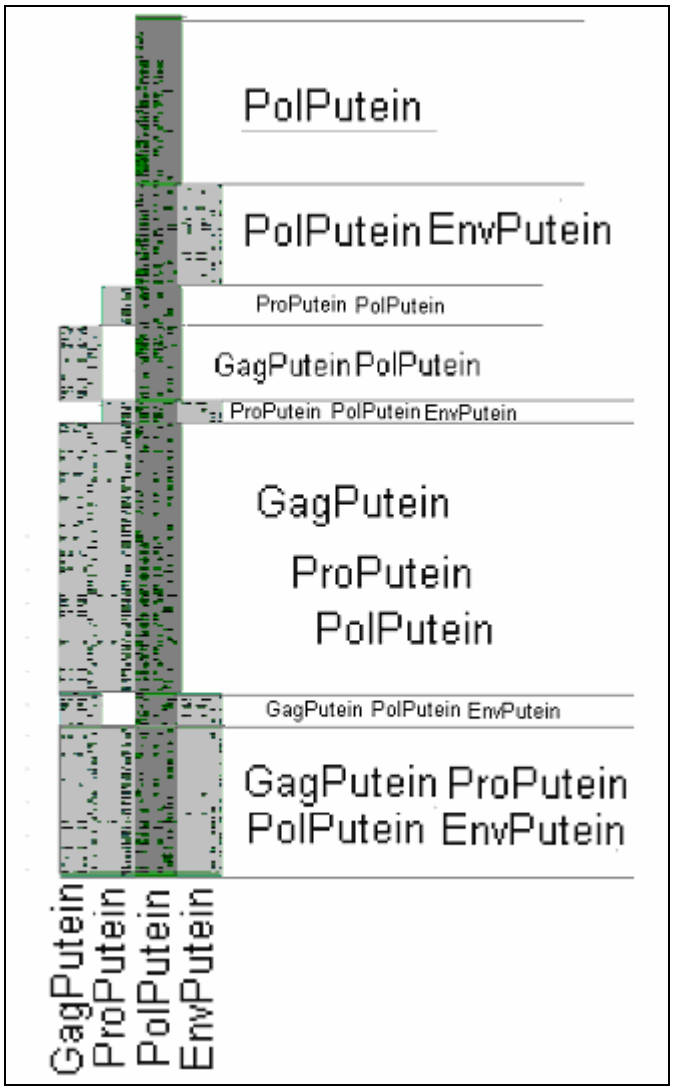


Fig 10. The drawing make visible a tower of all the 254 dog endogenous retrovirus elements, (all contains a pol gene) the most complete with all four genes take place in the bottom of the building.

Average score within the groups, defined based on different numbers of genes in the chains

All selected chains (except id 136, see above) in this study contain a *pol* gene. The *pol* gene is either the exclusive gene in the element or it could be present in combination with either *gag* or *pro* and *env* genes. Chains containing all four genes have the highest score and the chains with only the *pol* gene have the lowest score.

Table 1 Amount of chains in different groups of gene (*gag, pro, pol* and *env*) combinations and their average score.

Chain nb	Chain genes		Average score
48	*Pol*	putein	401,25
23	*Gag Pol*	putein	439,4348
30	*Pol Env*	putein	483,1333
12	*Pro Pol*	putein	559,1667
80	*Gag Pro Pol*	putein	805,875
6	*Pro Pol Env*	putein	838,3333
10	*Gag Pol Env*	putein	914,2
45	*Gag Pro Pol Env*	putein	1023,289
254	Tot		683,0853

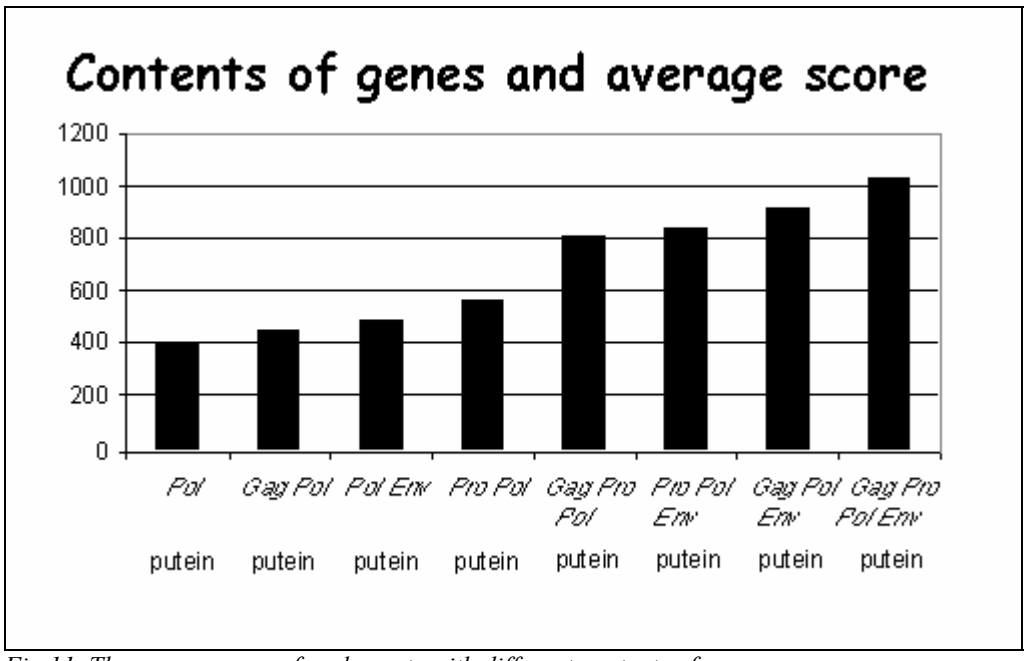
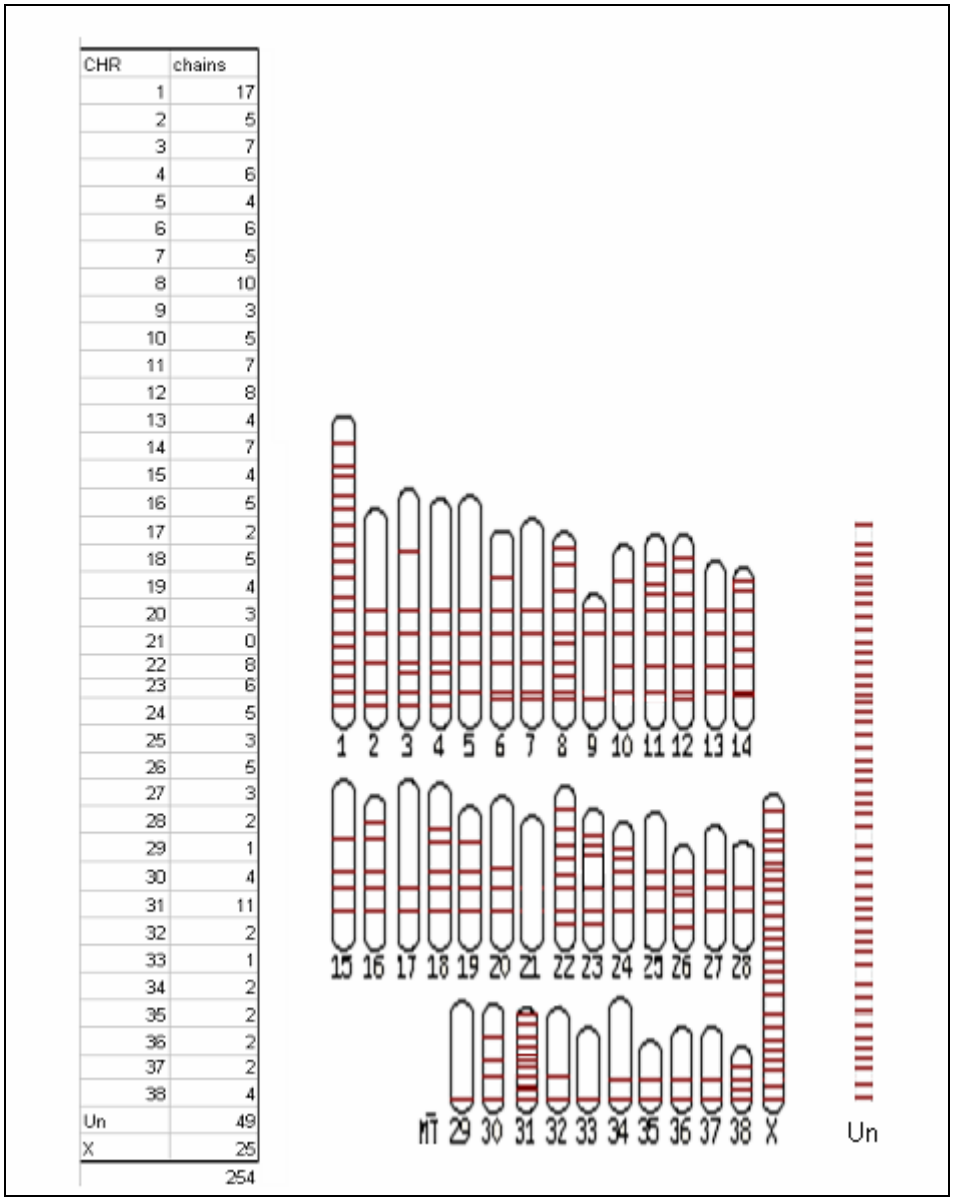


Fig 11. The average score for elements with different contents of genes.

Chromosomal distribution

The distribution of canine endogenous retrovirus elements between the dog's 38 chromosomes is unequal and 49 of the 254 *CFERV* chains (19.3%) retrieved from RetroTector are located on contigs with unknown chromosomal localization (see Fig. 12). The largest amount of endogenous retrovirus was identified on chromosome X, as expected which, have 25 elements (Fig. 12). The autosomes with the greatest amount of *CFERV* element are the chromosomes 8 and 31 that have 10 and 11 elements, respectively. The chromosome with the lowest occurrence of endogenous retroviruses is chromosome 21 which completely lacked *CFERV* and chromosomes 29 and 33 which have only one *CFERV* element each. It remains to be confirmed that these chromosomes are essentially lacking *CFERV*s or whether it reflects annotation-bias of these chromosomes.



CHR	chains
1	17
2	5
3	7
4	6
5	4
6	6
7	5
8	10
9	3
10	5
11	7
12	8
13	4
14	7
15	4
16	5
17	2
18	5
19	4
20	3
21	0
22	8
23	6
24	5
25	3
26	5
27	3
28	2
29	1
30	4
31	11
32	2
33	1
34	2
35	2
36	2
37	2
38	4
Un	49
X	25
	254

Fig 12. Distribution of endogenous retroviruses per chromosome and a schematic picture of CFERVs on the 38 autosomal chromosomes, the X chromosome and the CFERVs located on currently unknown chromosomes.

Rep Base Finds

The majority 115 elements (65 %) of the dogs' endogenous retroviruses, show similarity to *HERV* (Human Endogenous retrovirus). Of these elements, 18 (10 %) can be classified as *HERVF21*-like elements. The remaining 55 % of *CFERV* with similarity to *HERV*s is currently only classified as *HERV* –like. Further bioinformatic analyses will be performed to classify these elements. In figure 14, the close relation between dog CFERVs and the human *HERV*s is shown. *HERVF* is binding tRNA (in *PBS*) with phenylanalin (TTC) in the initiation of translation and is found to be expressed in human placenta and in human cancer cells (Kjellman et al., 1999). *HERVF* is a recently discovered type of *ERV* in human and Old World Primates (Kjellman et al., 1999), with an estimated that integration occurred more than 60 million years ago.

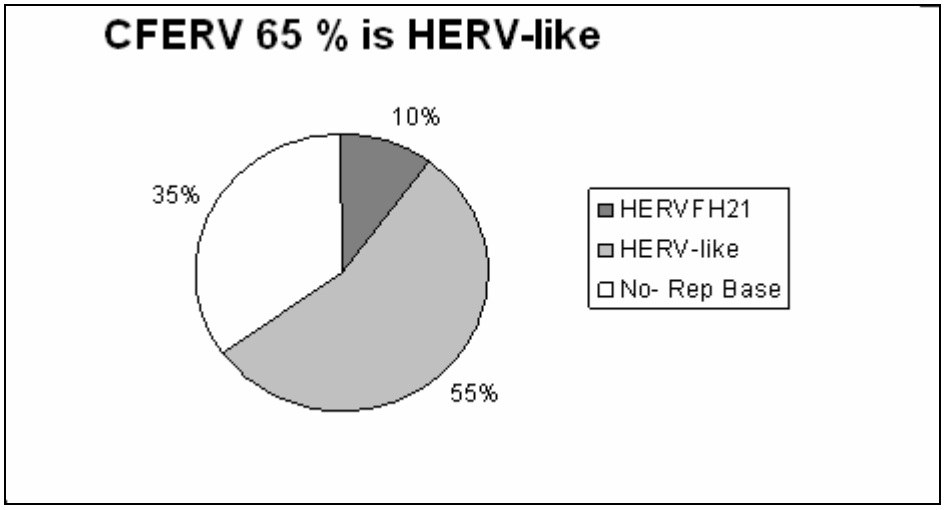


Fig 13. Approximately half (55 %) of the dogs ERVs is HERV-like and added to that 10 % that is HERVFH21-like, little more than 1/3 of the dogs ERVs have no Rep Base Finds.

Comparison between human *HERVF* and *HERVF*-like *CFERV*

A number of 18 *CFERVs* retrieved by Rep Base were found to be *HERVFH21*-like. Their *pol* sequences were aligned to human *HERVFH21 pol* sequences, (see unrooted cladogram in Fig.14 and Fig.15), and represent the similarity among the *HERVFH21 pol* sequences. Of these 18 *HERVFH21*-like *CFERV* elements, 14 were phylogenetically closer to the human *ERVFH21*. The 14 *HERVFH21*-like *CFERV* elements are remarkably highly scored (median 1240, 5) and all of them are as expected based on their similarity to *HERVFH21*, gammaretroviruses (Table 2).

Table 2, The 14 closest *HERVFH21*-like *CFERV* elements that show high score

CFERV element ID	Chain genus	score
1	c	1286
45	c	906
92	c	1232
151	c	1646
196	c	782
240	c	1542
289	c	595
301	c	1330
341	c	1249
415	c	1342
759	c	777
1093	c	695
1209	c	825
1168	c	1403

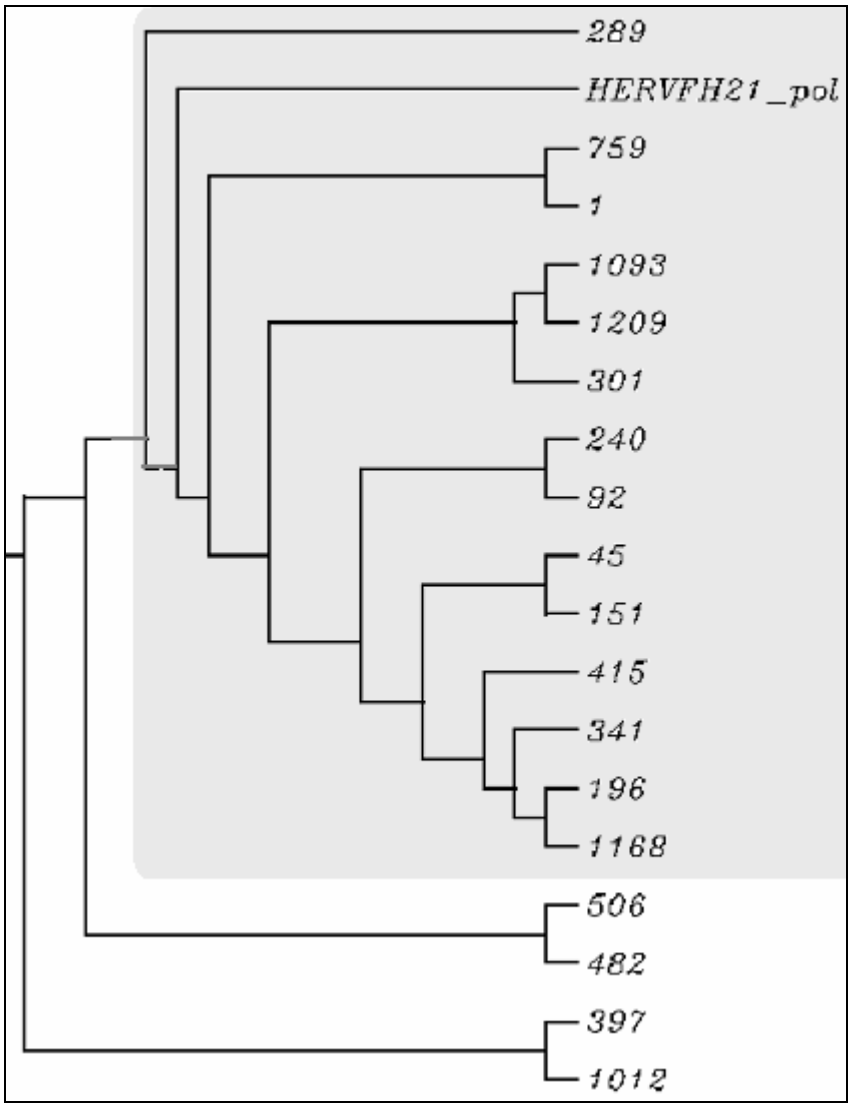


Fig 14. HERVFH21 pol sequence from human aligned to 18 HERVF-like pol sequences from dog in a cladogram. The grey area includes the closest related sequences. (Molecular Phylogenetic Tree by Neighbor-joining method CLUSTAL W (1.83))

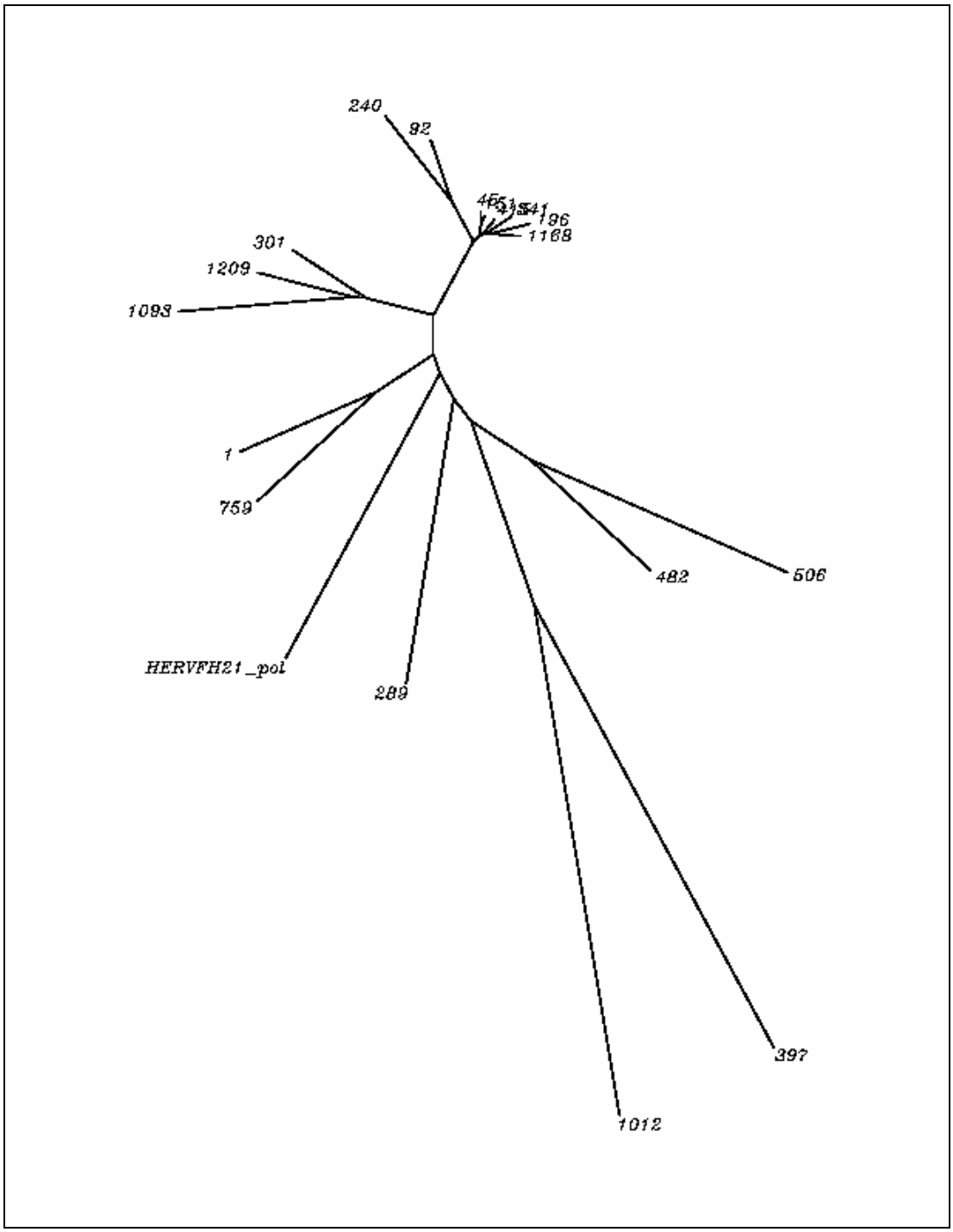


Fig 15. HERVFH21 pol sequences from human and 18 HERVF-like pol sequences from dog in an unrooted tree. (CLUSTAL W (1.83))

Class I gammaretroviruses (type C)

*CFERV*s belonging to the Gammaretroviruses represent the largest group of retroviruses in the dog genome. A similar situation is true for the human genome (Jern, 2005). There are 237 retrovirus elements of class I, gammaretroviruses with scores above 299 (RetroTector©). In Tables 3 to 6, only the chains with a chromosome location are listed.

Table 3 The gamma chains 1008-1771 score RetroTector©.

Gamma Chains –1008-1771 score

Score	select chain id	Chain Genus	C H R	Rep Base Finds	Stop Codons	Shifts	LTR Divergen	Env Puteinid	Pol Puteinid	Select pol id	Pro Puteinid	Gag Puteinid
1771	1205	C	X	HERVR	5	12	7.73196	2586	2589	2589	2590	2588, 2587
1761	539	C	15	HERVR	2	3	44.898		1115	1115	1116	1114, 1113
1667	118	C	3	HERVR	2	3	null		238	238	239	237
1646	151	C	3	HERVFH	0	3		300	302	302	303	301
1637	167	C	4	HERVR	3	3	46.875	351	353	353	354	352
1617	127	C	3	HERVR	5	3	4.92753	263	266, 265	266	267	264
1552	455	C	12	HERVR	2	5	null		916, 915	916	917	914
1548	377	C	9	HERVR	3	2	null		762	762	763	761, 760
1547	39	C	1	HERVR	2	4	null		83	83	84	82
1542	240	C	6	HERVFH	1	2	0.895524	492	494	494	495	493
1498	432	C	11	HERVR	3	6	6.91144	860	863	863	864	862
1488	287	C	7	HERVR	3	6	null		574, 573	574	575	572
1467	463	C	12	HERVR	7	7	null		939, 938	939	940	937
1454	211	C	5	HERVR	2	4	null		441	441	442	440, 439
1433	407	C	10	HERVR	4	7	6.10687	806	809, 808	809	810	807
1403	1168	C	X	HERVFH	0	3	0.925928	2488	2490	2490	2491	2489
1396	21	C	1	HERVR	4	11	11.0276		40, 39	40	41	38
1352	1081	C	36	HERVR	15	14	48.3516	2321	2324, 2323	2324	2325	2322
1351	293	C	8	HERVR	6	9	null		585	585	586	584
1342	415	C	11	HERVFH	0	1	null		827	827	828	826
1341	1158	C	X	HERVR	4	7	null		2459, 2458	2459	2460	2457
1330	301	C	8	HERVFH	1	5	5.14139	607	610, 609	610	611	608
1328	823	C	24	HERVR	5	9			1733, 1732	1733	1734	1731
1320	642	C	18	HERVR	4	8	null	1325	1328, 1327	1328	1329	1326
1313	171	C	4	HERVR	9	4	6.58228	358	360	360	361	359
1299	760	C	22	HERVR	7	7	2.53165	1573	1576, 1575	1576	1577	1574
1286	1	C	1	HERVFH	2	8	7.44681	1	4, 3	4	5	2
1258	1060	C	34	HERVR	5	13	null		2266, 2265	2266	2267	2264
1249	341	C	8	HERVFH	0	3	6.83761	693	697, 696	697	698	695, 694
1243	1173	C	X	HERVR	0	2	43.5897	2503	2506	2506	2507	2505, 2504
1241	1184	C	X	HERVR	3	6	null		2536, 2535	2536	2537	2534
1239	690	C	20		11	6	41.9048	1428	1430	1430	1431	1429
1232	92	C	2	HERVFH	0	2	null	185	187	187	188	186
1227	1139	C	X	HERVR	10	7	8.61539		2422, 2421	2422	2423	2420
1218	1181	C	X	HERVR	3	7	43.4343	2528	2530	2530	2531	2529
1217	1013	C	31	HERVR	2	7	44.4444	2145	2151	2151	2152	2150
1205	347	C	9	HERVR	5	12	null	712	716, 715	716	717	714, 713
1193	1196	C	X	HERVR	1	11	null		2564, 2563	2564	2565	2562
1135	1100	C	37		13	9	15.6489		2350, 2349	2350	2351	2348
1124	529	C	15	HERVR	6	4	null		1090	1090		1089
1097	821	C	24	HERVR	0	7	null		1725	1725	1726	1724
1090	883	C	26	HERVR	8	6	33.2288		1852, 1851	1852	1853	1850, 1849
1085	385	C	10	HERVR	8	4	9.19118	774	776	776	777	775
1036	175	C	4		8	7	null		372	372	373	371
1020	571	C	16	HERVR	4	8	5.95238		1178, 1177	1178	1179	1176, 1175
1015	517	C	14	HERVR	8	4	null		1071, 1070	1071	1072	1069, 1068
1008	1149	C	X	HERVR	0	8	43.5644	2436	2439, 2438	2439	2440	2437

Table 4 The gamma chains 663-992 score RetroTector©.

Gamma Chains – 663-992 score

Score	select chain id	Chain Genus	CHR	Rep Base Finds	Stop Codons	Shifts	LTR Divergen	Env Puteinid	Pol Puteinid	Select pol id	Pro Puteinid	Gag Puteinid
992	443	C	11	HERVR	5	7	null	890	894, 893	894		892, 891
981	33	C	1	HERVR	8	6			73, 72	73		71
971	609	C	17	HERVR	11	14	null		1250	1250	1251	1249
971	667	C	19	HERVR	4	2	null		1380	1380	1381	1379, 1378
938	1073	C	35	HERVR	4	8	3.69231	2309	2311, 2310	2311		
927	56	C	1	HERVR	3	3	null		126, 125	126	127	124
926	374	C	9	HERVR	2	3	null		755	755	756	754
915	745	C	22	HERVR	3	5	13.9073	1531	1533, 1532	1533	1534	
906	45	C	1	HERVFH	0	3	null		104	104	105	103, 102
903	178	C	4	HERVR	0	0	null		379	379	380	378
898	436	C	11	HERVR	1	4	44.8	873	876, 875	876	877	874
892	990	C	31	HERVR	6	5	null		2090, 2089	2090	2091	2088
891	1197	C	X	HERVR	10	8	null		2567	2567	2568	2566
884	887	C	26	HERVR	9	7	null		1870, 1869	1870	1871	1868, 1867
873	788	C	23	HERVR	3	5	null		1654	1654	1655	1653, 1652
859	459	C	12	HERVR	4	3	8.05687	931	933	933	934	932
859	1104	C	38	HERVR	2	4	27.5862	2361	2364, 2363	2364	2365	2362
852	1157	C	X	HERVR	3	5	null		2456, 2455	2456		2454
843	978	C	30	HERVR	7	14	null		2062	2062	2063	2061, 2060
840	416	C	11	HERVR	1	4	null	830, 829	833, 832	833	834	831
835	72	C	2	HERVR	5	9	null		151	151	152	150
828	14	C	1	HERVR	13	7	null		33	33	34	32
826	1019	C	31	HERVR	13	14	null		2167	2167	2168	2166
825	1209	C		HERVFH	1	6	null	2600	2604, 2603	2604	2605	2602, 2601
804	308	C	8	HERVR	2	2	null	620	622	622	623	621
804	824	C	24	HERVR	14	9	null		1737, 1736	1737	1738	1735
802	703	C	20	HERVR	3	2	null		1452, 1451	1452	1453	1450
800	457	C	12	HERVR	3	3	null		929, 928	929	930	927
798	322	C	8	HERVR	7	9	null		654, 653	654	655	652
782	196	C	5	HERVFH	2	2	5.32787	415, 414	416	416		
780	1002	C	31	HERVR	19	10	null		2121	2121	2122	2120
779	1198	C	X	HERVR	9	6	null		2572, 2571	2572	2573	2570, 2569
777	759	C	22	HERVFH	5	10	38.1356	1569	1571, 1570	1571	1572	
751	312	C	8	HERVR	18	14	null		629, 628	629	630	627, 626
750	1114	C	38	HERVR	12	13	14.8305		2387, 2386	2387	2388	2385
749	201	C	5	HERVR	2	6	null		428, 427	428	429	426
743	767	C	23	HERVR	15	12	null		1608, 1607	1608	1609	1606, 1605
741	1187	C	X	HERVR	12	8	null		2546	2546	2547	2545, 2544
737	1122	C	X	HERVR	6	9	null		2404, 2403	2404	2405	
727	502	C	14	HERVR	20	12	null		1023, 1022	1022	1024	1021
727	1159	C	X	HERVR	9	7	null	2462, 2461	2464, 2463	2464		
711	677	C	19	HERVR	6	9	null		1403, 1402	1403	1404	1401
698	147	C	3	HERVR	2	2	null		294	294	295	293
695	1093	C	37	HERVFH	4	6	null		2339, 2338	2339	2340	
691	199	C	5		7	13	null		421, 420	421	422	419
690	26	CD	1	HERVR	2	6	null		56, 55, 60, 5	60	61, 57	54, 53, 58
678	591	C	16	HERVR	11	8	null	1213	1217, 1216	1217	1218	1215, 1214
663	988	C	30		4	4	10.8808	2078	2081, 2080	2081	2082	2079

Table 5 The gamma chains 422-348 score RetroTector©.

Gamma Chains – 422-648 score

Score	select chain id	Chain Genus	C H R	Rep Base Finds	Stop Codons	Shifts	LTR Divergen	Env Puteinid	Pol Puteinid	Select pol id	Pro Puteinid	Gag Puteinid
648	230	C	6	HERVR	6	11	null	472	474	474	475	473
647	3	C	1		12	14	null		8, 7	8	9	
617	513	C	14	HERVR	10	8	null		1060, 1059	1060	1061	1058
616	6	C	1	HERVR	8	10	null		15, 14	15	16	
616	152	C	3		19	17	11.7371	304	308, 307	308	309	306, 305
615	1107	C	38		7	5	14.0984	2371	2374	2374	2375	2373, 2372
611	63	C	1	HERVR	7	7	9.02256		141, 140	141	142	139
609	616	C	17	HERVR	4	5	10.9091	1264	1266	1266	1267	1265
607	9	C	1	HERVR	4	6	8.70967	21	23, 22	23		
595	289	C	7	HERVFH	8	11	null		580, 579	580	581	578, 577
593	804	C	24		9	12	7.40741	1699	1703, 1702	1703	1704	1701, 1700
585	224	C	6	HERVR	5	10	5.41872	464	468, 467	468		
584	232	C	6		1	5	16.4179	478, 477	481	481		480, 479
581	884	C	26		2	10	0.150377		1856	1856		1855, 1854
567	651	C	18		2	7	11.7021	1343, 1342	1345	1345	1346	1344
566	1118	C	X	HERVR	10	12	null		2399, 2398	2399	2400	2397
563	935	C	28	HERVR	3	2	null		1973	1973	1974	1972
557	337	C	8	HERVR	3	13	null		686, 685	686	687	684
547	69	C	1	HERVR	6	1	null		145	145	146	
546	761	CD	22		5	8	null	1579, 1578,	1584, 1581,	1584		
542	568	C	16	HERVR	5	9	null		1170, 1169	1170	1171	1168
534	482	C	13	HERVFH	2	3	20.3488	987	990, 989	990		988
532	471	C	12	HERVR	6	9	null	956	958, 957	958		
524	966	C	30		1	4	3.9801	2039	2042, 2041	2042		2040
523	404	C	10	HERVR	2	4	null		803	803	804	802, 801
514	827	C	24	HERVR	11	8	null		1748, 1747	1748	1749	1746
509	165	C	4		22	7	41.9643	336	339, 338	338	340	337
509	1165	CD		HERVR	6	6	11.4458		2483, 2482,	2483		2478, 2477, 248
503	514	C	14	HERVR	3	2	null		1064	1064		1063, 1062
496	584	C	16	HERVR	13	16	47.3118	1202	1205, 1204	1205	1206	1203
494	630	C	18		3	5	null		1299, 1298	1299		
491	848	C	25		13	8	18.8235	1793, 1792	1796, 1795	1796		1794
487	166	CD	4		8	11	null		344, 343, 34	344	345, 350	342, 341, 347,
486	315	C	8	HERVR	6	8	null	635, 634	638	638	639	637, 636
469	781	C	23		20	13	null		1637, 1636	1637	1638	1635
469	1036	C	33		10	9	6.12245	2218	2220, 2219	2220		
468	34	CD	1	HERVR	3	2	null		75, 77	75		74, 76
467	515	C	14	HERVR	8	13	null		1066, 1065	1066		
464	1044	C	34		14	9	41.7266	2239, 2238	2241, 2240	2241		
439	421	C	11		22	14	45.098		842, 841	842		
437	426	C	11	HERVR	6	5	null		852, 851	852	853	850
436	673	C	19	HERVR	17	14	null		1398, 1397	1398		
435	1000	C	31	HERVR	4	11	null		2116, 2115	2116		
432	103	C	2		11	14	null	213, 212	214	214		
428	1156	C	X		26	11	21.1009	2450	2453, 2452	2453		2451
428	506	CD	14	HERVFH21	9	3	null	1036, 1040	1039, 1042,	1038		1037, 1041
423	284	C	7	HERVR	3	10	null		568, 567	568		
422	769	C	23		5	15	22.9358	1614	1617	1617		1616, 1615
422	1031	CD	32		3	4	null		2206, 2202,	2206	2203, 2207	2200, 2205, 220

Table 6 The gamma chains300-414 score from RetroTector©.

Gamma Chains – 300-414 score

Score	select chain id	Chain Genus	C H R	Rep Base Finds	Stop Codons	Shifts	LTR Divergen	Env Puteinid	Pol Puteinid	Select pol id	Pro Puteinid	Gag Puteinid
414	910	C	27		19	13	48.8		1925, 1924	1925		
412	523	C	14	HERVR	5	6	null		1084, 1083	1084		
408	499	C	13	HERVR	4	3	null		1014	1014	1015	
408	1071	C	35		18	12	null	2304, 2303	2306, 2305	2306	2307	
407	532	C	15		13	9	49.4624	1099	1101, 1100	1101		
405	1162	CD	X	HERVR	2	4	null		2471, 2470,	2471		2469, 2468, 247
400	1207	C	X		17	14	16.8142	2593	2596, 2595	2596	2597	2594
397	286	C	7	HERVR	6	14	null		571, 570	571		
397	691	C	20	HERVR	1	2	null		1433, 1432	1433	1434	
392	995	C	31		10	9	null		2108, 2107	2108	2109	2106
391	1178	C	X	HERVR	10	11	null	2518, 2517	2520, 2519	2520		
390	553	C	15		12	14	null		1139, 1138	1139		
388	654	C	18		15	12	24.7525	1349, 1348	1351, 1350	1351		
385	20	C	1	HERVR	1	1	null		37	37		36
385	889	C	26		3	11	null		1878, 1877	1878		1876, 1875
382	340	C	8	HERVR	14	14	15.4412		692, 691	692		
368	51	C	1	HERVR	8	5	null		114	114		
368	911	C	27	HERVR	13	9	null		1927, 1926	1927		
366	424	C	11	HERVR	6	11	null		846, 845	846		
360	655	C	18	HERVR	6	9	null		1355, 1354	1355	1356	1353, 1352
358	501	C	13		14	16	18.4397	1018	1020, 1019	1020		
358	762	CD	22		5	8	null		1590, 1587,	1590		1585, 1588
351	834	C	25		7	16	15.0794	1763	1765, 1764	1765	1766	
347	242	CD	6		8	14	26.4706	497, 502	506, 501, 50	506		499, 498, 504,
339	578	C	16		18	11	null		1190, 1189	1190		1188, 1187
338	225	C	6	HERVR	4	10	null		466, 465	466		
337	1016	C	31		10	10	null		2159	2155		2158, 2157
332	320	C	8		23	15	null		649, 648	649		
329	926	C	28		19	17	null	1956	1958, 1957	1958		
329	755	CD	22		6	11	31.1475	1556	1563, 1562,	1563	1560, 1564	1558, 1557, 156
322	744	C	22	HERVR	2	9	null		1530, 1529	1530		
320	494	C	13		1	5	12.0219	1005, 1004	1006	1006		
320	837	C	25		13	11	7.52212	1775, 1774	1777, 1776	1777		
318	254	C	7	HERVR	3	7	null		522, 521	522		
317	123	C	3	HERVR 479	7	7	null		252, 251	252	253	250, 249

Score proportions in Gammaretroviruses

The number of *C* elements with scores above 1000 (RetroTector©) is 48 (20 %) and in the group 300 to 1000 score the number is 189 (80 %).

Completeness in CF gammaretroviruses

Only 44 (19 %) element are completes with all four genes (*gag, pro, pol* and *env*). The largest group of the gamma elements consists of 88 chains (37%) that contain three genes. The group that contains elements with two genes consists of 58 chains (24 %) and gammaretroviral

elements with a single *pol* gene are a group of 47 (20 %). Half of the chains with unknown chromosomal localization remain in the group with just one *pol* gene.

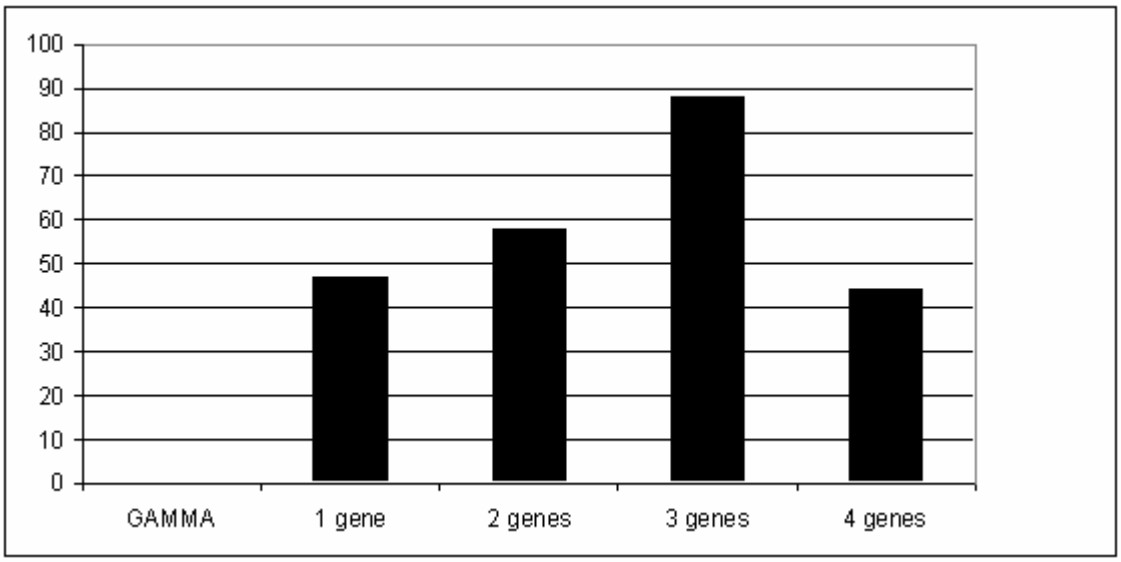


Fig 16. Division of groups of gammaretroviruses that contain from one, to four genes in their chains. The largest group contains three genes. The classes with one to four genes per element are almost normally distributed for gammaretrovirus.

Proportions of genes within the *CF* gammaretroviruses

Of all 190 C chains with a *pol* gene, 136 (72 %) have a *gag* gene, 126 (66 %) have a *pro* gene and 75 (39 %) have an *env* gene. For the number of 23 C chains score > 1000 there were 7 chains with 5' *LTR* and 3' *LTR*, 5 chains with just one 5' *LTR* and 11 chains (almost 50%) that completely lacks *LTR*s. It is possible that it exist chains where RetroTector© could detect neither the *env* genes nor *LTR*s.

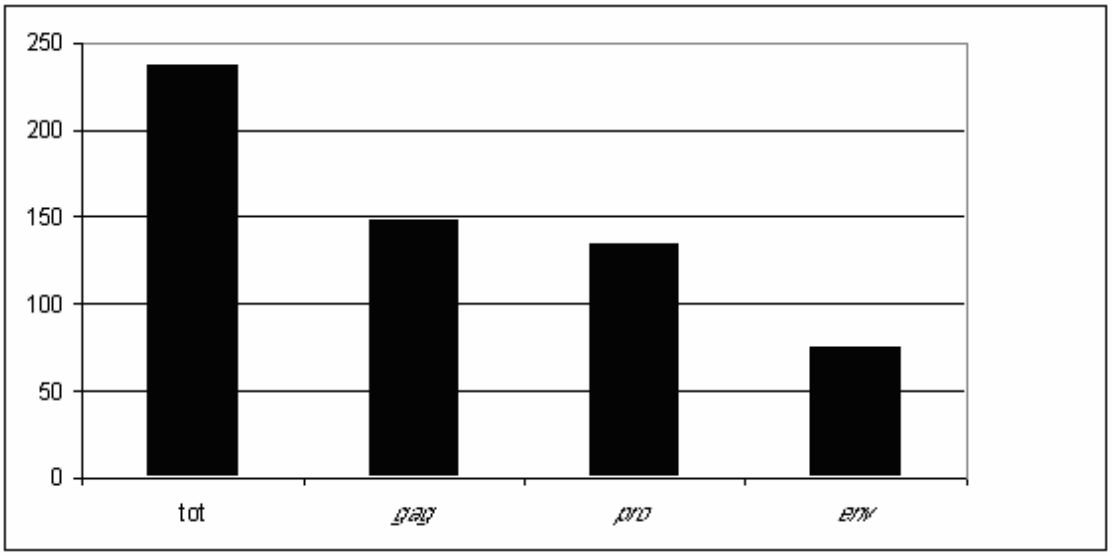


Fig 17. All CFERV C elements contain a pol gene. Gag (62,4 %), Pro (57 %) and Env (32%) genes occur in an apperent decreasing order (according to RetroTector© difficulty to find env and 3'LTR in CanFam2.0) .

Class II betaretroviruses, β and alpharetroviruses, α
alpharetroviruses are absent in dog

Betaretroviruses constitute a group of *CFERV* with only 14 elements (*6* % of all *CFERV* that were detected), Of these beta *CFERVs,* only one chain (id 478 in RetroTector©) remains from the high scoring group. No betaretrovirus in dogs seems to be apparently intact, none pass the score for group 1 *LTR* < 5 % and stop and shifts 0-3.

Table 7. Beta chains selected from RetroTector©

Beta Chains

Score	select chain id	Chain Genus	CHR	RepBase Finds	Stop Codons	Shifts	LTR Divergen	Env Puteinid	Pol Puteinid	Select pol id	Pro Puteinid	Gag Puteinid
1512	478	B	12		3	4	3.57143	974	976	976	978, 977	975
957	1014	B	31		3	9	null		2148, 2147	2148	2149	2146
745	991	B	31	HERVK3245 10336013 63;	14	8	null		2094, 2093	2094	2095	2092
735	1177	B	X		13	12	10.8108	2513	2515, 2514	2515	2516	
674	900	B	27		6	9	17.7778	1891	1893	1893	1894	1892
608	80	B	2		10	11	1.68539	165	167, 166	167		
572	1022	B	32		16	9	null		2173, 2172	2173	2174	
492	470	B	12		10	4	3.41464	953	955, 954	955		
430	953	B	29		13	7	null		2012	2012		2011, 2010
371	665	B	19		12	12	null	1372, 1371	1374, 1373	1374		
364	791	BC	23		14	10	24.6795	1668	1670, 1672, 166	1670		
351	1186	B	X		5	10	null	2542	2543	2543		
319	1012	BS	31	HERVFH21 41 33464058 54;	13	5	0.552487	2142	2144, 2143	2144		
305	397	B	10	HERVFH21 44 42534399 50;	11	9	21.3873	794	796, 795	796		

Proportions of genes and completeness within the *CF* Betaretroviruses
There are no beta chains with only one *pol* gene, which is a large distinction from the *C*-elements. The largest group, representing nine elements (64%) contain two genes, one *pol* and either one of *gag* or *env* genes. Three (21 %) chains contain three genes and only two (14%) elements contain the complete four genes.

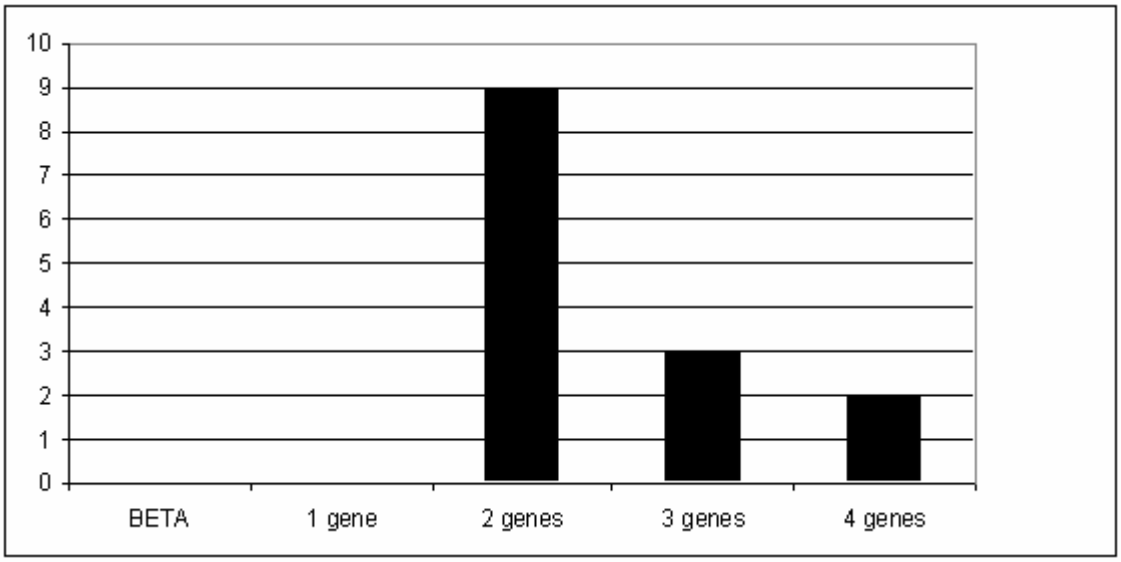


Fig 18. The largest betavirus group contains 2 genes one pol gene and one of the others.

Within the group of beta chains (with a *pol* gene) there are 10 (71 %) that also contain an *env* gene, and six (43 %) with a *pro* gene and five (36 %) elements that also contains a *gag* gene. Chain id number 478 (id number in RetroTector©,) is the only beta retrovirus, over 1000 score. The number of beta chains between 300 and 1000 score is 13.

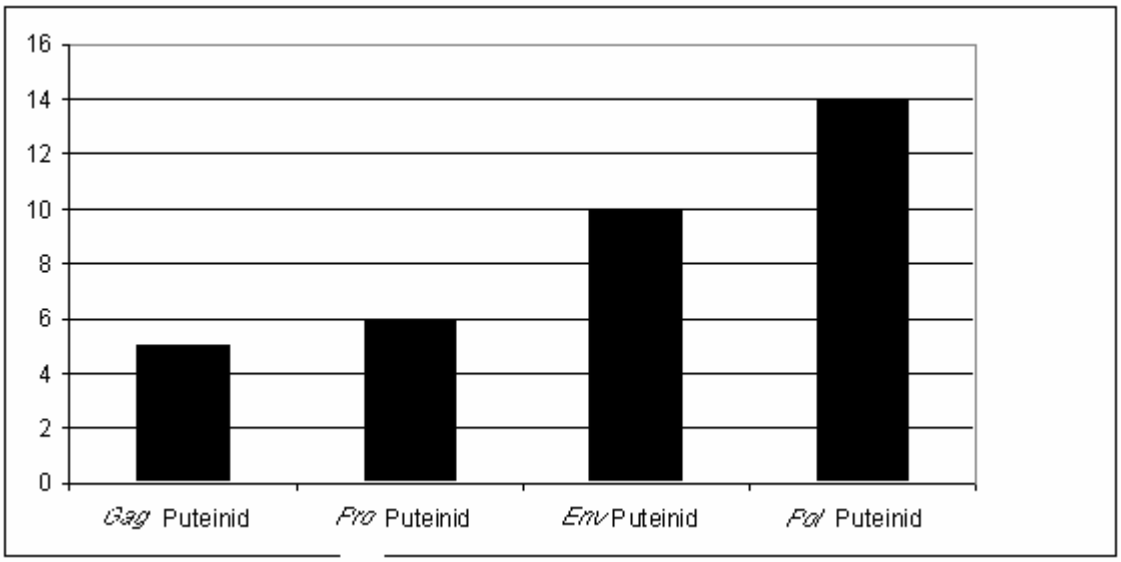


Fig19. All selected CF B element contain a pol gene. Env, Pro and Gag –genes occur in decreasing order

Class III Spuma ς retroviruses

Retrotector detected only one spuma- like *CFERV*-element with a *pol* gene in dog genome. Chain ID 95 (RetroTector©), CHR 2. The score is low (325) and it contains three (*Pro*, *Pol* and *Env*) of the four genes. The *LTR* divergence is large (20 %) and the stop and shifts are 25 and 12 respective, therefore, the *CFERV* spuma-like virus most likely constitutes the evolutionary oldest group of *CFERV*. The *CF* spuma-viruses were not closely related to other spuma viruses (see Fig 20-21.)

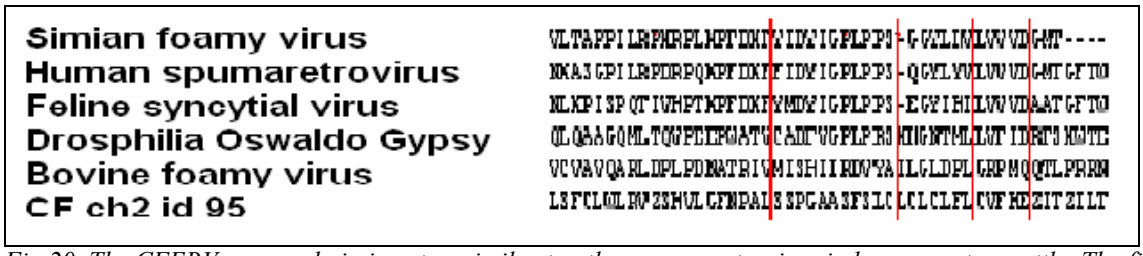


Fig 20. The CFERV spuma chain is not so similar to other spuma retrovirus in human, cat or cattle. The figure shown the best hits. The gypsy pol was included for rooting the alignment and the following NJ tree.

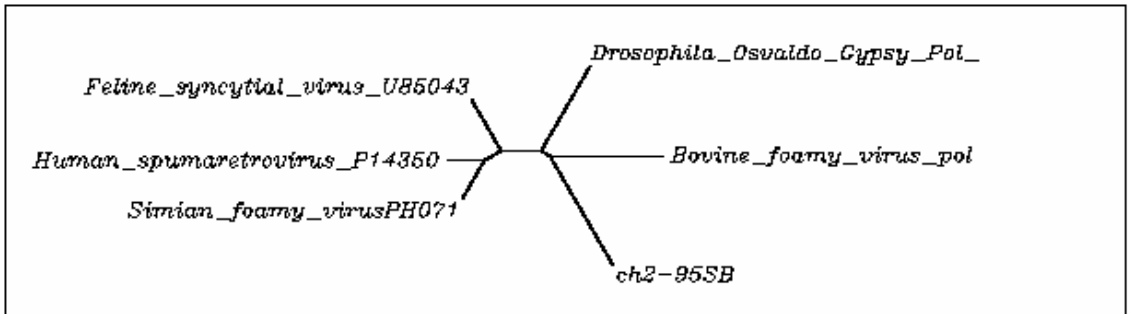


Fig 21. Spuma retrovirus in different species, cat, human. primates, cattle and dog. The gypsy chain was included as an outgroup.

Tab 2 CF spuma retrovirus from RetroTector©, the different id numbers in the genes represent different hits. Only one id for each gene was chosen with the highest average score at any given position.

Spuma Chains

Score	select chain id	Chain Genus	CHR	RepBaseFinds	Stop Codons	Shifts	LTRDivergence	Env Puteinid	Pol Puteinid	Select pol id	Pro Puteinid	Gag Puteinid
325	95	SB	2		25	12	20	191, 190, 194	196, 195, 192	196	193, 197	

CF Delta δ –like retroviruses

Two *CFERV* chains that are delta-like were retrieved. Both these two chains have low scores (371 and 467) and the elements lack *env* genes. Their ID numbers are 456 and 765 (RetroTector©), and they are located at chromosome 12 and 23. Some observations were done on RetroTector©, the chosen *pol* id 920 got the highest average score, that is more often seen for the highest *pol* id number (in this case 925) and for chain id 765 one last number was missed for two gene identification number (*pol* id 160X and *gag* id 159X). The 2 delta-like *CFERV* chains were marked D (Delta) C (Gamma) by RetroTector©, chain id 765 was C 0,9 and D 0,91. The chain has *pol* id C and even *pol* id D. In Fig 22 all 16 *CFERV pol* with genus d and one human delta-like *pol* sequence shows the relationship. All of the CFERV delta-like chains did not cluster and therefore the alignment did not confirm that all these chains belong to CFERV delta genus.

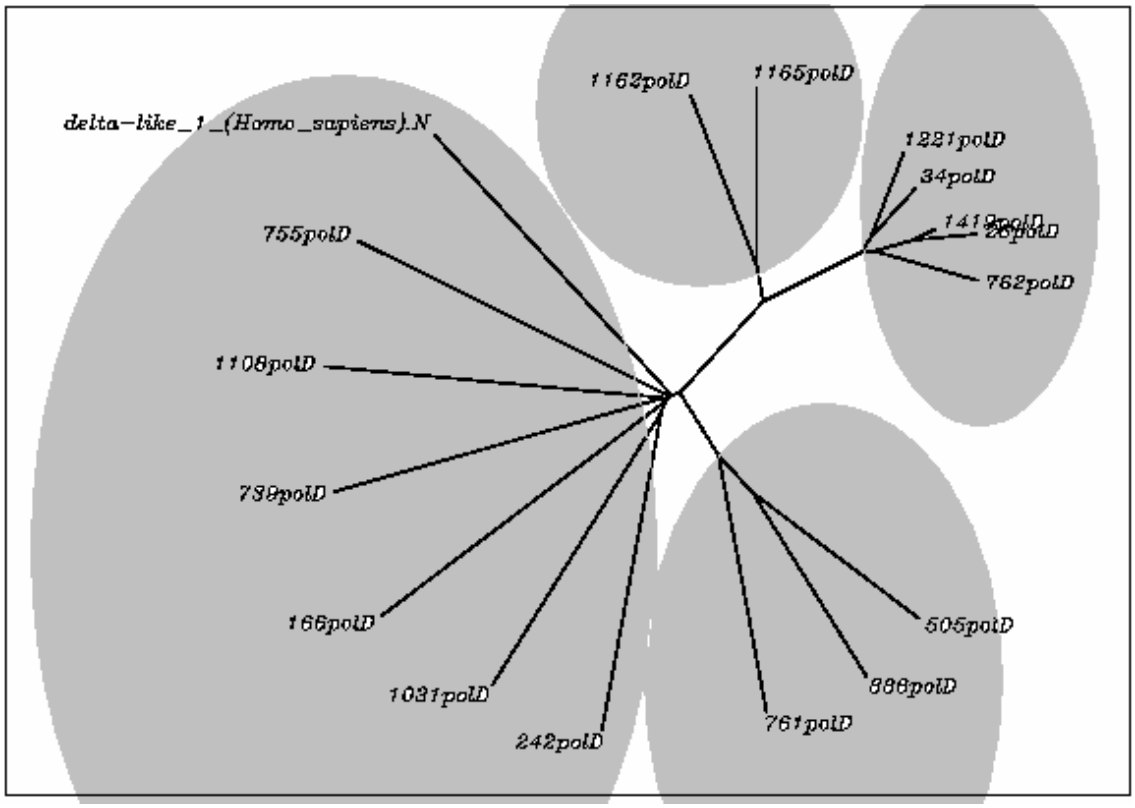


Fig 22. The delta-like CFERV cluster in four groups, the largest group contain a delta-like chain from human.

Tab 3. *CFERV* delta-like virus detected by RetroTector©. For chain id 456 an observation was done, the Pol Putein ID 920 is not the highest number but was chosen because its average score was highest.

Delta Chains

Score	select chain id	Chain Genus	CHR	RepBaseFinds	Stop Codons	Shifts	LTRDivergence	Env Puteinid	Pol Puteinid	Select pol id	Pro Puteinid	Gag Puteinid
371	456	DC	12	HERV9 3566 23579369 51	18	20	null		920, 925, 924	920	926, 921	918,923, 922
467	765	DC	23	HERVR 4806 11077607 57	22	17	null		1598, 1602, 160X	1602	1603, 1599	1596, 1600, 159X

Comparison of *CFERV* and endogenous retroviruses from other species

The contents of *ERV* in both the dog and chicken genome are less compared to the content of *HERV* in the human genome. Elements retrieved with RetroTector© > 299 score were 3164 in the human genome and 262 in the chicken genome (Jern., 2005) and in this study, 254 in the canine genome. *CFERV* aligned to *ERV* in many other species show high conformity. ERV 21 *Python molurus* (AF 500296) aligned to *CFERV* beta-retrovirus id 991 (RT) share 50 % identical puteins on a part of the sequence that is 36 amino acids long. 31 of 43 amino acids are identical for *CFERV* gamma-retrovirus id 988 (RT) and *ERV9* PH1 RT. Because the near companionship between human and dog for over 14 000 years there is a possibility that horisontal contagion might have occurred. *HERV H* consensus *pol* and *CFERV* id 988 (RT) have 70 % puteins at the same position. Exactly the same CF chains (ID 991, 1177, 900, 478

and 1014) in the alignment above are aligned to *Crocodyl niloticus*. In five of the *CFERVs* and in the *Crocodyl niloticus ERV*-chain, there is 100 % identity in six unique places (Fig. 24).

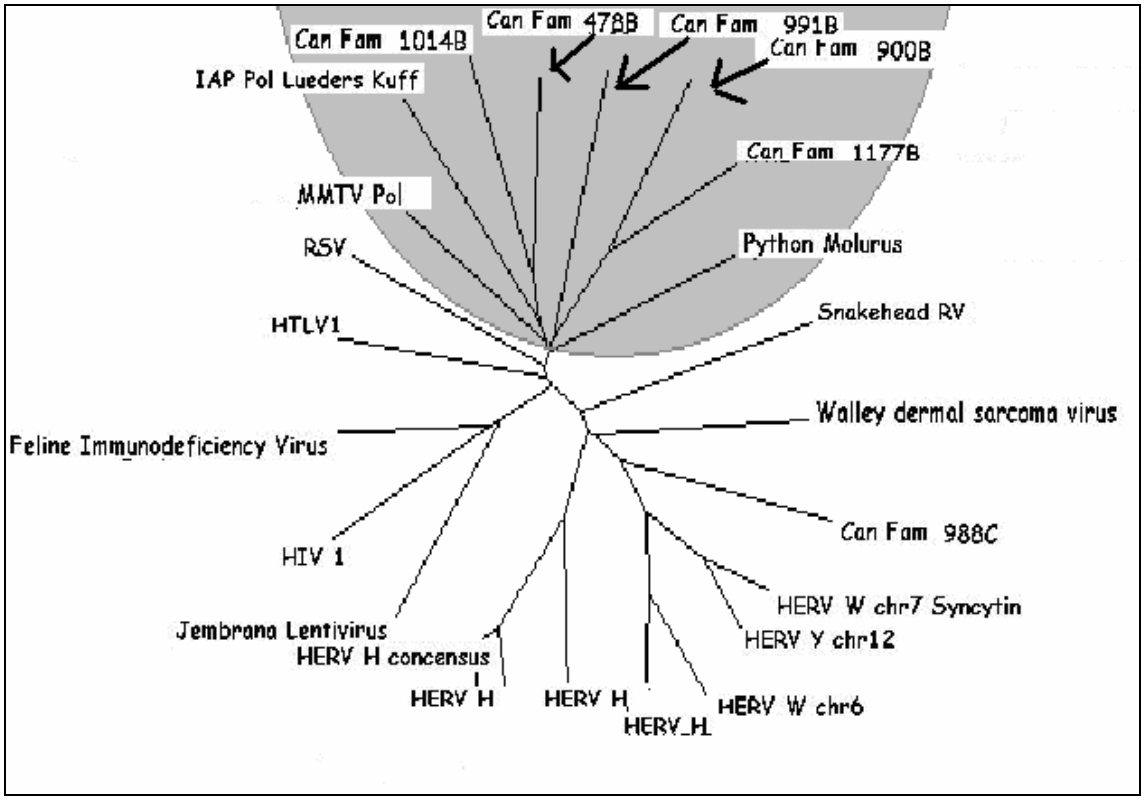


Fig 23. RV from Python, human, dog, cat, fish and mouse, in an unrooted Neighbor-Joining tree.The grey area marks closer relatives around the Python Pol sequence. Notice that the branches are moveable at the nodes.

```
                    Can Fam  1014B FFTIPLAPQNYERFAFSIPSINHMEPAKRFH--------WKVLPQGMANSPTLCQEFVAR
                    Can Fam  478B FFTIPLAPQDCERFAFSIPSINHVEPAKRFH--------ZKVLPQGMANSSTLCQEFVAR
                IAP Pol Lueders Kuff FFSIPLCPRDRPRFAFTIPSINSDEPDNRYQ--------WKVLPQGMSNSPTMCQLYVQE
                        MMTV Pol FFNIKLHPEDCKRFAFSVPSPNFKRPYQRFQ--------WKVLPQGMKNSPTLCQKFVDK
                    Can Fam  900B ----PLHQVGRRKFAFSSPCINSGAPLKRYZ--------WKVLPQGMVNSPSVCQDFVDH
                    Can Fam  1177B FFVIPLHQDDRENFAFSLSCVNSRSPLKRYQWILLKRYQZKVLPQEMVNSFSICQDFVDH
                    Can Fam  991B LFTIPLATQDKKQFTFSVPSINNIEPIQRYQ--------WKVLPQGMASSPTKCZHYMAL
                Python Molurus R  FFSIPLHPKDKHIFAFTVPALNNSQPTSRYQ--------WKVLPQGMLNSPTMCQYFVSQ
                        RSV Pol FFSIPLAEQDREAFAFTLPSVNNQAPARRFQ--------WKVLPQGMTCSPTICQLVVGQ
                        HIV 1 Pol YFSVPLDEDFRKYTAFTIPSINNETPGIRYQ--------YNVLPQGWKGSPAIFQSSMTK
            Feline Immunodeficiency Virus Pol YFTIPLDPDYQQYTAFTLPSKNNQGPGRRYV--------WKSLPQGWVLSPLIYQSTLDN
                Jembrana Lentivirus Pol YFTIPLDENFRQYTAFSVVPVNREGPLERYH--------WNVLPQGWVCSPAIYQTTTQE
                HERV W chr7 Syncytin FFCIPVHPDSQFLFAFEDTS----NPTSQLT--------WTILPQGFRDSPHLFGQALPR
                        HERV Y chr12 FFCIPVHPDSQFLFAFRDPS----NLTSQLP--------WTVLPQEFRDSPHLFGQALAQ
                        HERV W chr6 FFYIPLYSDSQFLFTFEDPT----DHTSQLT--------WMALPG-FRDSAHLFDQALAQ
                        ERV9 PHI R FFCVPLHSDSQFLFAFEDPT----NHTSQLT--------WMVLPQGFRDSPHLFGQALAK
                    Can Fam  988C FPHVPVQADSQQLFAFQDLI----GQNIQLT--------WAALPQGFRDSPHSFGQALAK
                        HERV H FFTIPLHPSSQRLFAFTRLT--LTPIRLQIT--------WAVLPQGFTDRPHYFSQAQVS
                HERV H concensus FFTIPLHPSSQPLFAFTWTDP-DTHQAQQIT--------WAVLPQGFTDSPHYFSQALSH
            Walley dermal sarcoma virus FFSVPIHKDSQYLFAFTFEG-------HQYT--------WTVLPQGFIHSPTLFSQALYQ
                        HTLV1 FFQIPLPKQFQPYFAFTVPQQCNYGPGTRYA--------WRVLPQGFKNSPTLFEMQLAH
                    Snakehead RV Pol FWSIRLEEECQYLFAFTFDT-------QQYT--------WTRLPQGFHASPGIFHQALYN
                                     :         :*          :           **
```

Fig 23 b. Alignment of 22 ERV (Pols) from six different species.

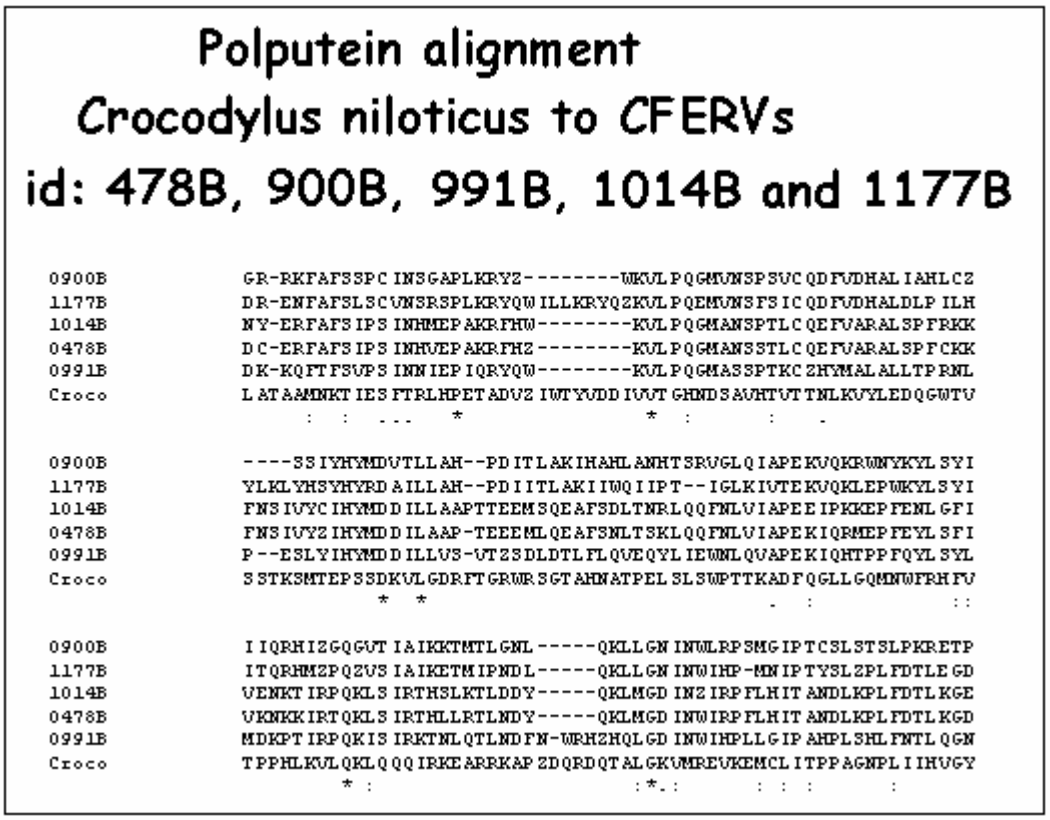


Fig 24. CFERV pol sequences aligned to Crocodylus niloticus.

Highly scored *CFERV* in all four genera were aligned to human (*HERVW* chr7 9105739 syncytin *pol*, *HERV H* consensus *pol, HERVFH21 pol, BaEV M7 pol*), sheep (Jaagsiekte *Pol* M80216), cat (Feline Immunodeficiency Virus *Pol* DQ192583), reptile (*Python molurus RV Pol AF500296, Crocodylus niloticus* Polputein AJ438130), mouse (*MMTV Pol NC 001503*), marsupial (Opossum type D *Pol AF224725*) fish (*Snakehead RV pol NC 001724*, Walleye dermal sarcoma virus *pol NC 001867*), and insect (*Drosophila Osvaldo Gypsy Pol AJ133521*). The resulted unrooted neighbor joining tree shows similarities to the unrooted *Pol* based neighbor joining tree with the seven retroviral genera and the *ERV* class definitions, created previously (Jern et al., 2005), if accepted that the branches are free to move around the nodes. The Gamma and the Beta elements are evolutionary distant retroviruses as represented in the Phylogenetic analysis where they cluster at separate and well set apart branches in the derived tree as shown here (Fig. 25).

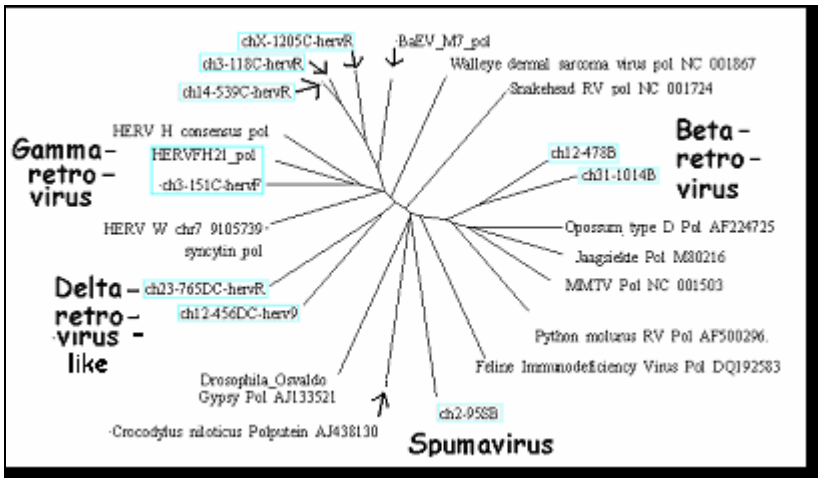


Fig25. Nine of the highest scored CFERV polputein sequences in all four genera gamma, beta, delta-like and spumavirus, aligned to polputein reference sequenses from different species shows a connection of all ERVs in an unrooted retroviral neighbor joining (NJ) tree

LTR divergence and ORF

The age of integration in host genomes for ERVs is commonly estimated as the nucleotide divergence between the 5'- and 3' *LTR* of a provirus. *LTR* divergence could be a measurement for tracking the age of the elements. When the element is old the divergence is high. At the time of insertion in the host genome the *LTRs* are identical. When the insertion is recent there is low divergence between the two *LTRs* flanking a provirus. For 39 % of *CFERV* the *LTR* divergence is known, (see Fig. 26).

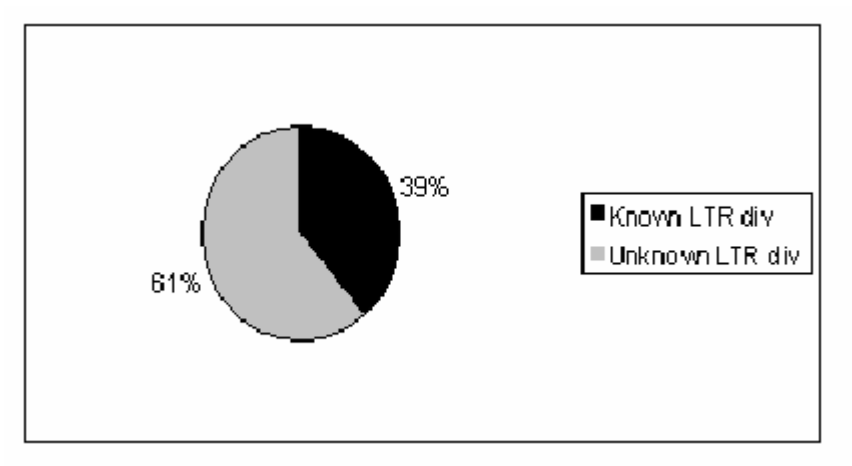


Fig 26. CFERV: distribution of known and unknown LTR divergence.

For chains with known *LTR*- divergence the chains were grouped in three groups. Most of them (70 %) in group 3, wich have >10 % divergence and >10 stops and shifts. For group 2 with divergence 5-10 % and 4-10 stop and shift the number of chains is (26 %) and for the youngest with possible functional elements in group 1 the number of chains is (4 %)

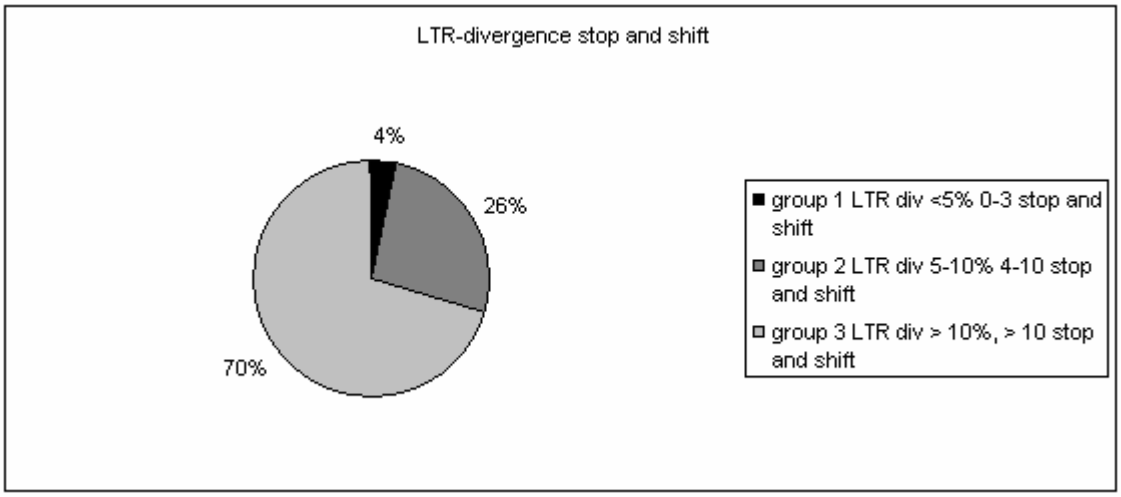


Fig27. The three groups of different LTR divergence.

Chain ID 240 C (RetroTector©,) with score 1542 is a complete *CFERV* with all 4 genes, low *LTR*- divergence, 1 stop and 2 shifts. ID 240 has Rep Base Finds *HERVFH21*.

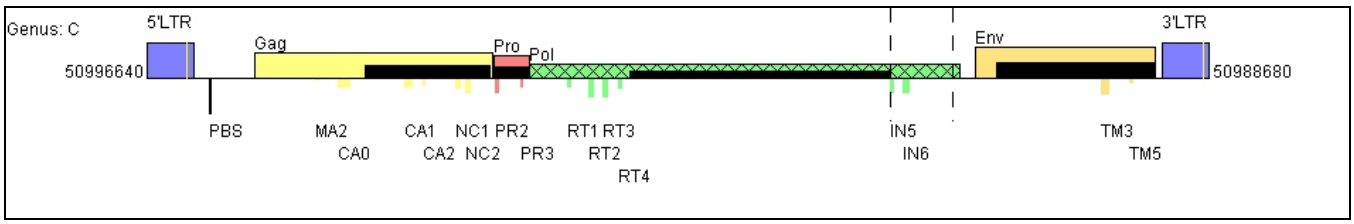


Fig 28. ID 240 a complete CFERV in the youngest group (1). Figure from RetroTector.

Discussion

The number of *CFERV* element was drastically decreased from the first list, 670 elements, (data not shown), made in October 2006 compared to that little amount, 254, of *CFERV* gained in March 2007. Twice during this study (in January and March 2007) the genomic sequence data (CamFam) were delivered in improved versions. Both these times the new versions resulted in simplified and improved chain collections. The amount of *L1* fragments and the number of Rep Base Finds, (for the chains) decreased (but did not disappear); apart from the first Rep Base Finds that was really wanted. The removal of *L1* and *ALU* elements was more succesfull than earlier but there are still some *L1*s that need to be eliminated in CanFam2crt10v070313 (CF2crt). The low amount of *CFERVs* in dog i.e. 254 unique chains score > 299 with RetroTector© in this study, was similar to the amount found in chicken, G*allus gallus* (gg01) with RetroTector© in a study made by Jern, P. in 2005, he found 262 elements. That confirm that the amounts of *ERV* in dog and chicken are very low compared with other animal genomes which corresponds well with statement done by Blomberg, J. (unpublished), the dog and chicken genomes amount of *ERV*s are a fifth of the amount in human and schimpanzee genomes. Tranposable elements is a part of the genome in which *ERV*s belongs. The amount of *TE* follow the size of the genome. In a small population the TE integration will increase because a fewer animal can not efficiently select against number of *TE* (Biémont et al., 2006) That indicate this is a statement improper for *Canis familiaris*, with its small amount of *CFERV*. According to the two important bottlenecks, the first when the wolf was domesticated, and the second when the breeds were shaped almost all domestic dog

populations (breeds) are small and have a limited genetic variation. Even if some breeds are large, the breeding program use obstructions from geographic and isolated breeding lines within the breeds. Tasha the boxer female used for CanFam2.0 is from a breed with limited genetic variation, and in her genome, the amount of *CFERV*s is low. It will be interesting to compare the amount and variation of CFERV in different dog breeds.

The gammaretrovirus and the betaretrovirus in dog are not following the same pattern according to gene contents. In the betaretrovirus genus the largest group has two genes (a *pol* gene and either one of *gag* or *env*) and in the gammaretrovirus genus the largest group contains three genes, (all selected sequences contain at least one *pol* gene). For C –elements, the *pol* gene, could be the one and only gene in the sequence, a large distinction from the B-elements. *Gag* and *pro* genes are more common in the C-elements than in the B-elements. *Env* genes are legio in B-elements but the least common gene in the C-elements. *Env* genes in dog are not the only part that is missed in the C- elements. Often both *env* genes and 3' *LTR* are missing, maybe as a consequence that RetroTector© have more difficulty to find these in the dog genome than in other genomes. Alternatively due to deletions of the 3'part of these CFERVs. Almost 50% of C chains score > 1000 lacks 3' *LTRs* and just a third of them have both 5' and 3' *LTRs*. The dog genome was considerably more difficult for RetroTector© to data mine and the evaluation of obtained results was challenging. *Env* genes are present in the functional elements and that confirm that B elements are more intact and in this study found to be more complete than the C element.

The unrooted tree of *CFERV*s, for all appropriate genera show high similarity to the basic orginal unrooted tree picture (Jern et al., 2005). When accepted that the branches are moveable at their nodes *CFERV*s of B and C chains were well separated in the unrooted tree. *CFERV*s show similarities to *ERV*s in other species. *HERVF*-like chains in dog clustered with *HERVF*s in the human genome.

RetroTector©

When using RetroTector©, the retrieved retrovirus chains and genes (*gag, pro, pol* and *env*) could not be counted nor collected at once. Because the sequences are fragmented into "chunks" and several copies appear. Every chain and protein sequence (gag, pro, pol and env) have to be sorted out to exclude that they were on the same place as another chain or gene with a higher chain or average score. By definition, only one ID, chain or protein coding gene, could occupy one unique position on each one of the chromosomes. Today those decisions have to be done after manual inspection until RetroTector©, will perform that job.

Other important observations that were done using RetroTector©: the chosen *pol* id number 920 obtained the highest average score that is usually the highest identification number which followed by the highest average score seen in RetroTector©. For the highest pol id number (in this case 925) on chain id 765 one last number was missed for two gene identification number (pol id 160X and gag id 159X).

Conclusions

The relative low abundance of endogenous retroviruses in the dog genome compared to rodent and primate genomes suggests that the dog, like chicken has been able to protect itself from large amounts of insertion of endogenous retroviruses. These both species have only 0,2 percent retroviruses compared with mouse, human and chimpanzee, species which all have a

lot more ERVs in their respective genomes. My study claimed that the dog endogenous retroviruses amount to 254 chains, 0,075 % of the entire genome that consist of 2. 38 Mbp.

This project's aim was to analyse the endogenous retroviruses in the dog's genome, CanFam (CF2c). The complexity of dog endogenous retroviruses was identified and classified to the different genera in chains and in proteins. *CFERV*s were classified and characterized using a bioinformatics approach. Phylogenetic studies were also performed to group the identified *CFERVs* in to retroviral genera. Retrovirus sequences were retrieved from GenBank and compared from the following species: dog, human, mouse, pyton and crocodile. In future studies, selected chains should be further characterized according to their functional capacity.

Using RetroTector©, the following classes were identified in the dog genome: beta, gamma, spuma and delta-like virus and all main putein sequences. The 254 *CFERV* elements that were retrieved by RetroTector© was based on the criterium that they should pass the lower limit of 300 score. On this basis, the amount of *CFERV*s that was found in this study is only a part (0,075 %) of the expected amount (0.2 %, Blomberg unpublished) of the endogenous retrovirouses in the dog's genome.

The gamma and betaretroviruses in dog are not following the same pattern according to division of the included genes (*gag, pro,pol* and *env*). (Fig 29). All selected sequences contain at least one *pol* gene. For the concern of the gamma element, the *pol* gene could be the exclusive gene in the sequence. This is the case for 11 % of the C-elements and that is a large distinction compared to the B-elements. There is no beta chain with only the *pol* gene. The Beta chains in dog are more complete compared to the gamma chains. This is the case also in other species where *Gag* and *pro* genes are more common in the C-elements than in the B-elements. *Env* genes are legio in B-elements but the least common gene in the C-element. A part of the C-elements lacks even 3' *LTR*, Presumably, RetroTector©, could have difficulties finding *Env* and 3' *LTR* in CanFam (CF2c).

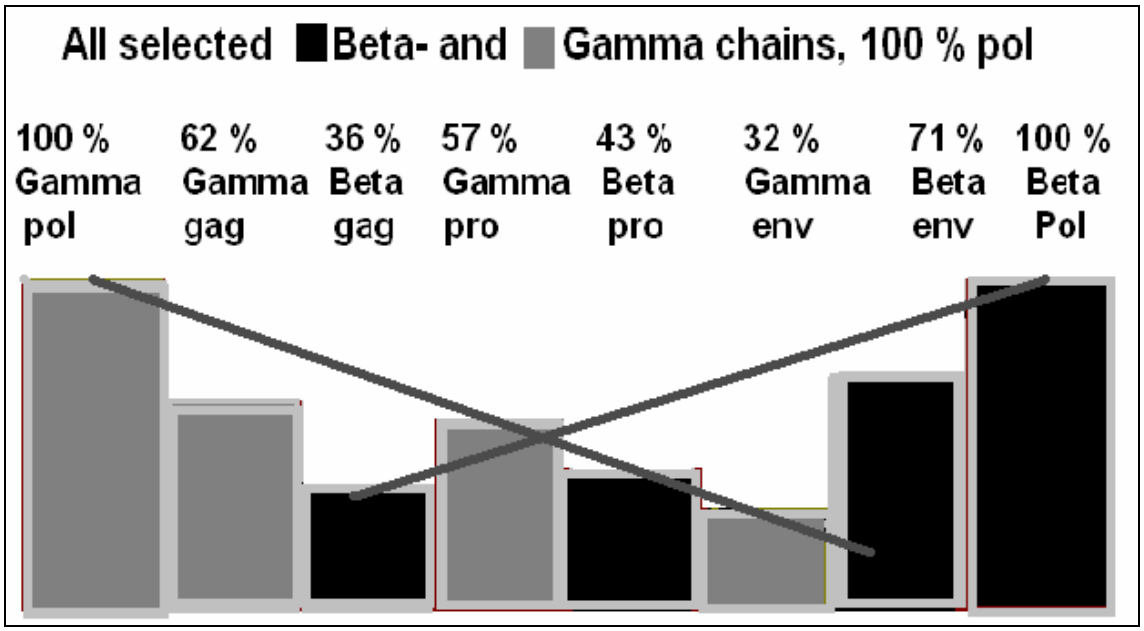


Fig 29. B-element and C-element goes the opposite way in ranking the genes after the basical gene, the pol gene.

In the dog and human genome there are closely related *ERV*s that cluster with *HERVFH21*-like elements suggesting a common origin for this type of retrovirus. Further studies should be done in this area. *HERVF* is like *HERVH* clustered within Gammaretrovirus and may have appeared in mammalia for 45 Myr ago.

Acknowledgements:

Great thanks for kindly support and guidance to my supervisor, Professor Göran Andersson, at the Department Animal Breeding and Genetics SLU. Great thanks for all help with RetroTector©, to senior lecture Göran Sperber at the Department Neuroscience Uppsala University. Great thanks for all answers, to Professor Jonas Blomberg at the Department Medical Science Uppsala University. Great thanks for inspired introduction to bioinformatic to Associate Professor Erik Bongcam-Rudloff at the Linnaeus Centre for Bioinformatics (LCB) Uppsala University.

Svensk sammanfattning

Syftet med min studie var att på molekylär nivå söka, klassificera och analysera hundens endogena retrovirus, *CFERV* med programmet RetroTector©. Från dessa data har komplexiteten av hundens *ERV* definierats inklusive dess proteiner. Kedjor och dess proteinsekvenser har jämförts med andra individer från samma art eller mellan individer från olika arter. Av hundens retrovirus insamlades alla kedjor över 300 poäng, av det bioinformatiska datorprogrammet, RetroTector©. Kopior och *L1*-liknande element (det är flera element som liknar retrovirus men inte uppfyller alla krav) sorterades bort. Kopior återfanns i två nivåer (först på kedjenivå, sedan på *pol*putein-nivå) på grund av att överlappade kedjor används i programmet för att inte riskera att missa något element. Kopior plockades bort med hjälp av (i RetroTector©) kända positioner i början och slutet av kedjor och *pol*-puteiner. *L1*-liknande element plockades bort med hjälp av Rep Base Finds funktionen i RetroTector©. Efter rensning återstod endast 254 unika sekvenser, med gränsen satt till minst 300 poäng och med den nödvändiga *pol*-genen av hundens retrovirus, *CFERV*, vilka analyserades vidare. Analysen resulterade i 237 gamma-retrovirus, 14 beta-retrovirus, 2 deltalika retrovirus och en spuma-retrovirus. Distributionen av ingående gener i retroviruselementen är normalfördelade med viss skevhet. De flesta (43 %) retroviruselement har tre gener, (en *pol* gen och två andra). En stor skillnad mellan hundens beta- och gamma-kedjor är att bland betakedjorna finns inte en enda kedja med endast en *pol*-gen till skillnad från Gamma-kedjorna, där uppgår antalet kedjor med en enda *pol*-gen till 21 stycken (11 %). Mängden av högpoängs *CFERV* som återfanns i denna studie var 254 stycken vilket utgör 0,075 % av hundens genom på ca 2,3 miljarder baspar. Många av hundens 254 retrovirus liknar varandra och överensstämmer väl med retrovirus i andra djurarter *HERVFH21* hos människa och 18 stycken *HERVF*-lika sekvenser i hund visar genom alignment i Clustal W på ett gemensamt ursprung. *HERVF i* människa och i gamla världens primater är en tidig *ERV*, kanske 60 miljoner år men *HERVF* upptäckes sent.

Endogena retrovirus kan smitta, liksom övrigt virus, horisontellt mellan arter som lever nära varandra. Det kan ske genom articifiella barriärer då olika arter tvingas leva på en mycket begränsad yta, som mellan olika arter i en djurpark eller som mellan människa och hennes husdjur. Hunden lever kanske lika nära människan som människor lever nära varandra, därför tros hund och människa ha gemensamma *ERV* som *HERVF* vilket bekräftas av denna studie. Horisontell smitta kan även förekomma genom predation, från byte till rovdjur liksom till

människa som äter kött av andra ryggradsdjur. Endogena retrovirus skiljer sig från exogena retrovirus genom att de är införlivade i genomet och ärvs ner på mendelsk väg, liksom de övriga generna. Det är den andra vägen för ”smitta” *ERV* förs vidare vertikalt genom arvet, från förälder till avkomma. Det är endast när ett retrovirus infekterar en könscell och blir en del av genomet som retroviruset blir endogent. *ERV* kan ses som ett minne av en tidigare infektion. Om retroviruset inte stör värden alltför mycket, det vill säga om värden kan överleva och fortplanta sig med viruset intakt och heller ingen selektion sker emot viruset, kan det efter cirka minst en million år bli fullständigt befäst i en art. Endogent retrovirus är en nukleotid-kedja som är 9-12 kb lång. För att uppfylla kriteriet retrovirus ska kedjan innehålla två *LTR*:er (en i varje ände) och helst de fyra generna *gag, pro, pol* och *env*. Dessa gener kodar för retrovirusets (protein). Ett retrovirus ser ut som följer: 5’ änden börjar med en 5’*LTR*, en cap som motsvarar den första nukleotiden i det mRNA som bildas, vidare en *PBS* vilken startar syntes genom att binda till tRNA. tRNA:ts aminosyra ger namn åt aktuell *ERV*, *HERVF* har tRNA med phenylanalin. Efter *PBS* kommer de fyra speciella generna: *gag*, *pro*, *pol* och *env*, som är avgörande för retrovirusets funktion. *Gag* kodar för strukturproteinen, matrix, capsid och nucleocapsid. *Pro* kodar för proteas. *Pol* kodar för ett integras och för omvänt transkriptas som möjliggör omvänd transkription (RNA till DNA), vilket är en absolut förutsättning för framgång för virus. *Env* kodar för struktur proteiner för retrovirusets yta/vägg och dess transportproteiner genom väggen. I 3’ änden efter *env*-genen avslutas kedjan med en LTR. Endast en dryg femtedel (22 %) av alla hundens endogena retrovirus är kompletta med alla de fyra generna. I denna studie har gränsen satts för minst en *pol* gen för att kvalificera som retrovirus och för möjlighet att jämföra *pol* sekvenser. *Pol*-genen har en särställning genom sin höga konservering över långa evolutionära tider och sin stabila proteinsekvens (mer stabil än motsvarande nukleotidkedja). Därför är det *Pol*-sekvenserna som används i alla fylogenetiska jämförelser inom hund och mellan hund och andra arter. De, från allra första början, identiska *LTR*:en som ”inhägnar” retroviruset kan användas som ett mått på virusets ålder. Det sker mutationer, insertioner och deletioner över tiden i all avsmassa, med hjälp av känd mutationshastighet och mätning av skillnaden (divergence) mellan de båda *LTR*:en kan ett mått på retroviruselementets ålder erhållas. Denna analys har inte utförts och beräknats i denna studie.

Hundens genom CanFam2.0 blev fullständigt kartlagt på Broad Institute MIT & Harvard och publicerades i december 2005 (Lindblad-Toh et al., 2005). Hundsekvensen anges som en högkvalitativ sekvens och den annoteras och förbättras fortlöpande. Den senaste tillgängliga versionen är, CanFam2crt10v070313 vilken har används i denna studie. Hundens retrovirus är inte beskrivet på ett utförligt sätt i litteraturen. Hunden och kycklingen har skyddat sig väl emot retrovirus jämfört med människa och schimpans.

En djupare studie med jämförelse av *CFERV* och kända patogena retrovirus i andra arter behöver göras liksom en djupare jämförande studie av *ERV* (*HERVF*) i människa och hund eftersom dessa arter under de senaste 14 000 åren stått varandra nära, rent fysiskt.

References

Andersson, A.C., 2002. Studies on Human Endogenous Retroviruses (HERVs) with Special Focus on ERV9. *Digital Comprehensive of Uppsala Dissertations from the Faculty of Medicine* 1165

Biémont, C and Vieira, C., 2006. Junk DNA as an evolutionary Force. *Nature Vol 443/5 October 2006*

Bock, M., and J. P. Stoye. 2000. Endogenous retroviruses and the human germline. Curr. Opin. Gwnet. Dev. 10:651-655.

Blomberg, J., Ushamechis, D.,Jern, P., 2004.Evolutionary Aspects of Human Endogenous Retroviral Sequences (HERVs) and Disease *Retroviruses and Primate Genome Evolution*, edited by Eugene D. Sverdlov. ©2004 Eurekah.com.

Hughes, J. F., Coffin, J. M., 2005. Human endogenous retroviral elements as indicators of ectopic recombination events in the primate genome. Genetics. *2005 Nov;171(3):1183-94. Epub 2005 Sep 12.*

Jern, P., Sperber, G. O., Blomberg, J., 2004. Definition and variation of human endogenous retrovirus H. *Virology 327 (2004) 93-110*

Jern, P., 2005. Divergent Patterns of Recent Retroviral Intergration in Human and Chimpanzee Genomes; Transmissions from Other Primates to Chimpanzees .*In progress*

Jern, P., 2005.Genomic Variation and Evolution of HERV-H and other Endogenous Retroviruses (ERVs). *Digital Comprehensive of Uppsala Dissertations from the Faculty of Medicine* 62

Jern, P., Sperber, G. O., Blomberg, J., 2005. Use of Endogenous Retroviral Sequences (ERVs) and structural markers for retroviral phylogenetic inference and taxonomy. *Retrovirology 2005, 2:50 doi:10.1186/1742-4690-2-50*

Kjellman, C., Sjögren, H. O., Widegren, B., 1999. HERV-F, a new group of human endogenous retrovirus sequences. *J. Gen. Virol. Sep*;80 (Pt 9):2383-92

Lavie, L., Medstrand, P., Schempp, W., Meese, E. and Mayer, J., 2004. Human Endogenous Retrovirus Family HERV-K(HML-5): Status, Evolution, and Reconstruction of an Ancient Betaretrovirus in the Human Genome. *J. Virol.* 78 (16): 8788–8798.

Lindblad-Toh, K., Wade, C. M., Mikkelsen, T. S., Karlsson, E. K., Jaffe, D. B., Kamal, M., Clamp, M., Chang, J. L., Kulbokas, E. J., Zody, M. C., Mauceli, E., Xie, X., Breen, M., Wayne, R. K., Ostrander, E. A., Ponting, F. G., Smith, D, R., deJong, P. J., Kirkness, E., Alvarez, P., Biagi, T., Brockman, W., Butler, J., Chin, C., Cook, A., Cuff, J., Daly, M. J., DeCaprio, D., Gnerre, Sante., Grabherr, M., Kellis, M., Kleber, M., Bardeleben, C., Goodstadt, L., Heger, A., Hitte, C., Kim, L., Koepfli, K-P., Parker, H. G., Pollinger, J. P., Searle, S. M. J., Sutter, N. B., Thomas, R., Webber, C., Lander, E. S., 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature Vol* 438|8 December 2005|doi:10.1038

Oja, M., Peltonen, J., Blomberg, J. and Kaski, S., 2007. Methods for estimating human endogenous retrovirus activities from EST databases. *BMC Bioinformatics 2007,* 8(Suppl 2):S11 doi:10.1186/1471-2105-8-S2-S11

Sperber, GO., Airola, T., Jern, P., and Blomberg, J. 2007. Automated recognition of retroviral sequences in genomic data RetroTector(C)Nucleic Acids Res. 2007 Jul 17; [Epub ahead of print]).

Weiss, R. A., The discovery of endogenous retrovirises 2006. *Retrovirology. 2006; 3: 67.* Published online 2006 October 3. doi: 10.1186/1742-4690-3-67.

http://www.ensembl.org/index.html

http://www.ensembl.org/Canis_familiaris/index.html

http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=rv.table.7768

http://www.ncbi.nlm.nih.gov/entrez/

http://bioweb.pasteur.fr/seqanal/interfaces/clustalw-simple.html

Reference sequences used in alignment
BaEV_M7_pol
Crocodylus_niloticus_Polputein_AJ438130
(insect)Drosophila_Osvaldo_Gypsy_Pol_AJ133521
ERV9_PH1_RT
(cat) Feline_Immunodeficiency_Virus_Pol_DQ192583
HERV_H_consensus_pol
HERV_H_RGH1_pol
HERV_H_RTVLH2_pol
HERV_W_chr6_141432567_ERV9_like_pol
HERV_W_chr7_9105739_syncytin_pol
HERV_Y_chr12_51022911_pol
HERVFH21_pol
HERVHRGH2_pol
HIV1_Pol_NC_001802
HTLV1_Pol_NC_001436
(mouse)IAP_Pol_Lueders_Kuff_MUSFLIAP
(sheep) Jaagsiekte_Pol_M80216
(cattle)Jembrana_Lentivirus_Pol_NC_001654
(mouse)MMTV_Pol_NC_001503
Opossum_type_D_Pol_AF224725
Python_molurus_RV_Pol_AF500296.
RSV_Pol_AF033808
(fish)Snakehead_RV_pol_NC_001724
(fish)Walleye_dermal_sarcoma_virus_pol_NC_001867