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

The identification and classification of endogenous retroviruses in the horse genome

Batmagnai Enkhbaatar

Department of Animal Breeding and Genetics
Examensarbete 375
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

The identification and classification of endogenous retroviruses in the horse genome

Batmagnai Enkhbaatar

Supervisor:
Erik Bongcam-Rudloff, SLU, Department of Animal Breeding and Genetics
Göran Andersson, SLU, Department of Animal Breeding and Genetics
Matthew Peter Kent, UMB, Department of Animal Breeding and Genetics
Examiner:
Gabriella Lindgren, SLU, Department of Animal Breeding and Genetics

Credits: 30 HEC
Course title: Degree project in Animal Science
Course code: EX0556
Programme: Erasmus Mundus Programme - EMABG
Level: Advanced, A2E

Place of publication: Uppsala
Year of publication: 2012
Name of series: Examensarbete 375
Department of Animal Breeding and Genetics, SLU
On-line publication: http://epsilon.slu.se

Key words: equine genome, endogenous retrovirus, transposable elements

THE IDENTIFICATION AND CLASSIFICATION OF ENDOGENOUS RETROVIRUSES IN THE HORSE GENOME

BATMAGNAI ENKHBAATAR
870818-P790

THESIS ANIMAL BREEDING AND GENETICS (EX0556)
JUNE 2012

Swedish University of Agricultural Sciences

SUPERVISORS
Erik Bongcam-Rudloff Associate Professor at SLU
Göran Andersson Professor at SLU
Matthew Peter Kent Professor at UMB

Education and Culture
Erasmus Mundus

CONTENTS

Contents 3
Abstract 1
1. Background 1
1.1 Endogenous retrovirus. 2
1.2 Equine SINE elements 4
1.3 The life cycle of retrovirus 4
1.4 Estimation of the ERV age 5
1.5 The amount of ERVs vary in different mammals 6
1.6 The nature of integrations related with host genes. 7
1.7 Classification of Endogenous retroviruses 7
2 Objectives of the study 9
3 Materials and methods 9
3.1 Experimental approach 10
3.1.1 SINE-PCR[30] 10
3.1.2 Pan-pol PCR amplification 11
3.1.3 Polymorphism of EcERV Beta1 region between 13 different breeds 13
3.2 Bioinformatics approach 14
3.2.1 LTR_STRUC 14
3.2.2 NCBI-BLAST 16
3.2.3 Retrotector© 16
4 Results 16
4.2 Experimental results: 16
4.1.1 SINE-PCR results 17
4.1.2 Pan-pol PCR results 17
4.1.3 Sequencing results: 17
4.1.4 Polymorphism of EcERV Beta1 19
4.2 Bioinformatics results 19
4.2.1 LTR_STRUC results 19
4.2.2 Chromosomal distribution 22
4.2.3 Phylogenetic analysis of Equine endogenous retroviruses 24
4.2.4 Unique integrations 26
5 Conclusions 28
6 Discussion 29
7 Further analysis: 30
References: 30
Appendices 32
Appendix 1. Equine endogenous retrovirus candidates 32
Appendix 2. Unique integrations 38
Appendix 3. Sequencing results 41
Acknowledgements: 42

Abbreviations

ALV	avian leukosis virus
BLV	bovine leukemia virus
bp	base pairs (nucleotides)
CA	capsid protein
cDNA	complementary deoxyribonucleic acid
DNA	deoxyribonucleic acid
Env	envelope
ERV	endogenous retrovirus
EcERV	endogenous retrovirus of Equus caballus
EIAV	Equine infectious anemia virus
Gag	group specific antigen
IN	integrase domain
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
MA	matrix protein
MLV	murine leukemia virus
MMTV	mouse mammary tumour virus
NC	nucleocapsid protein
ORF	open reading frame
PCR	polymerase chain reaction
PBS	primer binding site
Putein	putative protein
Pol	polymerase gene
Pro	protease gene
RV	retrovirus
SU	surface unit
TE	transposable elements
tRNA	transfer ribonucleic acid
U3	unique 3’-sequence
U5	unique 5’-sequence
XRV	exogenous retrovirus

ABSTRACT

Endogenous retroviruses (ERVs) are sequences that derived from ancient retroviral infections of germ cells and integrated in humans, mammals and other vertebrates millions years ago. These *ERVs* are inherited according to Mendelian expectations in the same way as all other genes in the genome. Size of complete endogenous retrovirus is between 8-12 kb long in average and contains *gag, pro, pol* and *env* genes that always occur in the same order. Coding sequences are flanked by two LTRs (Long Terminal Repeat sequences). Most ERVs are defective that are carrying multitude of inactivating mutations. However some ERVs still have open reading frames in their genome. These ERVs settle close to functional genes or within the genes and can influence or control functions of the host genes using their LTRs. Most integration has deleterious effects. However some integration could be example of positive co-adaptation as syncitin which is involved to form the syncytial layer of the placenta. The first equine endogenous beta retrovirus which is EcERV-Beta1 has been found in 2011 by Antoinette C. van der Kuyl[1]. The first known beta retrovirus and few *pol* gene similar to foamy retrovirus were only known endogenous retroviruses fixed in the domestic horse (Equus caballus) genome. Our aim of the study was to identify other endogenous retrovirus sequences in an equine genome and classify them into groups. Based on the high number of SINEs (Equine Repetitive Element) in the horse genome we hypothesized that certain ERVs will be located sufficiently close to SINEs that they will be amplified using an unbiased SINE-PCR approach with degenerate primers. The nearest SINE element was located 5.5 k bp upstream at the 5'of the EcERV-Beta1. Pan-pol PCR was also used to find novel ERVs based on 640 bp long region of *pol* gene which is the most conserved region of ERVs. 27 complete and novel ERVs that are 13 beta, 13 gamma, 1 spuma and 249 candidate endogenous retroviruses have been revealed using LTR_STRUC tool and double checked by Retrotector© online tool and NCBI-BLAST tool. It was proven that EcERV-Beta1 which has 2 LTRs with 1% divergence between LTRs has a polymorphism among 13 different breeds.

1. BACKGROUND

The infections of first exogenous retroviruses into the germ cell could have appeared at any time over an extended evolutionary time-scale between 2 to 70 million years ago.[16] During the evolutionary time life form of endogenous retroviruses has eventually been changed from parasitic infectious type to symbiotic passive type as a part of the genome.[4] A new retroviral integration takes one million years in order to be fixed within a population.

Many extant species have been analyzed for their endogenous retroviral content, and even the extinct woolly mammoth has been shown to contain endogenous proviral fragments in its genome[23]. Surprisingly, information on endogenous retroviruses fixed in the domestic horse (Equus caballus) genome is scarce[1]. A few short *pol*-gene fragments with similarity to foamy viruses and the first EcERV Beta1 are the only endogenous retrovirus sequences from horses published today[24,1]. The first horse ERV, the full length beta retrovirus genome was retrieved from a horse chromosome 5 contig by Antoinette C. van der Kuyl and published in 2011. We pursued to find out all other EcERVs.

Our study consists of 2 sections, bioinformatics and experimental sections. Following approaches were used in the experimental section.

SINE-PCR: SINEs was used as templates for identifying novel ERVs because it is likely that several ERVs are located in the vicinity of the SINEs. As a positive control the first step was to find the nearest SINE element in the flanking region of known equine beta endogenous retrovirus (EcERV-Beta1). The idea of using SINE-PCR approach was rooted principally in the high density occurrence of SINE elements in the mammalian genome. Human SINE elements which are called Alu elements make up a large portion about 11 percent in human genome. 1 Alu occurs in every 10 kb of DNA in human genome. Horses (Equus caballus) have abundant SINE elements as well as other mammals. Recent study has estimated that $5*10^4$ copies of Equine Repetitive Element-1 are in horse genome.[2] To find the nearest SINE elements in the vicinity of EcERV-beta1 degenerate primers were designed using the multiple alignments of previously known SINEs. We have cloned the PCR products using TOPO TA cloning kit and sequenced.

The Pan-PCR approach has universal, degenerate primers which are called 5'MOP-2 and 3'MOP-2 that can amplify approximately 640 bp product of *pol* gene which is the most conserved region. The location of *pol* gene allows us to determine novel ERV from horse genome. This approach has been successfully used in other species genome like human, swine, and avian genome etc.

Recent integrations are likely to be polymorphic between different breeds. EcERV-beta1 has 2 LTRs with 1% divergence therefore this integration has occurred 2.5 million years ago (mya). Hence we have tested polymorphism of EcERV-beta1 between 13 different breeds.

LTR_STRUC was the main tool on bioinformatics part of the study and a limit was set for elements at >0.3 score in range between 0.3 and 2. The latest available version of the horse genome, EquCab2 sequence was used in the experiment. Repetitions were sorted out and excluded from further analysis. Retrotector© online tool was used for scrutinizing the results of LTR_STRUC tool.

1.1 Endogenous retrovirus.

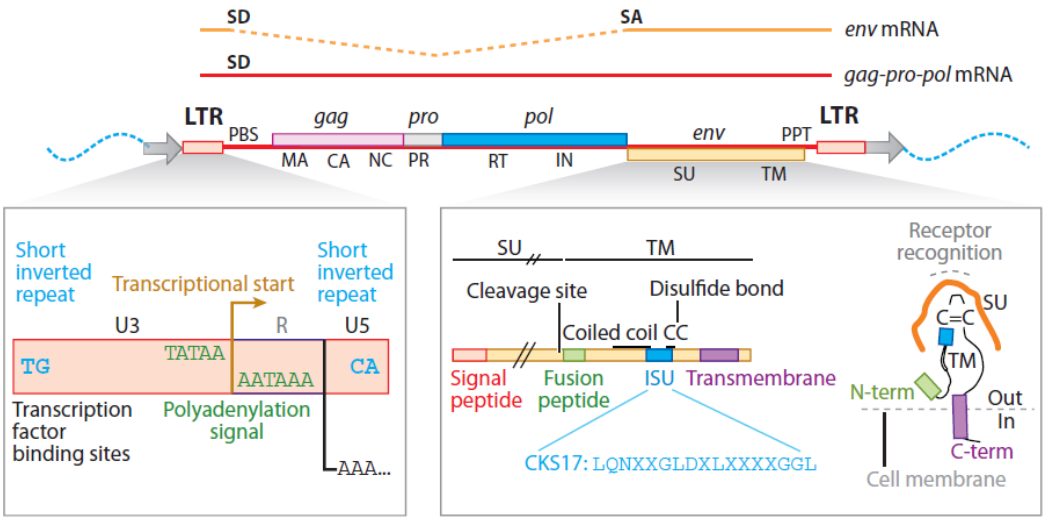


Figure 1. Provirus structure: Large arrows indicate 4-6 bp target site duplications formed during integration of the viral DNA. Simple retrovirus mRNAs are shown above. Abbreviations: PBS, primer

binding site; ISU or CKS17, immunosuppressive domain; SD, splice donor; SA, splice acceptor; ppt, polypurine tract. Viral genes (proteins): gag (MA, matrix; CA, capsid; NC, nucleocapsid); pro (PR, protease); pol (RT, reverse transcriptase; IN, integrase); env (SU, surface protein; TM, transmembrane protein).[3]

Adapted from "Effects of retroviruses on host genome function. Patric Jern, John M. Coffin Annual Review of Genetics Vol. 42: 709-732"

Endogenous retfroviruses have been studied since late 1960s. Human T-cell leukaemia virus, the first pathogenic human retrovirus was discovered in 1981 and Human immunodeficiency virus was discovered in 1983. Complete ERVs have *gag, pro, pol* and *env* genes flanked by 2 LTRs on both side of the proviral genes as shown on Figure 1. Two LTRs have the characteristic start (TG…) and stop (…CA) sequences.[11] Retroviral transcription of mRNA initiates in the R region of the 5' *LTR*. R region is a short repeated sequence at each end of the genome during the reverse transcription in order to ensure correct end-to-end transfer in growing chain. U5, on the other hand, is a short unique sequence between R and PBS. U3 is a sequence between PPT and R, which has signal that provirus can use in transcription. R is the terminal repeated sequence at 3' end. *Gag*, the first gene encodes the structural polyproteins that is cleaved into the three structural proteins forming the inner part of the virion: matrix (MA), capsid (CA) and nucleocapsid (NC). The genetic arrangement is a necessity for the translated proteins to interact in a specified order, and to guide the next proteins into positions in order to assemble the virion correctly from the outside to the inside.[6] *Gag* is the second most conserved region after *pol*. The second gene is *pro* which encodes a protease required for cleaving the different retroviral polyproteins into active subunits and located between *gag* and the *pol*. The third gene is *pol*, which is the most conserved gene, encodes reverse transcriptase and integrase. The last gene is *env*, gene encodes the structural proteins for surface and transmembrane proteins of the retroviral envelope. Envelope proteins are crucial for infection of cells of the host or cells from another individual because envelope proteins are used for fusion and retroviral entry.[11]

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 retrovirus mRNA located between the first nucleotide (*i.e.* the start site of transcription), and *gag* gene. PBS (primer binding site) consists of 18 bases complementary to 3' end of tRNA primer. PPT (polypurine tract) is primer for plus-stranded DNA synthesis during reverse transcription and located in the 3' end of the mRNA between *env* gene and the 3' LTR.

Retroelements

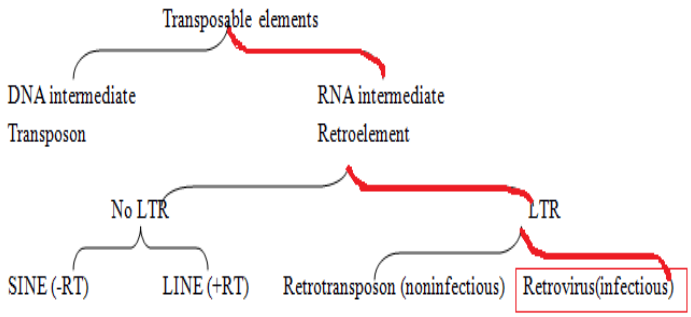


Figure 2. Retroelement classification[3] *contributed by "Jonas Blomberg"*

Almost half of the mammalian genome (45% to 48%) comprises transposons or remnants of transposons. Retroelements are transposable genetic DNA sequences that pass an intermediate RNA stage in their replication cycle. Around 42% of the human genome is made up of retrotransposons while DNA transposons account for about 2-3%. Endogenous retrovirus belongs to retroelement with LTR. **Non-LTR retrotransposons** consist of two sub-types, long interspersed elements (LINEs) and short interspersed elements (SINEs).

1.2 Equine SINE elements

Short INterspersed Elements are short DNA sequences that represent reverse-transcribed RNA molecules originally transcribed by RNA polymerase III into tRNA, rRNA, and other small nuclear RNAs. A typical SINE consists of three parts: 5'-terminal 'head', 'body', and 3'-terminal 'tail'. SINEs do not encode a functional reverse transcriptase protein and rely on other mobile elements for transposition. The most common SINEs in primates are called Alu sequences. With about 1,500,000 copies, SINEs make up about 11% of the human genome.[2] The density of SINEs in the human genome is high and on average a SINE occurs at every 10Kb of DNA in human genome. Sometimes these copies carry part of the cellular DNA with them, rearranging the genome. While historically viewed as "junk DNA", recent research suggests that in some rare cases both LINEs and SINEs were incorporated into novel genes, so as to evolve new functionality. Sometimes they affect the function of neighboring genes just through their presence. For example, they can interfere with regulatory sequences. Equine SINE elements are called ERE1, ERE2 etc. ERE is abbreviation of Equine Repetitive Element and the average length of them is 230 bp. While historically viewed as "junk DNA", recent research suggests that in some rare cases both LINEs and SINEs were incorporated into novel genes, so as to evolve new functionality. Many microsatellites are closely associated with and generated by retrotransposons like LINEs and SINEs (Arcot *et al*. 1995; Nadir *et al*. 1996; Wilder and Hollocher 2001; Yandava *et al*. 1997).
SINEs are transmitted vertically but not horizontally, and the probability of independent emergence of the same SINE families in unrelated species is negligible. The equine SINE family (ERE-1) has several feature characteristics of tRNA-derived retroposons. Therefore equine SINEs may have originated from $tRNA^{ser}$.[10] Many of these repeats were generated through the activity of transposable elements or transposons that can insert their copies into new chromosomal locations. SINEs on the other hand, co-opt the LINE machinery and function as nonautonomous retroelements.

1.3 The life cycle of retrovirus

ERVs are generally only infectious for a short time after integration as they acquire many inactivating mutations during host DNA replication. The replication cycle (Figure 3) begins with the binding of surface protein (SU) to one of the cell receptors which forces transmembrane protein (TM) to come in a close contact with the cell membrane. Following the cell and virus fusion, the virion core is released into cytoplasm where the genome of the retrovirus is reverse transcribed into a double-stranded DNA clone using the ssRNA like a template. Reverse transcriptase has a high error rate because the enzyme

is highly error-prone, and it makes many mistakes in copying viral RNA into DNA. The preintegration complex (PIC) including the retroviral DNA and integrase (IN) along with some cellular factors is formed, transported into the nucleus and subsequently integrated into the host's genome by the viral encoded IN. Integrated viral DNA is called provirus. Retroviral transcription starts at the 5' U3-R junction and the 3' polyadenylation site if placed at the 3' R-U5 junction. The major splice donor site downstream of the primer binding site (PBS) is used for generation of subgenomic mRNAs, including env. After translation, the polyproteins Gag and Gag-Pro-Pol localize to the cell membrane into which the Env protein attaches by its C-terminal transmembrane domain.[2] As the virion matures, the polyproteins are cleaved into functional subunits within the capsid and thus the infection continues spreading to neighbouring cells by budding of the virion from the cell surface.

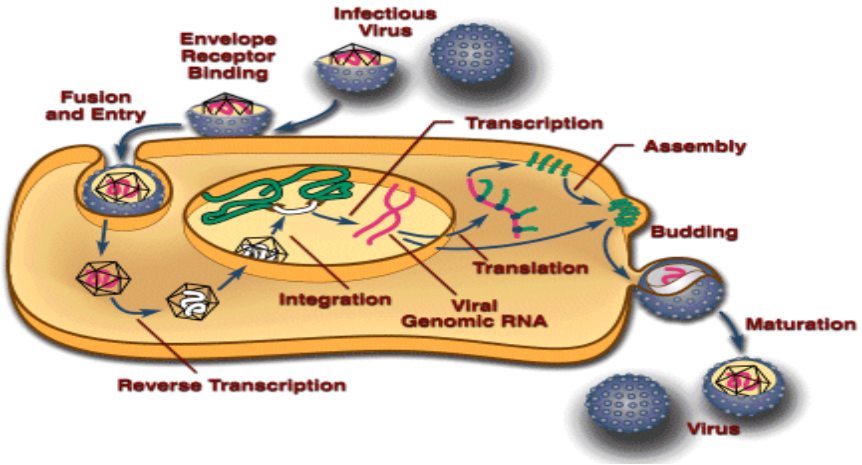


Figure 3: Life-cycle of the retrovirus
Adapted from "Methods in Cell biology Vol.52 Charles P, Emerson, H. Lee Sweeney 1997, page 180"

1.4 Estimation of the ERV age

Estimations of the ERV age are based on the calculated LTR divergence because LTRs were identical at the time of integration. LTRs have been diverged as they acquire many random mutations during the evolutionary time. A neutral nucleotide substitution rate is 0.2% per million years (my) for retrotransposons. For example: 5% nucleotide sequence divergence between two LTRs would represent an integration that occurred around 12.5 million years ago (mya) whereas an integration with a 10% divergence would have occurred around 25 mya. Up to date, ERVs older than 125 mya cannot be found in current genomes

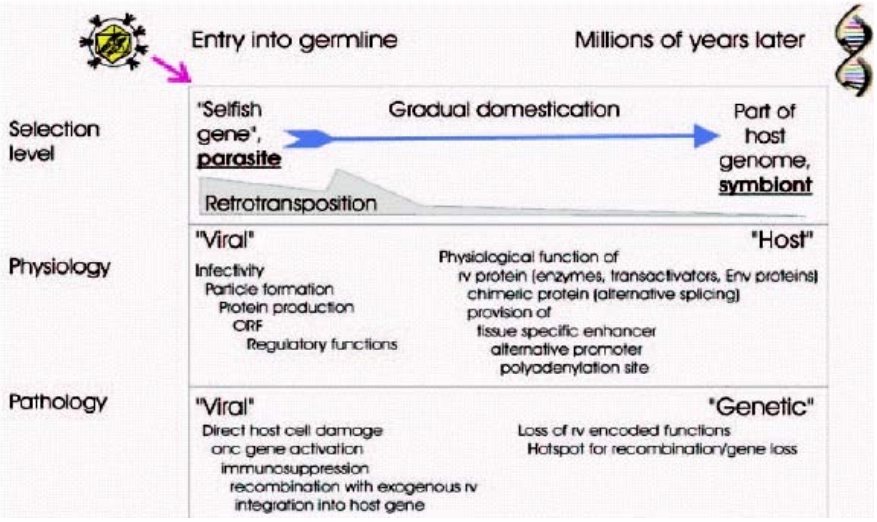


Figure 4. *Events following endogenization of a retrovirus.*[4]
adapted from "Evolutionary Aspects of Human Endogenous Retroviral Sequences (HERVs) and Disease Jonas Blomberg, Dmitrijs Ushameckis, and Patric Jern. Madame Curie Bioscience Database 2000-."

Figure 4 shows that gradual domestication of retrovirus during the evolutionary time. When an exogenous retrovirus infected the host it may be considered as selfish gene but after thousands of years of co-evolution with the host the virus became part of host genome. They have changed their life type from parasitic to symbiotic type.

1.5 The amount of ERVs vary in different mammals

Retroviruses have challenged and infected all orders of eukaryotic life during evolution. Endogenous retroviruses in different species have the same genomic organizations. The first ERVs were identified in chicken, mice and cats. In mouse, some retroviruses including Mouse Mammary Tumour Virus (MMTV, beta retrovirus) and Murine Leukaemia Virus (MLV gamma retrovirus) are known to cause diseases. The ERV analysis of the different species supports the notion that different mammals interact distinctively with endogenous retroviruses. The equine genome has been effective in protection from extensive retroviruses integration. The amount of equine ERVs was approximately same with dog's ERVs which Martinez Barrio et al (2011) have found.

Table 1. *Estimated presence of ERVs in different organisms.*[5]

Species	Genome Assembly	Chains present	Genome percentage	Assembly depth of coverage*
Zebrafish (*Danio rerio*)	danRer5/4/3	2048	0.8%	6.5–7x
Red jungle fowl (*Gallus gallus*)	gg01	260	0.2%	6.6x
Opposum (*Monodelphis domestica*)	monDom5/4/1	7456	~2%	7.33x
Dog (*Canis familiaris*)	canFam2	407	<0.15%	7.5x
Mouse (*Mus musculus*)	mm9/8/7	7582	~2%	7.7x
Rhesus macaque (*Macaca mulatta*)	Mmul_1	2690**	<0.8%	5.1x
Chimpanzee (*Pan troglodytes*)	panTro1/2	2919**	<0.8%	4–6x
Human (*Homo sapiens*)	hg16	3149	0.8%	4–5x

Adapted from "Martínez Barrio Á, Ekerljung M, Jern P, Benachenhou F, Sperber GO, et al. (2011) The First Sequenced Carnivore Genome Shows Complex Host-Endogenous Retrovirus Relationships. PLoS ONE 6(5): e19832. doi:10.1371/journal.pone.0019832"

1.6 The nature of integrations related with host genes.

Retroviral LTRs which are 500-600 nucleotides long contain strong transcriptional regulatory elements such as compact, mobile promoter, enhancer sequences hormone responsive elements, and polyadenylation signals that may alter the expression of cellular genes adjacent to integrated proviruses. LTRs are commonly found upstream of genes in antisense orientation or downstream in sense orientation[6]. One of the most well-known cases of tissue-specific promoter is the expression of amylase in the human salivary glands by an integration of HERV-E in reverse orientation upstream of the gene[8]. LTRs are underrepresented within and in the vicinity of genes [9].

Integrated proviruses may activate cellular gene expression either in somatic cells or following germ-line infection. Effects on cellular gene expression following retroviral integration into somatic cells will be detected only if there is a resultant phenotype that offers a selective advantage to the cell. Most commonly, this has been detected as increased cellular growth associated with oncogenes. Germ line integration can result in more subtle changes in gene expression, such as the development of new mechanisms of tissue-specific gene regulation.[6]

Retroviruses make use of cellular machinery that was intended for other purposes. Retroviruses are known to use 2 different mechanisms to accomplish the transport of unspliced or partially spliced RNA. At least some simple retroviruses have a structure in their RNAs-the constitutive transport element (CTE)-that interacts with cell machinery to allow export of a fraction of the unspliced RNA.

However retroviruses not always play a destructive mechanism in the genome organization.

Syncytin is the best known example of co-adaptation between viruses and the host. Syncytin mediates placental cytotrophoblast fusion *in vivo*, and thus plays an important role in human placental morphogenesis.[7] Host species have co-evolved with ERVs over millions of years and developed multiple defence mechanisms such as co-suppression, CpG methylation and cytidine deamination. (reviewed in Jern and Coffin 2009) Host-retrovirus interactions influence the genomic landscape and have contributed substantially to mammalian genome evolution.

1.7 Classification of Endogenous retroviruses

ERV classification and grouping originally was based on sequence similarity between the proviral PBS and the host tRNA [6]. However, it is inconsistent for many other ERV groups that have alternative PBSes [12].

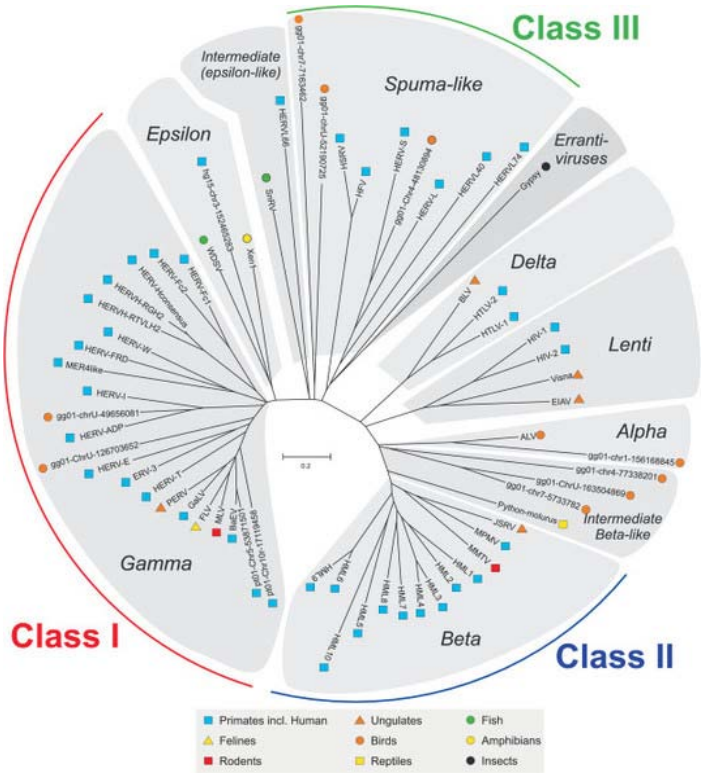


Figure 18. General tree of retroviruses based on the Pol proteins of exogenous and endogenous retroviral sequences. (Jern et al., 2005) [12]

Adapted from "Genomic Variation and Evolution of HERV-H and other Endogenous Retroviruses (ERVs) Jern P, 2005"

The classification of ERV using Retrotector© tool 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 and its large size (800–1100 aa) provides adequate information for a relatively detailed classification. (Jern et al., 2005) This is facilitated by the program Retrotector©© [Sperber G.O. et al], which reconstructs probable Pol proteins ("puteins") from different reading frames in the often damaged gene candidates.

Retrotector© uses its own motif bases to classify ERVs. Retroviruses are classified into 7 groups.

Endogenous retroviruses are not formally included in this classification system, and are broadly classified into three classes, on the basis of relatedness to exogenous genera:

- Class I are most similar to the gammaretroviruses
- Class II are most similar to the betaretroviruses and alpharetroviruses
- Class III are most similar to the spumaviruses

Simple retroviruses are only maintaining the most essential genes, *gag*, *pro*, *pol* and *env*, coding for the virion proteins. Beta, Gamma retroviruses belong to simple organization. Some retroviruses have additional genes. These endogenous retroviruses are grouped as complex retroviruses. For example: Equine Foamy virus (Spumavirus) and Equine infectious anemia virus (Lentivirus) belong to complex

organization The complex retroviruses have evolved to use accessory genes and their gene products. The accessory genes are overlapping with the essential genes, although usually in different reading frames, thus keeping the retroviral genome compact (Petropoulos, 1997; Vogt, 1997a; Vogt, 1997b).

2 OBJECTIVES OF THE STUDY

The main objective of our research was to identify endogenous retroviruses in horse genome using experimental approaches and bioinformatics tools then classify them into groups based on pol region and primer binding site (PBS).
In order to accomplish the main objective we pursued the following tasks.

- To develop an unbiased SINE-PCR approach that uses SINE element as a template to find out new ERVs
- To execute Pan-pol PCR which has been successfully used to find novel ERVs in the other species
- To execute bioinformatics analysis using LTR_STRUC tool and Retrotector© online tool

During the study new tasks emerged from the results.

- To check the polymorphism of EcERV-Beta1 in 13 different breeds.
- To determine the relations with host gene or neighbour genes of the candidate ERVs by positioning them on the horse genome.

3 MATERIALS AND METHODS

3 Mongolian and 2 Thoroughbred horse samples were used in SINE-PCR and Pan-PCRs. Genomic sequence of Thoroughbred horse, Twilight was used for bioinformatics part. 26 samples from 13 different breeds were used for testing polymerases of EcERVs as shown on Table 2.

Table 2. Samples of 13 different breeds

Number from each breed	Breed	Type	Breed origin
3	Standardbred	Trotting race horse	Sweden
2	North Swedish Draft	Draft	Sweden
1	Morgan Horse	Leisure	USA
2	Swedish Warmblood	Sport horse	Sweden
1	Swedish Ardenne	Draft	Belgium
2	Gotland Pony	Pony	Sweden
2	Shetland Pony	Pony	UK
2	Connemara	Pony	Ireland
3	Icelandic Horse	Riding, gaited	Pony size
2	Welsh Mountain	Pony	UK

	Pony		
2	Knabstruber	Riding/Light draft	Denmark
2	Faeroe Pony	Leisure	Faeroe Islands
2	Thoroughbred	Race horse	UK

We used bioinformatics and experimental approaches to find endogenous retrovirouses from horse genome.

3.1 Experimental approach

In order to find new ERVs we have used an unbiased SINE-PCR approach. And the following PCR approaches have been performed for identifying novel EcERVs and testing a polymorphism of EcERV beta1.

3.1.1 SINE-PCR[30]

SINE elements are abundant and widespread in all chromosomes of equine genome therefore we assumed that SINEs can be used as templates for finding novel ERVs because it is likely that several ERVs are located in the vicinity of the SINEs. The first step was to find the nearest SINE element of EcERV Beta1. Degenerate primers were designed from the conserved region of all known horse SINE elements using the multiple alignments of all equine repetitive elements which were archived in GIRI database.

Forward primer was designed as degenerate primer and targeted on 10 K bp region upstream of the EcERV Beta1. (Figure 6.)

Forward primer: CCRGBGTTTCGYTGGTTCV Tm=60.1 C. 19bp

where R stands for A or G; B stands for G, T or C; Y stands for C or T; V stands for G, A or C;

Reverse primer: CTAGAGAGGGGCAAAAACTTCTC Tm=59.9 C. 23bp

Reverse primer was designed on the PBS-flanking region of full-length ERV. Degeneracy rate is 2*3*2*3=36 times. We have predicted a SINE element in 10 k bp based and the nearest SINE element was found in 5.5 k bp upstream of 5‘ EcERV-beta1. The second step was to clone and separate PCR products since we used degenerate primers. We cloned and sequenced them to determine locations of the SINEs on the whole genome sequence by BLAST.

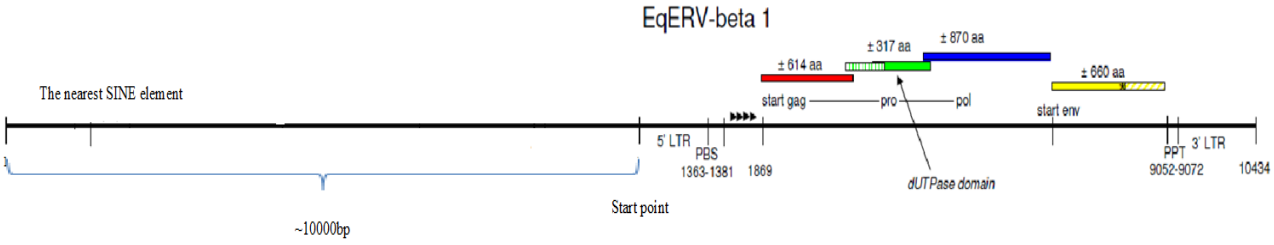


Figure 5. The expected nearest SINE element of EcERV beta1.

Touch down PCR (Table 3) was performed with 120 ng of genomic DNAwith standard reagents and 2.5 U of Ampli taq Gold for 40 cycles.

Table 3. Cycle parameters of Touch down PCR amplification

	Temperature	Time	Cycles
Initial denaturation	95°C	4 min	1X
Denaturation	95°C	30 sec	
Annealing	50°C	30 sec	20X
Elongation	72°C	90 sec	
Denaturation	95°C	30 sec	
Annealing	50°C	30 sec	20X
Elongation	72°C	90+20 sec cycle elongation for each successive cycle	
Final Elongation	72°C	10 min	1X
Cooling	4°C	unlimited time	

3.1.2 Pan-pol PCR amplification

Pan-pol[31] PCR was used to amplify 640 bp of conserved pol region. This approach has been successfully used on avian, baboon, human and swine genomes previously. The elongation was extended by 20 seconds in every successive cycle in Pan-pol PCR amplification as a Touch down PCR. And the annealing temperature was 45°C. It decreases PCR stringency and allowed us to separate a right band.

The degenerate oligonucleotides were used:

5‘ MOP-2 (5‘-CCWTGGAATACTCCYRTWTT-3‘)

3‘ MOP-2 (5‘-GTCKGAACCAATTWATATYYCC-3‘),

where R stands for A or G, Y stands for C or T, K stands for G or T, and W stands for A or T.
PCR products were cloned and sequenced.

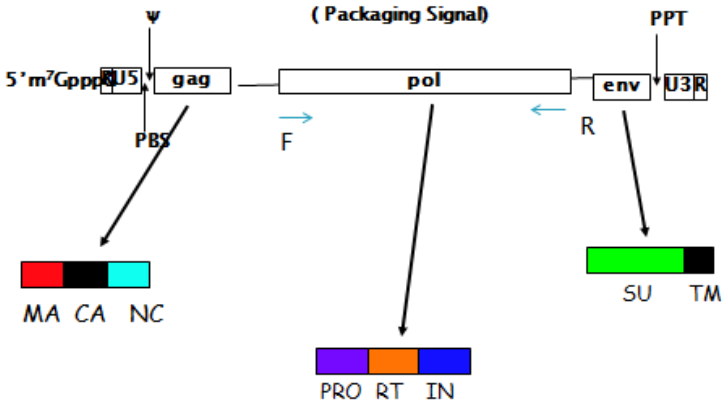


Figure 6. Approach to amplify pol region.

Degenerate PCR primer

All known horse SINE element families were aligned, which have been archived in Genetic Information Research Institute, using ClustalW. Degenerate primers were designed for finding the nearest SINE element in the vicinity of the first beta endogenous retrovirus using multiple alignment of ERE1 elements which are the most widespread horse SINE elements. Domestic horse (Equus caballus) has $5*10^4$ copies of ERE1 SINE elements.[2]

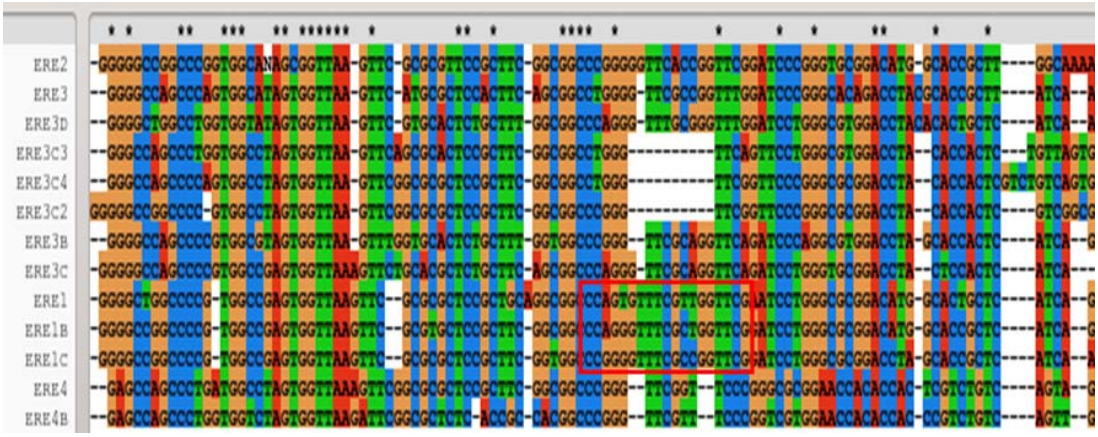


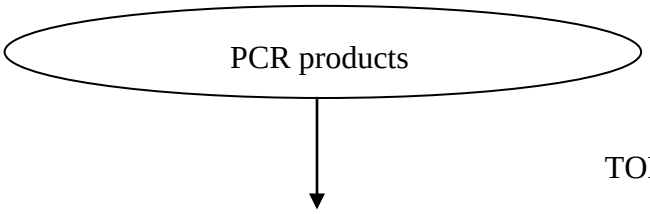
Figure 7. The multiple alignment of the Equine repetitive elements (ERE)

It was difficult to find a region to design a primer with suitable melting temperature in such short sequence with many gaps indeed. It does not include conserved region much but it was the only place which was adequate for primer requirements. Degenerate PCR has proven to be a very powerful tool to find "new" genes or gene families. Most genes come in families which share structural similarities. By aligning the sequences from a number of related SINE elements we have determined which parts are conserved and which are variable. Degenerate primers were designed for finding the nearest SINE element.

Cloning of PCR products:

SINE and Pan pol PCR products were both cloned and sequenced. SINE-PCR product was 5.5 kb.

Pan-pol PCR with degenerate primers corresponding to conserved regions of known retrovirus pol genes was performed and yielded products of the expected size of approximately 640 bp. The products of the expected size were cut by scalpel from the 0.8% agarose gel after gel electroporation and gel purified using SNAP Mini prep Kit. After that the products were ligated with pCR2.1-TOPO vector. TOPO TA Cloning Kit for sequencing was used according to the protocol.



TOPO®-Cloning Reaction:

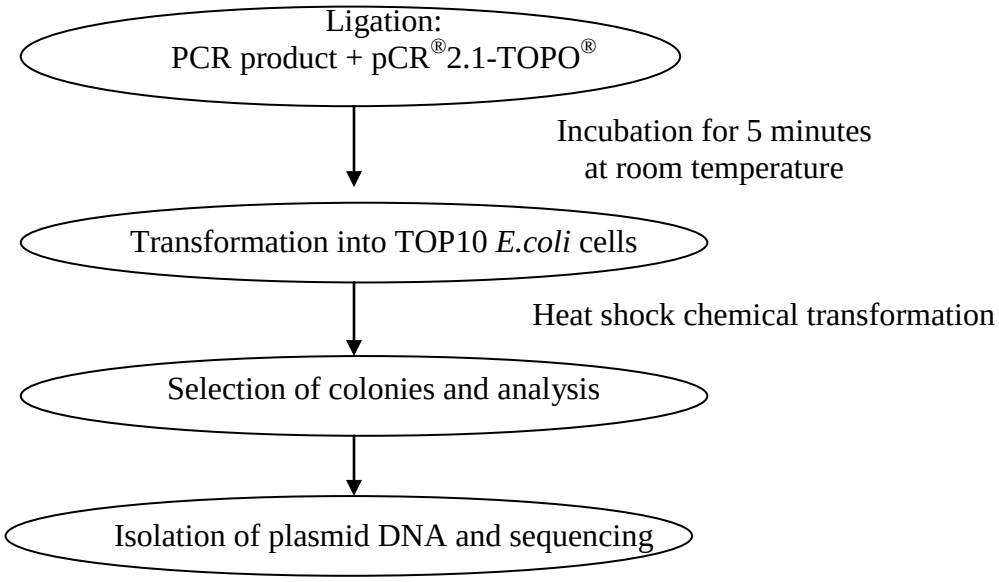


Figure 8. Experimental outline for cloning and sequencing

Ligated DNA samples of Pan-pol PCR and SINE PCR products have been transformed to the TOP10 E.coli cells and been grown on the medium overnight. 3 colonies from each dish, total 15 colonies have been cultured on LB medium at 37°C incubator with shaker overnight. Plasmid DNA were isolated using S.N.A.P™ MiniPrep Kit. After these steps, Plasmid DNAs were prepared successfully for sequencing

Sequencing of Cloned plasmid inserts

M13forward and M13 reverse primers were used to generate a nucleotide sequence of the DNA insert cloned into pCR-TOPO2.1.

M13 Forward (−20) 5´-GTAAAACGACGGCCAG-3´
M13 Reverse 5´-CAGGAAACAGCTATGAC-3´.

Plasmids DNAs were directly sequenced by BigDye® Direct Cycle Sequencing Kit according to the protocol. Since plasmid has its own M13 tails PCR amplification was not required before sequencing.

3.1.3 Polymorphism of EcERV Beta1 region between 13 different breeds

3360 bp segment has been amplified which includes whole pol region. It was possible to analyze polymorphism between breeds. For that purpose 13 different horse breeds were used.

Left primer: GTCTCAAGCCTCCTTCGAGC
Start: 3641 Length: 20 bp Tm: 60.5 C GC: 60.0 %
Right primer: TCCACAAAGGAGAGGAAGCG
Start: 7000 Length: 20 bp Tm: 59.7 C GC: 55.0 %

The LTR divergence of the first beta ERV is 1% which is relatively recent integration. LTR divergence is the crucial factor for polymorphism. 25 samples from 13 different breeds were used for this purpose. Long range PCR amplification was used for pol region.

Table 4. Cycle parameters of Long range PCR

	Temperature	Time	Cycles
Initial denaturation	94°C	2 min	1
Denaturation	94°C	10 sec	
Annealing	60°C	30 sec	10
Elongation	68°C	8 min	
Denaturation	94°C	10 sec	
Annealing	60°C	30 sec	25
Elongation	68°C	8 min+20 sec cycle elongation for each successive cycle	
Final Elongation	68°C	7 min	1
Cooling	4°C	unlimited time	

3.2 Bioinformatics approach

There are several bioinformatics approaches for finding endogenous retroviruses from mammalian genome. For example: Retrotector©, LTR_STRUC, RepeatMasker, Retrosearch (for human genome only), HERVd (for human genome only) etc. These tools are based on four different approaches: repeat finding methods, homology-based methods, structure-based methods, and comparative genomic methods.[18] Each tool has strengths and weaknesses, and the best results are obtained by using a combination of them. [21]

We have used LTR_STRUC and Retrotector© online tool because they are available tools to detect ERVs not only from human genome but also from genomes of other species.

3.2.1 LTR_STRUC

LTR_STRUC was the main tool on bioinformatics part of the study and a limit was set for elements at >0.3 score. The latest available version of the horse genome is EquCab2 sequence which is assembled by Whole Genome Shotgun (WGS) sequencing in September of 2007. A female thoroughbred named "Twilight" was selected as the representative horse for genome sequencing. Horse genome is 2.5-2.7 Gbp [24] and it is somewhat larger than the dog genome (2.5 G bp) and smaller than human and bovine genome (2.9 Gbp). Repetitions were sorted out and excluded from further analysis. For scrutinizing results of LTR_STRUC tool as ERV candidates, Retrotector© online tool was used.

LTR_STRUC has been used in bovine genome by Koldo Garcia-Etxebarria et al. [13] and it was proven as powerful tool among different LTR based approaches. We have used it as a main tool for mining all horse chromosomes.
LTR_STRUC tool is invented by McCarthy, E.M., and J.F. McDonald at the Department of Genetics, University of Georgia, Athens, USA in 2003. [20]
LTR_STRUC retrieves endogenous retroviruses (ERV) and generates report files (in text format) only for hits generating a score in excess of the cutoff score. These files contain a detailed analysis of each hit. They include all the information enumerated in Table 2.

Table 5. Information in LTR STRUC output files

1. Name of source contig,
2. Location of element within contig,
3. Score for current hit,
4. Lengths of contig, element, LTRs and largest ORF,
5. Nucleotide sequences for the whole transposon, TSRs,
6. LTRs, PBS, PPT, dinucleotides terminating the LTRs,
7. orientation of the transposon within the contig, (determined by relative positions of PBS and PPT),
8. Sequences for all ORFs (longer than 50 amino acids),
9. Intra-element percent identify of LTRs.
10. An alignment of the putative LTRs

The LTR STRUC is written in Visual C++ (Microsoft version 6.0) and runs on PC platforms. The search algorithm used by LTR STRUC seeks certain generic structural features of retroelements. It relies on structure of LTR ends and characteristics.
Retroviruses are difficult to detect using sequence similarity, because they are diverse, having only small portions. of their genomes in common. For example, it is estimated that there are 10^{60} variants of HIV[17]. One explanation for this variability is that when RNA is converted to DNA using reverse transcriptase, there is no error correction such as there is when DNA is copied. Also, due to a lack of selective pressure, many endogenous retroviruses are heavily mutated to the point of being defective or even completely non-functional. These defective retroviruses are still of interest, however, because of their past influence on the genome, their value as molecular fossils, and because they can function with the help of other retroviruses.[18]
An alternative to detection using sequence similarity is detection based on structure.
Retrovirus genomes have a consistent structure. They range in size from 5000 to 20,000 nucleotides.

- A typical LTR retrotransposon has a structure called TG..CA box on both side of the chain, with TG at the 5’ extremity of 5’ LTR and CA at the 3' extremity of 3'LTR.
- TSR Region: TSR (target site repeat) is a 4~6 bp short direct repeat string flanking the 5' and 3' extremities of an element. It is the sign of insertion of transposable elements.
- PBS: Near 3' end of the 5'LTR, there is a ~18bp sequence complemented to the 3' tail of some tRNA. The site is very important because tRNA binding process is first step of initiating reverse transcription.

- PPT: Polypurine tract is a short rich purine segment, about 11~15 bp in length. Like PBS, this region is important for reverse transcription.
- Protein domains: In a typical virus genome there are three polygenes: gag, pol and env. Among them, pol is most conserved. Inside pol there are three important domains: IN (integrase), RT (reverse transcriptase) and RH (RNase H), which are enzymes for reverse transcription and insertion. RT and IN are regarded essential for autonomous LTR elements to fulfill their function.

These signals may become blur or even undetectable for evolutionary events.
Parameter set was configured when chromosomes were examining. The genomic material was a Thoroughbred mare's DNA sample.

3.2.2 NCBI-BLAST

NCBI-BLAST search for endogenous retrovirus was used to double-check the candidates of ERVs.[25,26]
BLAT (The BLAST-like Alignment Tool) searches through the Horse (*Equus caballus*) Genome Browser Gateway of the Genome Bioinformatics Group of UC Santa Cruz.[27]

3.2.3 Retrotector©

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.
The program package Retrotector© (formerly RetroSpector) is designed to identify and characterize entire or fragmented endogenous retroviruses (ERVs) in genomic material, in a fashion robust to mutations and with considerable flexibility.
The program is written in Java and quite portable. It is in use under Windows, MacOS X and Linux.
For Retrotector© three types of algorithms have been developed:
"Fragment threading": whereby characteristic motifs are combined into chains, satisfying distance criteria.

A fast dynamic programming: Needleman-Wunsch type algorithm for checking similarity between two DNA base sequences.

A dynamic programming: Needleman-Wunsch type algorithm for fitting an amino acid sequence to a DNA base sequence, taking into account known related peptides and other factors suggesting the preferred reading frame. [19]
Retrotector© was used to characterize these LTR_STRUC results.

4 RESULTS

4.2 Experimental results:

The following results were obtained on experimental part during the study.

4.1.1 SINE-PCR results

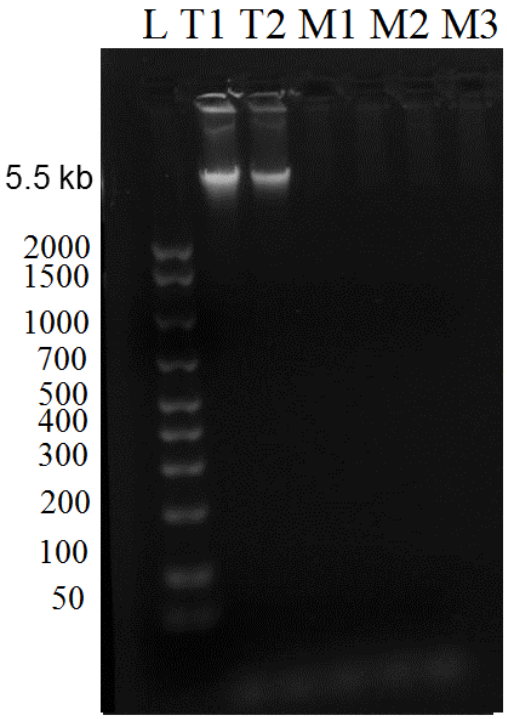


Figure 10. SINE-PCR result. L is ladder between 50-2000 bp. T represents thoroughbred horses. M represents Mongolian horse sample. Touch down PCR was used and the nearest SINE element has been found in 5.5 kb far from the ERV. It was quite large fragment between SINE element and EcERV Beta1. By sequencing this fragment we can find the location of the SINE element.

Mongolian horse samples could be fragmented because we could not get results on the samples. These 3 Mongolian horse samples were tested by microsatellite primers and M2 and M3 have been amplified but not M1. Therefore M1 has been proven that it has degraded. Although M2 and M3 have DNA however the longer pieces of DNA could be fragmented. That could be a reason. Or it could be due to a polymorphism between different breeds.

4.1.2 Pan-pol PCR results

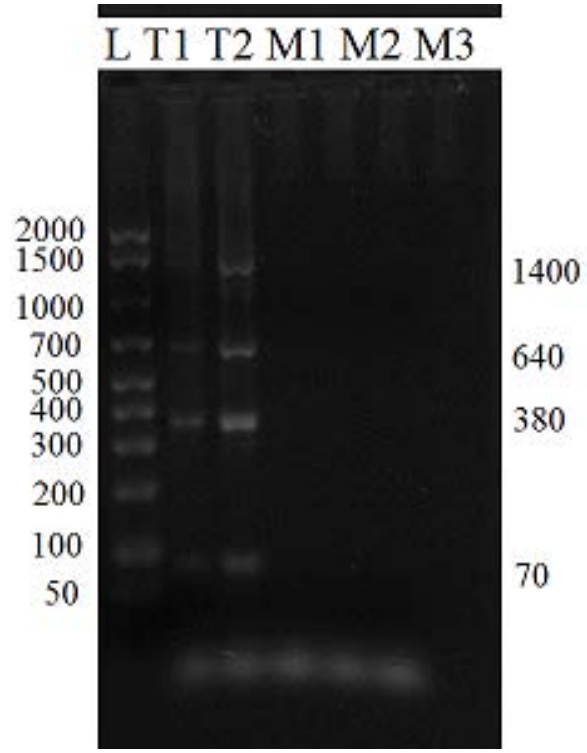


Figure 12. Pan-PCR results. L is ladder between 50-2000 bp. T represents thoroughbred horses. M represents Mongolian horse sample.

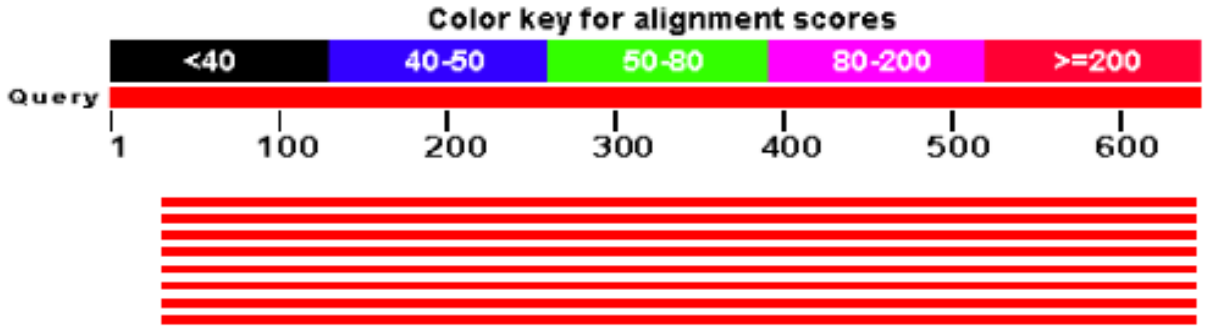
Touch down PCR was used and annealing temperature was at 45^{o}C. 10 Mongolian horse hair root samples were extracted according to routine lab protocol. But 260/280 ratio was low.

Mongolian horse samples were tested by microsatellite primers and they amplified well except M1. Perhaps the negative result on Mongolian horse DNA is because the DNA is degraded and consequently does not amplify the longer product. Or it could be due to the polymorphism of different breeds.

4.1.3 Sequencing results:

Pan-pol sequence:

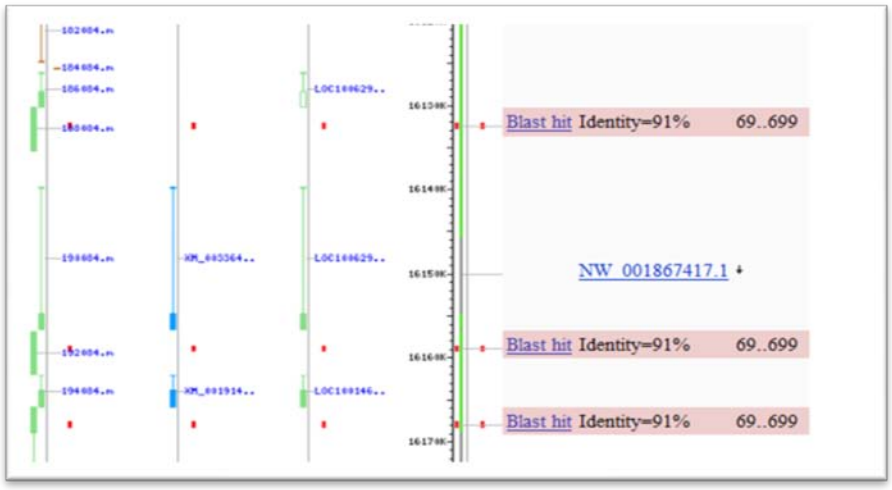
Pan-pol PCR products were cloned and sequenced. These sequencing products were blasted with equine genome by NCBI-BLAST.



Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
NW_001875267.1	Equus caballus breed thoroughbred unplaced genomic scaffold, EquCab	1109	1109	94%	0.0	99%	
NW_001877028.1	Equus caballus breed thoroughbred unplaced genomic scaffold, EquCab	1103	1103	94%	0.0	99%	
NW_001868414.1	Equus caballus breed thoroughbred unplaced genomic scaffold, EquCab	1092	1092	94%	0.0	99%	
NW_001867864.1	Equus caballus breed thoroughbred unplaced genomic scaffold, EquCab	1044	1044	94%	0.0	97%	
NW_001867490.1	Equus caballus breed thoroughbred unplaced genomic scaffold, EquCab	965	965	94%	0.0	95%	
NW_001876907.1	Equus caballus breed thoroughbred unplaced genomic scaffold, EquCab	959	959	94%	0.0	95%	
NW_001867534.1	Equus caballus breed thoroughbred unplaced genomic scaffold, EquCab	928	928	94%	0.0	94%	
NW_001867417.1	Equus caballus breed thoroughbred chromosome 5 genomic scaffold, Eq	856	2557	94%	0.0	92%	

7 endogenous retroviruses have been found from unplaced genomic scaffold and 3 ERVs from chromosome number 5 by sequence of Pan PCR products. These 640 bp products are overlapped with pol regions of EcERVs of chromosome 5. (Sequence was enclosed in Appendix 3)



SINE-sequence:

SINE-PCR products were sequenced but we sequenced 5.5 kb region between known beta retrovirus and its nearest SINE element instead of 230 bp SINE element. In order to find other ERVs the nearest SINE element of EcERV beta1 needs to be re-cloned and sequenced.

4.1.4 Polymorphism of EcERV Beta1

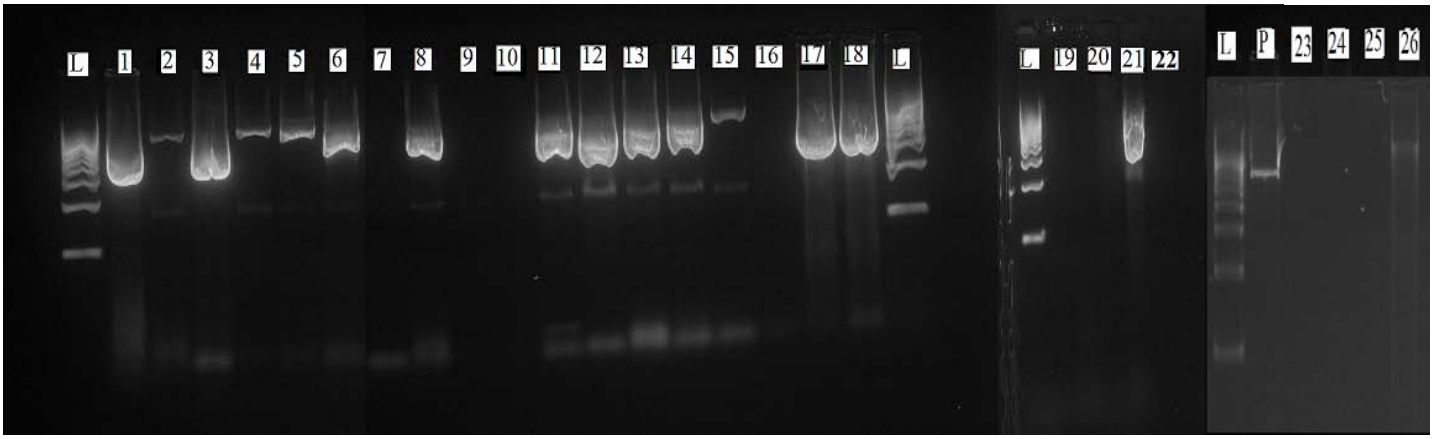


Figure 13 . Polymorphism of EcERV Beta1
L is 500 -5000 bp ladder, 1,3-Thoroughbred horse1, 2,11,-North Swedish warmblood , 4,8-,10 Standardbred, 5,26-Shetland pony, 6-Morgan horse, 7,16-Gotland pony (Sweden), 9,19-Icelandic horse, 12,21-North Swedish draft, 13-Swedish Ardenne, 14, 15-Connemara pony, 17,18-Faeroe pony, 20-202, 22, 23-Knabstruber, 24, 25- Welsh pony, P is a sample of thoroughbred horse that was used as a positive control in the second PCR

From above figure we can see that the expected region was not amplified at the same size. And some breeds do not contain this EcERV Beta1 retrovirus or they have accumulated mutations on that region. Therefore we can say that there is a polymorphism between breeds. Some of them have 2 products and they could be possible candidates. There is definitely a polymorphism in different breeds because although some samples do not have the expected band but they have primer dimer in the bottom which can prove the PCR was performed well. Sample number 26 has slightly larger fragment than the positive control . The insertion could be the cause of different size of the products.
Thoroughbred horses (1 and 3) have same products. Standardbreds (4 and 8) have 2 products but number 10 was not amplified. The reason of sample 10 could be bad DNA quality or fragmented DNA. Morgan (6) horse has the band. Shetland pony has the band (5 and 26) Connemara pony (14 and 15) Swedish Ardenne (13) has the ERV, Faeroe pony (17 and 18) have the ERV. Swedish draft (12 and 21) has the ERV, North Swedish warmblood (2) and Swedish Warmblood (11) have the ERV, Knabstruber does not have (22 and 23) Icelandic horses do not have this product (9,19). Gotland pony does not have (7 and 16) 202 (20) does not have. Welsh pony does not have (24 and 25). This polymorphism could be due to the geographic of different breed's distribution.

4.2 Bioinformatics results

Main result of the bioinformatics approach was that 27 complete and novel ERVs were found.

4.2.1 LTR_STRUC results

LTR_STRUC denotes all retrieved chains and putein sequences unique ID numbers but some of them are copies of each other. A total of 276 unique EcERVs were identified and every calculation and analysis on bioinformatics part were based on these 276 selected endogenous retrovirus sequences from LTR_STRUC, used Twilight thoroughbred mare genome by LTR_STRUC©. The average EcERV is 8.3 kb long and the amount of 276 EcERV (2299577 bp) is 0, 085 % (based on the 276 chains real lengths) in the horse genome that consist of 2.68 Gb.

Score

Each hit is assigned a score which depends on the degree of fulfillment of the chains from LTR_STRUC. Significance level of the score was 0.3. There were 276 unique (based on LTR position) *Equus caballus* endogenous retrovirus elements with scores > 0.3 detected by LTR_STRUC. The chains were ordered in two groups, one group with scores ranging between 0.75 and 2. In the other group the chains with scores ranging between 0.3 and 0.75. The number of ECERV elements with scores over 0.75 is 53 (19.2 %). The group with chains less than 0.75 score consists of chains from 0.3 to 0.75 score and the number of chains here is 223.

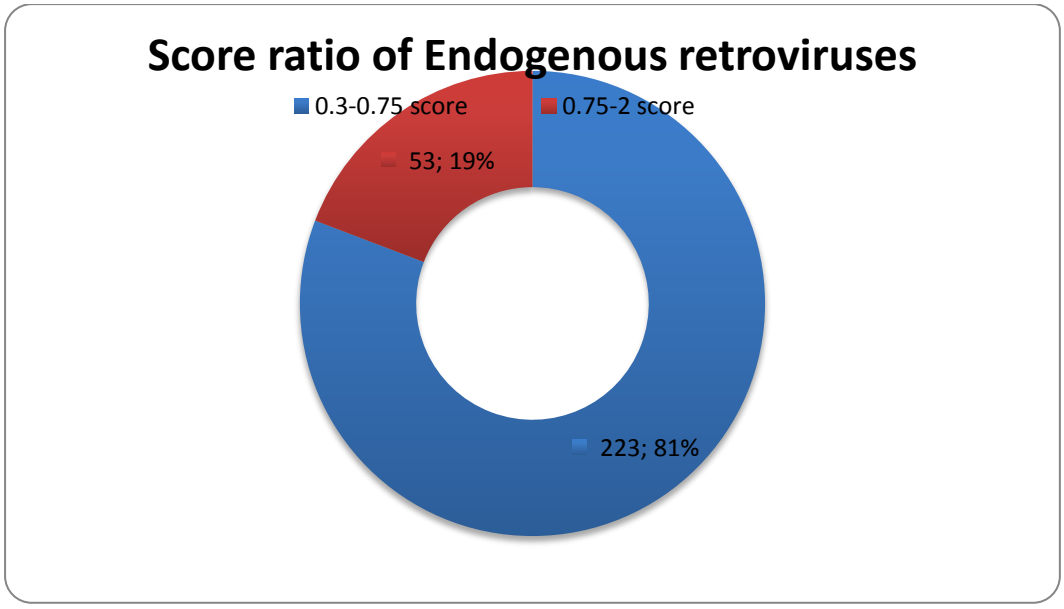


Figure 14. Almost 4/5 of Equus caballus endogenous retroviruses consist of element with scores between 0.3 – 0.75 and approximately a fifth of the elements are with scores between 0.75-2.

chr120000_RT2_B2_L2_2 on chromosome 1 and chr1820000_RT1_B2_L2_2 on chromosome 18 and chr620000_RT2_B7_L7_7 on chromosome 6, chr520000_RT3_B7_L7_8 on chromosome 5 are the highest scored (the maximum cut-off score, 2). It was proven that the highest scored ones occur more common in the genome by NCBI-BLAST tool. For example chr520000_RT3_B7_L7_8 on chromosome 5 has 99% similarity versions in other chromosomes as 11, 29, 1, 15, 6 etc.

25 complete endogenous retroviruses were discovered while all candidates were examined by Retrotector© online tool. These ERVs were abundantly present on all other chromosomes except chromosome 29 and 31.

Table 6. 62 Novel equine endogenous retroviruses

Chr.	Genus	pro	pol	env	gag	LTR_STRUC score	Overall length of transposons	LTR PAIR HOMOLOGY	RETROTECTOR© SCORE
1	C	-	-	+	+	0.38	7439 bp	97.90%	
1	C	N/D	+	+	N/D	0.61	7261 bp	91.50%	
1	C	+	+	+	+	0.95	8919 bp	98.40%	1182
1	B	N/D	+	+	+	1.1	8098 bp	97.70%	1178
1	C	+	+	+	+	1.74	8620 bp	90.80%	1312
1	B	N/D	+	+	+	2	7232 bp	96.50%	1147
2	C	+	+	+	N/D	0.33	14934 bp	90.50%	
2	C	+	+	N/D	N/D	0.42	7234 bp	90.80%	
2	C	N/D	+	+	+	0.64	8657 bp	83.00%	261
2	B	N/D	+	+	N/D	0.66	6881 bp	96.90%	
2	C	N/D	+	+	+	0.76	8182 bp	99.00%	373
2	C	N/D	+	+	N/D	0.89	7814 bp	99.80%	
4	C 65%	N/D	+	+	N/D	0.67	8675 bp	99.70%	
5	B	N/D	N/D	+	N/D	0.62	5080 bp	95.50%	
5	B	+	+	+	+	0.8	10439 bp	99.00%	2787
5	C	+	+	+	+	1.19	8428 bp	99.60%	1545
5	B	N/D	+	+	+	2	7570 bp	97.80%	1518
6	B	N/D	+	+	+	2	7724 bp	93.10%	1096
7	S	N/D	+	+	N/D	0.4	7391 bp	70.80%	
7	B	+	+	+	+	0.74	7724 bp	95.50%	1765
8	C	N/D	+	N/D	N/D	0.38	17897 bp	88.90%	
8	B	N/D	+	N/D	+	0.59	7000 bp	87.40%	
8	B	+	+	N/D	+	1.1	7076 bp	96.70%	
9	B	+	+	+	+	0.7	7004 bp	96.40%	1056
9	C	+	+	+	+	1	10009 bp	86.80%	655
10	C	N/D	+	N/D	N/D	0.61	7261 bp	91.50%	
10	C	+	+	+	+	1.8	14203 bp	96.00%	985
11	B	N/D	+	+	N/D	0.44	8225 bp	96.70%	
11	C	+	N/D	N/D	N/D	0.79	4853 bp	91.30%	
11	B	N/D	+	+	+	1.22	8247 bp	92.20%	1294
11	C	N/D	+	+	+	1.48	10806 bp	92.10%	590
13	C 58%	N/D	+	N/D	+	0.72	7339 bp	95.60%	
13	C	+	+	+	N/D	0.75	9756 bp	92.10%	
14	B 82%	N/D	+	+	+	0.66	8658 bp	96.70%	
14	C	+	+	?	+	0.95	6288 bp	88.60%	

15	B	N/D	+	+	+	0.74	8413 bp	87.90%	1192
16	C 65%	N/D	+	+	N/D	0.67	8675 bp	99.70%	
17	C	+	+	+	+	1.48	8658 bp	93.20%	1308
18	S 94%	N/D	+	+	+	0.66	6247 bp	95.50%	255
18	B 99%	N/D	+	+	+	2	7724 bp	93.10%	1096
19	S 87%	N/D	+	+	N/D	0.4	7391 bp	70.80%	
19	C 97%	N/D	+	N/D	N/D	0.5	11657 bp	85.70%	
20	C 91%	+	+	?	+	0.58	5611 bp	91.60%	
22	B 83%	N/D	+	+	N/D	0.66	7769 bp	90.80%	
22	C 94%	N/D	+	+	N/D	0.66	7555 bp	97.90%	
22	55% B	N/D	+	N/D	+	0.74	7986 bp	99.00%	
22	C 83%	N/D	+	N/D	+	0.76	6099 bp	91.50%	
24	C 97%	N/D	+	N/D	+	0.75	6326 bp	93.70%	
25	C	+	+	?	+	0.46	13475 bp	83.50%	398
25	D	N/D	+	N/D	+	0.74	8208 bp	94.50%	
25	C 76%	N/D	+	+	N/D	1.32	14964 bp	96.20%	
26	C 96%	N/D	+	N/D	+	0.62	6308 bp	92.00%	
26	B	N/D	+	+	+	0.66	7504 bp	92.70%	293
26	B 93%	+	+	+	+	0.73	6319 bp	82.80%	414
26	C 99%	+	+	?	+	1.15	11790 bp	94.70%	
27	C 86%	N/D	+	+	+	0.66	6780 bp	89.90%	407
28	B 99%	N/D	+	+	+	1.1	8098 bp	97.70%	1178
30	C 99%	+	+	+	+	1.74	8620 bp	90.80%	1312
X	B 66%	N/D	+	+	N/D	0.67	7808 bp	90.10%	
X	C 96%	+	+	+	+	0.93	11220 bp	90.50%	596
X	C	+	+	+	N/D	1	8456 bp	93.50%	963
X	C 76%	N/D	+	+	N/D	1.32	14964 bp	96.20%	

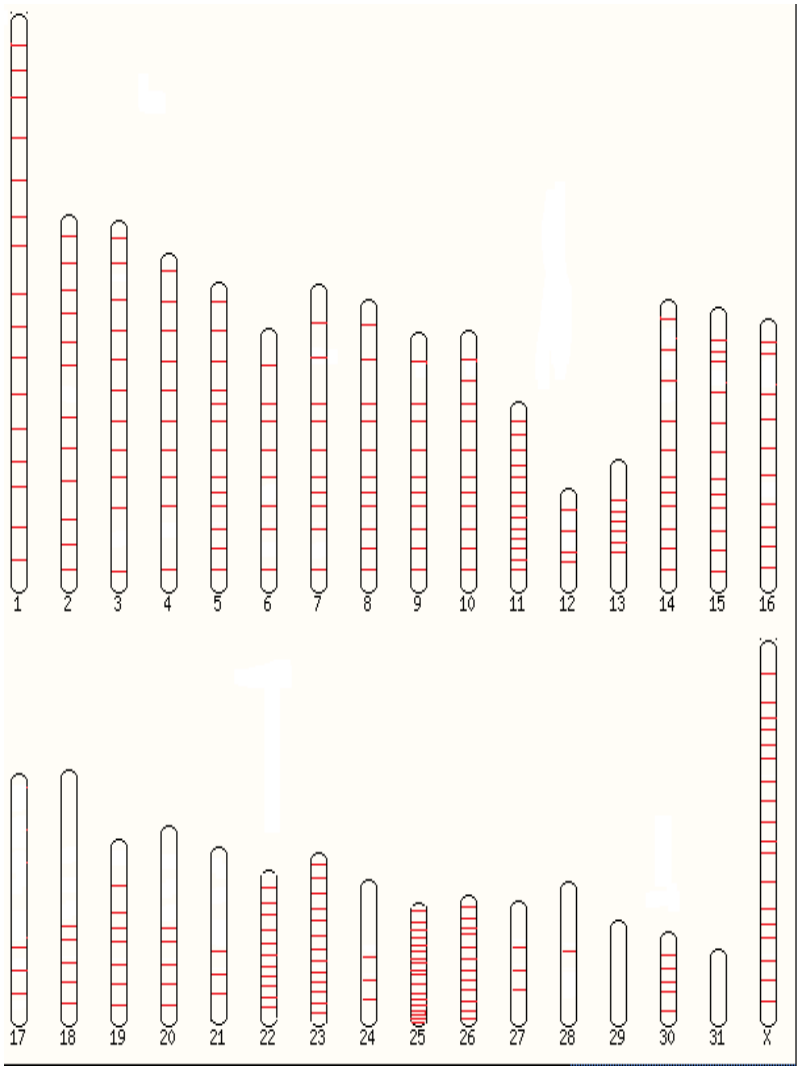
High scored ERVs are tabulated in Table 5. And low scored ERVs are enclosed in Appendix 1. Where C is gamma B is beta and S is spuma ERVs. 27 complete endogenous retroviruses (highlighted in Table 5.) have been revealed from this study using LTR_STRUC tool that are 13 beta, 13 gamma, 1 spuma ERVs. Previously known the first beta retrovirus EcERV Beta1 which is highlighted by blue color has also been found within 27 ERVs and we used it as a positive control in the *in silico* analysis. Retrotector© score was quite high (1007.1) in average among complete 27 ERVs.

4.2.2 Chromosomal distribution

The distribution of equine endogenous retrovirus elements between the equine 32 chromosomes is unequal and there could be some more equine ERV on contigs with unknown chromosomal localization because we searched for equine ERV chains on contigs with all known chromosomal localization. We have found some 95-99% identical variants on contigs with unknown

chromosomal localization when we blasted 27 novel and complete ERVs that we found with equine whole genome sequence using NCBI-BLAST tool.

The largest amount of endogenous retrovirus was identified on chromosome X and 25 which have 18 elements (Fig.). The chromosome with the lowest occurrence of endogenous retroviruses is chromosome 29 and 31 which completely lacked equine ERV . LTR_STRUC could not find equine ERVs from unidentified part of the chromosome. It remains to be confirmed that these chromosomes are essentially lacking equine ERVs or whether it reflects annotation-bias of these chromosomes. For the first 16 and X chromosome the frequency of EcERVs seemed to be correlated with length of the chromosome but the trend has not been shown between chromosomes from 17 to 31.



chromosome	chains
1	16
2	12
3	11
4	10
5	13
6	8
7	10
8	11
9	10
10	11
11	13
12	4
13	6
14	11
15	12
16	10
17	3
18	5
19	7
20	5
21	3
22	11
23	14
24	3
25	18
26	12
27	3
28	1
29	0
30	5
31	0
X	18
Un	0
total	276

Figure 16. Chromosomal distribution of endogenous retroviruses in horse genome

***LTR* divergence and ORF**

The chains were grouped in three groups by LTR- divergence. Most of them (49 %) in group 3, which 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 (29 %) and for the youngest with possible functional elements in group 1 the number of chains is (22 %)

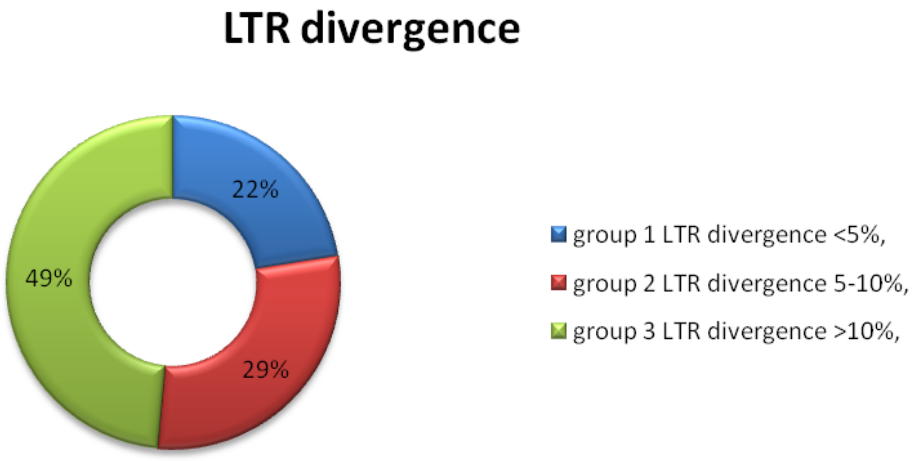


Figure 17. Three groups with different LTR divergence.

4.2.3 Phylogenetic analysis of Equine endogenous retroviruses

Figure 18.Alignment of translated pol sequences of novel horse ERVs on chromosome 7 and 1, EcERV Beta1, beta- and delta retroviruses of different species. The conserved YMDD motif is shown with asterisks. Chr7_7406 and chr120000 are novel and complete ERVs.

Phylogenetic analysis of the translated pol product of equine endogenous retroviruses was performed using the phylogeny option in ClustalW2 program and reference sequences from beta and delta retroviruses. There is a highly conserved YMDD motif surrounded by red line. The highly conserved motif also had some variants as YXDD (where X=M,V, or I) on the other candidates. Novel EcERVs have also conserved this motif and other amino acids as L, GL, K, Q which were in the same position among other species (highlighted with asterisks). Novel EcERVs, found from chromosome 1 and 7

were slightly different from each other. From this alignment we can see the accumulated dot mutations.

```
Chr7.7406       SPTICQQLVGDVLLTF-SKYPTIQLYYMDDILLAAPTKNLSLMAYQQLIELLKGKGLLV 1125
chr11.12        SPTICQQLVGDVLLTFRSKYPTIQPYYMDDILLAAPTKNFSLTAYQQLIELLKGKGLLK 1331
chr111          SPTICQQLVGEVLLTFRSKYPTIQLYYMDDIISSAN--KFFINSTP-TINIAKRQRPAY 1306
chr2811         SPTICQQLVGEVLLTFRSKYPTIQLYYMDDIISSAN--KFFINSTP-TINIAKRQRPAY 1306
chr62000        SPTICQQLVGDVLLTFQRKYPTIQLYYMDDILLAVPTKNFSLTAYQ-LIELLKGKGLLM 1319
chr18.2000      SPTICQQLVGDVLLTFQRKYPTIQLYYMDDILLAVPTKNFSLTAYQ-LIELLKGKGLLM 1319
chr5.2000       SPTICQQLVGDVLLTF-SKYPTIQLYYMDDVLLAAPTKNFSLTAYQQLTELLKGKGLLM 1110
chr12000        SPTICQQLVGDVLLTFRSKYPAIQLYYMDDILLAAPTKNFSLTAYQQLIELLKGKGLLM 1152
chr5.7965       SPTLCQKFVDMALMQVRQKYPSIYLIYMDDLLLAHQDRSYLQQVLQETVEALTLYGLQV 1607
EqERV           SPTLCQKFVDMALMQVRQKYPSIYLIYMDDLLLAHQDRSYLQQVLQETVEALTLYGLQV 1607
Chr117357       SPHLFGRALGKDLWDV--QLPVGGLIQYVDDILICSPTKENSHHNTVLVLNFFPNRDIAC 1281
chr3017         SPHLFGRALGKDLWDV--QLPVGGLIQYVDDILICSPTKENSHHNTVLVLNFFPNRDIAC 1281
chr5.119        SPHLFGSALGKDLRDV--QLPVGGLIQYVDDILICSPTKENSDHNTVLVLNFLAQRGYRV 1283
chr101849       SPHLFGSALGNDLQDV--QLPVGGLIQYVDDILICSPTKDSSDQNTVLVLNSLAPQRYRV 2604
chr17.14        SPTLFNETLAKDLRGL--QLNQGTLLQYVDDLLIASPSYQHCLNNTIIMPNHLAWCGYKV 1320
chr26.7254      SPTICHGLIADLATWA--HSDAVPVFQYIDDVMLTSDSLADLEQADDSLRRHLATCGWAV 1042
                ** :    :         :        *:**::                .
```

Figure 19. Alignment of translated pol sequences between only horse ERVs, The conserved YXDD (where X=M,V, or I) motif is shown with asterisks. YIDD motif was the fewest one while YMDD was the most common in all other candidates.

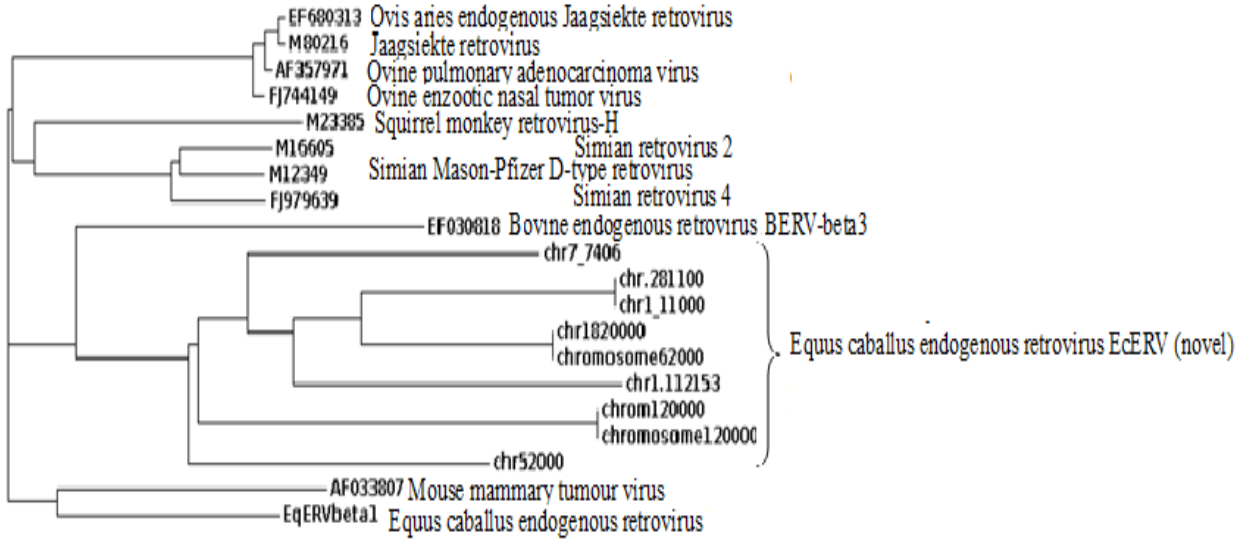


Figure 19.Phylogenetic tree of translated pol sequences of novel EcERVs, EcERV Beta1, beta- and delta retroviruses. Genbank accession numbers are indicated.

Novel EcERVs were branched closer to each other and resemble most closely with Bovine ERV-beta3 in the phylogenetic tree (Figure 19). And EcERV-beta1 was branched with murine retrovirus, Mouse mammary tumour virus.

Beta retroviruses from sheep and primates are only distantly related to EcERVs and EcERV-beta1.

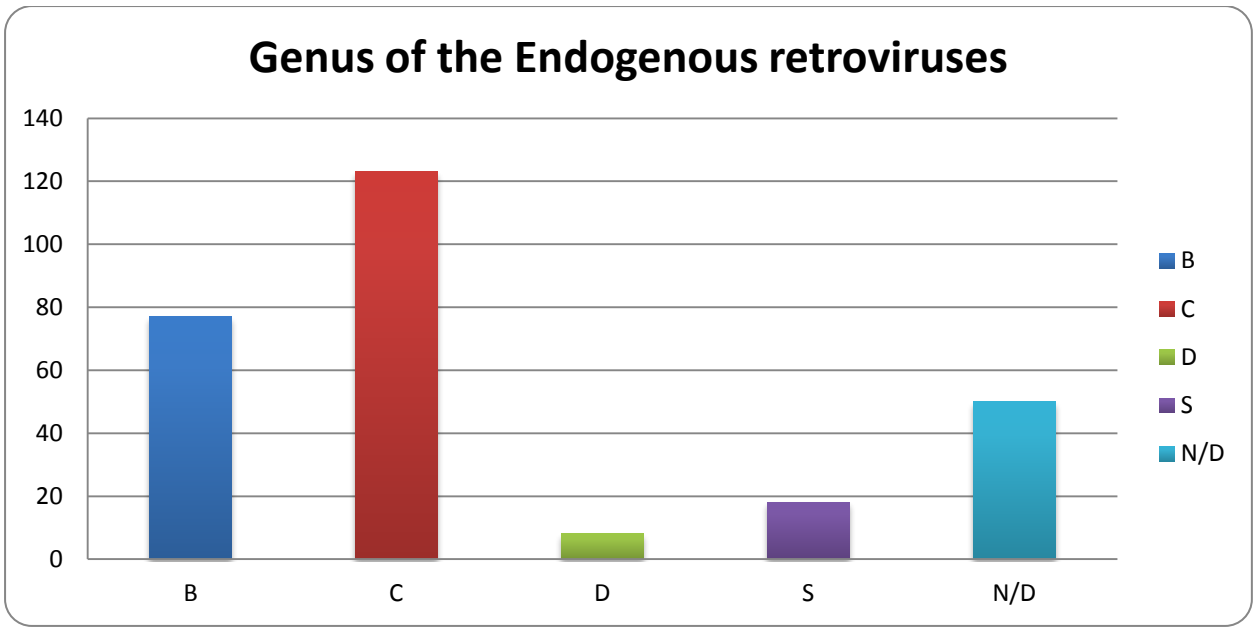


Figure 19. Genus of the Equine endogenous retroviruses (based on the results of Retrotector© online tool)

Figure 10 shows that Gammaretroviruses (44,2%) dominated between genus of equine ERVs. The second wide-spread genus was beta retroviruses (28.6%) from the results of our study. Delta (2.5%) and spumaretroviruses (6.5%) were minorities of the EcERVs. 18.1% of candidates have not been determined. 226 out of 276 candidates were determined. If we calculate percentage again by not including undefined candidates, 53,9% or 122 were gammaretroviruses, 34,9% or 79 were betaretroviruses. 3% or 7 were deltaretroviruses, 7.9% or 18 were spumaretroviruses from determined candidates.

4.2.4 Unique integrations

Retroviral colonization of the germ line can have a range of consequences for the host organism.[22]
These 27 complete endogenous retroviruses were scrutinized. By determining their locations on the horse genome we have found intriguing integrations on several chromosomes. Some integrations or their highly resemble variants are overlapped with known genes or they are neighbouring with functional genes. ERVs could alter function of adjacent genes using their Retroviral LTRs which are 500-600 nucleotides long contain strong cell-specific transcriptional regulatory elements such as compact, mobile promoter, enhancer sequences hormone responsive elements, and polyadenylation signals. LTRs are commonly found upstream of genes in antisense orientation or downstream in sense orientation[6]
Most interesting integrations that could be co-adaptation of ERV and host gene or cause of the diseases are shown on the Table 4. There are plenty of neighbour genes in other places and NCBP1 is one of them. We focused mostly on the integrations within the gene. There is another candidate on chromosome 5 which is named chr5_2000. Its 95% similar variant has been overlapped on

NW_001867366.1 Loc100073073 syntaxin 8 like. Syntaxin 8 impairs trafficking of cystic fibrosis transmembrane conductance regulator (CFTR) and inhibits its channel activity (human)[15]
The most interesting result has been found on chromosome 10. The ERV was on Loc 100630429 locus which is carcinoembryonic antigen related cell adhesion molecule 1 like. And the ERV was 14280 bp long which is the size of complete endogenous retrovirus. Our suspection is that there could be influence of ERVs on the function of the gene. (see appendix 2).
If we put all 276 candidates on the horse genome sequence, we could find some more integrations that might influence the adjacent genes because they all contain both LTRs. Table 8 shows the most interesting integrations with neighbouring or integrated genes their location and function of the genes

Table 8. ERVs could be related to host animal's gene

Location of the ERV	The name of the gene	Function of the gene	Location of the gene	Identity with complete ERV and length
4640-4647 k Neighbour genes,	NCBP1, nuclear cap binding protein	a RNA-binding protein which binds to the 5' cap of RNA polymerase II.	4647-4680 k NW_001867396,1	92%, 6923 bp
2830-2837 k The whole Integration has inserted within the gene	FBP2-fructose1,6-bisphosphatase 2	an enzyme in the liver that converts fructose-1,6-bisphosphate to fructose 6-phosphate in gluconeogenesis	2795-2845 k NW_001867394,1	90%, 7503 bp
11714-11721 k The whole Integration has inserted within the gene which covers exon9.	Loc100630182 cytochrome 450 4A11-like ***Equus caballus***	membrane-bound hemoproteins that contain heme groups and carry out electron transport.	11703-11721 k NW_001867402,1	97%, 7291 bp
50932-50941 k The whole Integration has inserted within the gene. The ERV was divided into 2 pieces	Loc100073073 syntaxin 8 [15]	play a role in determining the specificity of vesicular trafficking	50800-51500 k NW_001867366.1	99%, 7567
12306-12320 k The whole	Loc 100630429	carcinoembryonic antigen related cell adhesion molecule 1	12260-12360 k	100%,

Integration has inserted within the gene		like[28]		14280 bp
7977-7981 k The whole Integration has inserted within the gene	TMEM131 transmembrane protein	Many TPs function as gateways or "loading docks" to deny or permit the transport of specific substances across the biological membrane, to get into the cell, or out of the cell as in the case of waste byproducts.	7915-8115 k NW_001867379,1	100%, 8460 bp

5 CONCLUSIONS

It was shown that SINE-PCR approach is available to find novel ERVs from horse genome by finding the nearest SINE element of the known beta retrovirus which belongs to ERE1 family.

Pan-PCR has worked well on horse genome as well as other species genome. Seven EcERVs were found from unassembled region of horse genome and three ERVs were found from chromosome 5 as variants of EcERV beta1.

276 EcERV elements were discovered by LTR_STRUC tool based on the criterium that they should pass the lower limit of 0.3 score. 27 novel and complete EcERVs have been found which is about 10% of all candidates and the first beta ERV has also been found within them therefore we assumed it as a positive control for the *in silico* analysis.

Nine equine ERVs located on the unassembled part of horse genome have been found by NCBI-BLAST tool.

We have found 4 ERVs from chromosome 5 (Table 6) using LTR_STRUC tool and 9 ERVs from unassembled region using NCBI-BLAST tool and 3 out of 4 ERVs from chromosome 5 and 7 out of 9 ERVs from NCBI-BLAST results were same as what we have found from Pan-PCR result. Hence these results show that the bioinformatics and experimental work have complemented each other.

It could be too early to make a conclusion but the picture I have got from this study allows me to make the following inferences. The equine genome has been effective in protection from extensive retroviruses integration, the amount of ERVs in horse was a fifth of the ERV amount in species like human and chimpanzees. Functional study remains to be performed in the future. The relative low abundance of endogenous retroviruses in the horse genome compared to other species suggests that the horse has been able to protect itself from large amounts of insertion of endogenous retroviruses. From preliminary results we have found 276 candidates which is 0,085 % of whole genome that consist of 2.68 Gb and summed up that horse has being effective in protecting themselves from large amounts of retroviruses as chicken.

We have studied the polymorphism between breeds on EcERV-Beta1. Integrations with 1% divergence between LTRs have polymorphism between breeds.

The complexity of horse endogenous retroviruses was identified and classified to the retroviral genera. EcERVs were classified and characterized using a bioinformatics approach and experimental approaches. The highest scored chains were characterized according to their functional capacity.
Some variants have been known that they exist on functional genes like syntaxin 8, TMEM131 which encodes a Transmembrane protein
The most interesting result has been found on chromosome 10. The ERV was on Loc 100630429 locus which is carcinoembryonic antigen related cell adhesion molecule 1 like. The ERV was 14280 bp long which is the size of complete endogenous retrovirus. We suspect that there could be influence of ERV on the function of the gene.[28] It remains to be studied in the future.
Using Retrotector© online tool, the following classes were identified in the horse genome: 53,9% or 122 candidates were determined as gammaretroviruses, 34,9% or 79 were determined as betaretroviruses. 3% or 7 were determined as deltaretroviruses, 7.9% or 18 were determined as spumaretroviruses from determined endogenous retroviruses.
This study has demonstrated the importance of using multiple methods when trying to identify new ERVs and showed that the number of Equine ERVs is not as limited as previously thought.[29] Importance of the study is to contribute to the knowledge of ERVs' distribution between different species.

6 DISCUSSION

The integration polymorphism between breeds of EcERV Beta1 is an interesting result. The primers were designed on the bases of pol conserved gene of EcERV Beta1. Some breeds have 2 products while some have none. Most of the breeds possess same region as thoroughbred has. There were some bands with slightly different size that could be caused by insertion or deletion mutations. It could be due to primers that are not specific enough for testing a polymorphism.
Y chromosome has not been included in this analysis because the horse genome sequence that we mined was made from a mare. More ERVs could have been found if we included Y chromosome because it is assumed as "graveyard" of ERVs.[14] Chromosome 29 and 31 are lack of equine ERVs from this result. It remains to be confirmed that these chromosomes are essentially lacking e*quine ERV*s or whether it reflects annotation-bias of these chromosomes.
Previous published research on equine endogenous retrovirus was limited. In The Retroviridae book volume 2 page 258. "A number of studies have probed various equine tissues for the presence of endogenous retroviruses. (Rice et al, 1978, 1989. Rasty et al 1990, O'Rourke et al, 1991). None of the studies (Southern blots, PCR) have detected endogenous retrovirus sequences in tissues of equine origin, although more sensitive techniques such as nested PCR have not yet been used to search for equine retroelements. However, in 2011, 20 years later Van der Kuyl found the first Equine endogenous beta retrovirus using by Blast search. In the present study we have found 27 novel and complete ERVs with other candidates. The low amount of *EcERVs* in horse i.e. 276 unique chains scored more than 0.3 with LTR_STRUC in this study, was similar to the amount found in chicken, G*allus gallus* (gg01) with Retrotector© in a study made by Jern, P. and colleagues in 2005, 262 elements were identified and a similar amount found in dog[5], The result of LTR_STRUC tool confirms that the amounts of *ERV* in horse, dog and chicken are very low compared with human,

schimpanzee and bovine genomes. However improvement of bioinformatics tool could allow to obtain more EcERVs.

We have found several unique integrations within the functional genes that may be cause of cancer or imprint of co-adaptation between host and retrovirus. Companionship between human and horse for over 6000 years since the horse was domesticated in 4000 BC. There is a possibility that horizontal contagion might have occurred.

We suggest to annotate equine endogenous retrovirus as EcERV instead of EqERV. The abbreviation, EcERV represents Equus caballus Endogenous RetroVirus as same as CfERV (Canis familiaris Endogenous RetroVirus)

7 FURTHER ANALYSIS:

We have found that nine equine ERVs located on the unassembled part of horse genome. BLAST results show that they were 99% identical to each other. This information can be contributed to improve unassembled parts of equine genome.

Another idea is to develop detailed catalogue which can display the equine endogenous retroviruses (EcERVs).

The following investigations may be seminal for an improved understanding of the biological significance of EcERVs.

The search for complete proviruses should be continued for those which are only partially characterized.

Studies should be continued to investigate expression of EcERVs genes at the RNA and protein level.

Studies should be intensified to search for adjacent genes that are influenced or controlled by EcERV enhancers or promoters, by UTRs, or by polyadenylation signals.

REFERENCES:

.

1. Van der Kuyl, A.C. “Characterization of a Full-Length Endogenous Beta-Retrovirus, EqERV-Beta1, in the Genome of the Horse (Equus caballus).” Viruses 2011, 3, 620-628.
2. Richard Cordaux and Mark Batzer (October 2009). "The impact of retrotransposons on human genome evolution". Nature Reviews Genetics 10 (10): 691–703.
3. Patric Jern, John M. Coffin “Effects of retroviruses on host genome function.” Annual Review of Genetics Vol. 42: 709-732
4. Jonas Blomberg, Dmitrijs Ushameckis, and Patric Jern “Evolutionary Aspects of Human Endogenous Retroviral Sequences (HERVs) and Disease.” Madame Curie Bioscience Database chapter Viruses 2000, http://www.landesbioscience.com/curie/chapter/2214/
5. Martínez Barrio Á, Ekerljung M, Jern P, Benachenhou F, Sperber GO, et al. “The First Sequenced Carnivore Genome Shows Complex Host-Endogenous Retrovirus Relationships.” PLoS ONE 2011, 6(5): e19832. doi:10.1371/journal.pone.0019832
6. Coffin JM, Hughes SH, Varmus HE, “Retroviruses”. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press; 1997. Chapter 2, 5, 8.

7. Mi S, Lee X, Li X, Veldman GM, et al. ”Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis.” Nature, Feb 2000, 403(6771):785–9, doi: 10.1038/35001608.
8. Samuelson LC, Wiebauer K, Snow CM, & Meisler MH. “Retroviral and pseudogene insertion sites reveal the lineage of human salivary and pancreatic amylase genes from a single gene during primate evolution.” Mol Cell Biol, Jun 1990, 10(6):2513–20,.
9. Medstrand P, van de Lagemaat LN, & Mager DL. ”Retroelement distributions in the human genome: variations associated with age and proximity to genes.” Genome Res, 12(10):1483–95, Oct 2002. doi: 10.1101/gr.388902.
10. Masayuki Sakagami, Kazuhiko Ohshima, Harutaka Mukoyama, Hiroshi Yasue, Norihiro Okada “A Novel tRNA Species as an Origin of Short Interspersed Repetitive Elements (SINEs) : Equine SINEs May Have Originated From tRNASer,” Journal of Molecular Biology (1994) Volume: 239, Issue: 5, Pages: 731-735
11. Jern P “Genomic Variation and Evolution of HERV-H and other Endogenous Retroviruses (ERVs)” 2005 Vol 62, http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-5906
12. Jern P, Sperber GO, Blomberg J “Use of endogenous retroviral sequences (ERVs) and structural markers for retroviral phylogenetic inference and taxonomy.” Retrovirology. 2005 Aug 10;2:50.
13. Koldo Garcia-Etxebarria and Begona Marina Jugo “Genome-Wide Detection and Characterization of Endogenous Retroviruses in Bos Taurus.” Virology. Oct, 2010 p. 10852-10862
14. Kjellman C, Sjögren HO, Widegren B. “The Y chromosome: a graveyard for endogenous retroviruses.” 1995 Aug 19;161(2):163-70.
15. Bilan F, Thoreau V, et al. “Syntaxin 8 impairs trafficking of cystic fibrosis transmembrane conductance regulator (CFTR) and inhibits its channel activity (human)” April 15, 2004 J Cell Sci 117, 1923-1935.
16. Ekerljung, Marie “Molecular systematics: data mining of canine endogenous retroviruses, CFERV.” Dept. of Animal Breeding and Genetics, SLU. 2007, Vol. 295.
17. Luis P. Villarreal, “Viruses and the evolution of life,” ASM Press, Washington, D.C., 2005.
18. Wendy Ashlock, “Detecting retroviruses in genomic sequences and applying signal processing techniques to genomics,” Literature review, 2010
19. Retrotector online tool http://www.kvir.uu.se/Retrotector©/Retrotector©Project.html
20. LTR_STRUC version 1.1 http://www.mcdonaldlab.biology.gatech.edu/finalLTR.htm
21. S. Saha et al., “Empirical comparison of ab initio repeat finding programs,” Nucleic Acids Research 36, 2008, no. 7, 2284-2294.
22. Jonathan P. Stoye, “Studies of endogenous retroviruses reveal a continuing evolutionary saga” Nature Reviews, Microbiology Volume 10, June 2012 page no.397
23. Greenwood, A.D.; Lee, F.; Capelli, C.; DeSalle, R.; Tikhonov, A.; Marx, P.A.; MacPhee, R.D. “Evolution of endogenous retrovirus-like elements of the woolly mammoth (Mammuthus primigenius) and its relatives.” Mol. Biol. Evol. 2001, 18, 840–847.
24. Wade, C.M.; Giulotto, E.; Sigurdsson, S.; Zoli, M.; Gnerre, S.; Imsland, F.; Lear, T.L.; Adelson, D.L.; Bailey, E.; Bellone, R.R.; et al. “Genome sequence, comparative analysis, and population genetics of the domestic horse.” Science 2009, 326, 865–867.

25. Horse Genome Resources, NCBI. Available online: http://www.ncbi.nlm.nih.gov/projects/genome/guide/horse/ (accessed on 4 April, 2012).
26. NCBI Basic Local Alignment Search Tool BLAST. Available online: http://blast.ncbi.nlm.nih.gov/ (accessed on 4 April, 2012).
27. Horse (Equus caballus) Genome Browser Gateway of the Genome Bioinformatics Group of UC Santa Cruz. Available online: http://genome.ucsc.edu/cgi-bin/hgGateway?db=equCab2 (accessed on 4 January 2011).
28. Singer BB, Scheffrahn I, Heymann R, Sigmundsson K, Kammerer R, Obrink B. "Carcinoembryonic antigen-related cell adhesion molecule 1 expression and signaling in human, mouse, and rat leukocytes: evidence for replacement of the short cytoplasmic domain isoform by glycosylphosphatidylinositol-linked proteins in human leukocytes." J Immunol. 2002 May 15;168(10):5139-46
29. AvJay A. Levy "The Retroviridae book" volume 2 1993, page 258.
30. Olga R. Borodulina, Dmitri A. Kramerov "PCR-based approach to SINE isolation: Simple and complex SINEs" Gene 349 (2005) 197–205
31. Thomas Ericsson, Beth Oldmixon, Jonas Blomberg, Margaret Rosa, Clive Patience, and Göran Andersson "Identification of Novel Porcine Endogenous Betaretrovirus Sequences in Miniature Swine" J. Virol. March 2001 vol. 75 no. 6 2765-2770

APPENDICES

Appendix 1. Equine endogenous retrovirus candidates

Table . Rest 214 equine endogenous retroviruses EcERVs, found by LTR_STRUC tool (best scored ones are listed in Bioinformatics results section) Direct repeats, polypurine tracts sequences, primer binding site sequences, dinucleotides are found but not included in the table.

Chrom.1	Genus	CUT_OFF score	Overall length	LTR PAIR HOMOLOGY
1	C	0.3	9178 bp	92.60%
2	C	0.4	11835 bp	90.80%
3	C	0.42	7170 bp	93.00%
4	BC	0.44	1512 bp	94.80%
5	SC	0.47	14704 bp	82.60%
6	BC	0.52	17268 bp	80.40%
7	Unknown	0.6	4752 bp	83.30%
8	B	0.65	6654 bp	88.50%
9	S	0.76	7373 bp	97.10%
10	C	1	5426 bp	86.10%
Chrom.2	Genus	CUT_OFF score	Overall length	LTR PAIR HOMOLOGY
11	SC	0.33	2039 bp	90.00%
12	C	0.41	7008 bp	86.60%
13	CS	0.52	12692 bp	89.60%
14	BC	0.6	1781 bp	98.10%

15	C	0.67	6058 bp	97.50%
16	C	0.88	6918 bp	84.00%
Chrom.3	Genus	CUT_OFF score	Overall length	LTR PAIR HOMOLOGY
17	C	0.34	4589 bp	89.40%
18	BS	0.36	7730 bp	97.80%
19	SC	0.48	14061 bp	79.60%
20	CB	0.56	8894 bp	83.80%
21	SC	0.57	6589 bp	88.40%
22	C	0.57	9829 bp	82.30%
23	CB	0.64	4146 bp	95.30%
24	DB	0.67	6684 bp	98.40%
25	C	0.75	6089 bp	97.80%
26	B	0.81	1312 bp	99.00%
27	S	0.82	12236 bp	92.70%
Chrom.4	Genus	CUT_OFF score	Overall length	LTR PAIR HOMOLOGY
28	C	0.32	2514 bp	87.60%
29	CB	0.42	9016 bp	91.40%
30	CB	0.47	8474 bp	93.10%
31	C	0.69	12194 bp	88.00%
32	C	0.74	6527 bp	84.00%
33	C	0.74	6117 bp	93.60%
34	BC	0.75	4817 bp	85.60%
35	BS	0.92	8464 bp	90.60%
36	BC	0.99	6081 bp	97.70%
Chrom.5	Genus	CUT_OFF score	Overall length	LTR PAIR HOMOLOGY
37	C	0.31	2656 bp	85.10%
38	unknown	0.31	1732 bp	88.50%
39	unknown	0.35	7357 bp	80.80%
40	BS	0.37	13029 bp	92.00%
41	B	0.38	13691 bp	94.20%
42	CS	0.42	7500 bp	85.50%
43	BS	0.44	13030 bp	88.10%
44	SB	0.52	10524 bp	88.80%
45	CD	1.22	13704 bp	68.60%
Chrom.6	Genus	CUT_OFF score	Overall length	LTR PAIR HOMOLOGY
46	unknown	0.3	1442 bp	91.80%
47	unknown	0.36	3097 bp	93.10%
48	BC	0.39	17331 bp	64.10%
49	C	0.39	10284 bp	83.60%
50	C	0.41	8225 bp	83.00%
51	BC	0.55	6312 bp	89.80%
52	CB	0.85	14325 bp	91.40%

Chrom.7	Genus	CUT_OFF score	Overall length	LTR PAIR HOMOLOGY
53	C	0.3	15117 bp	92.70%
54	BC	0.34	15672 bp	91.20%
55	C	0.35	20540 bp	88.10%
56	C	0.38	6192 bp	85.10%
57	C	0.38	1667 bp	91.70%
58	DC	0.4	6841 bp	85.40%
59	B	0.74	8025 bp	99.60%
60	C	0.74	7137 bp	98.30%
Chrom.8	Genus	CUT_OFF score	Overall length	LTR PAIR HOMOLOGY
61	unknown	0.3	2149 bp	76.30%
62	unknown	0.32	1314 bp	94.60%
63	B	0.43	9737 bp	81.20%
64	unknown	0.49	1437 bp	88.00%
65	C	0.66	6228 bp	88.50%
66	B	0.67	7978 bp	97.90%
67	SC	0.68	21595 bp	91.40%
68	DC	0.74	8149 bp	97.80%
Chrom.9	Genus	CUT_OFF score	Overall length	LTR PAIR HOMOLOGY
69	S	0.33	8914 bp	70.40%
70	unknown	0.36	4662 bp	93.10%
71	unknown	0.38	2509 bp	94.00%
72	B	0.5	12123 bp	95.00%
73	C	0.52	24320 bp	83.30%
74	C	0.53	8383 bp	93.70%
75	CS solo ltr	0.66	7622 bp	91.60%
76	B	0.66	6675 bp	97.60%
Chrom.10	Genus	CUT_OFF score	Overall length	LTR PAIR HOMOLOGY
77	CB	0.32	3802 bp	92.30%
78	C	0.33	7196 bp	90.30%
79	C	0.36	7709 bp	85.20%
80	C	0.36	19940 bp	80.50%
81	C	0.47	8472 bp	96.50%
82	CB	0.6	8680 bp	69.90%
83	B	0.65	6654 bp	88.50%
84	SC	0.66	7623 bp	93.30%
85	C	0.71	13421 bp	86.30%
Chrom.11	Genus	CUT_OFF score	Overall length	LTR PAIR HOMOLOGY
86	S	0.31	10558 bp	76.20%
87	BC	0.32	20225 bp	90.60%
88	C	0.42	10492 bp	90.60%
89	C	0.45	11362 bp	87.70%

90	C	0.45	9199 bp	81.30%
91	unknown	0.53	15784 bp	84.20%
92	SC	0.55	15055 bp	91.20%
93	SC	0.55	14241 bp	44.10%
94	BC	0.6	13872 bp	88.60%
Chrom.12	Genus	CUT_OFF score	Overall length	LTR PAIR HOMOLOGY
95	CS solo ltr	0.37	1749 bp	98.40%
96	C	0.58	18607 bp	89.30%
97	BC	0.69	3542 bp	91.40%
98	C	1	2511 bp	99.10%
Chrom.13	Genus	CUT_OFF score	Overall length	LTR PAIR HOMOLOGY
99	unknown	0.43	5971 bp	89.90%
100	CB	0.53	13799 bp	86.20%
101	unknown	0.56	1359 bp	91.60%
102	B	0.61	8840 bp	98.20%
Chrom.14	Genus	CUT_OFF score	Overall length	LTR PAIR HOMOLOGY
103	C	0.32	7976 bp	82.50%
104	CS	0.37	12489 bp	84.10%
105	CB	0.49	12503 bp	88.90%
106	CB solo LTR	0.5	17522 bp	91.40%
107	BC	0.58	7866 bp	91.90%
108	BC	0.64	8883 bp	85.50%
109	B	0.66	7527 bp	98.50%
110	C	0.66	15148 bp	88.30%
111	C	0.74	6074 bp	95.80%
Chrom.15	Genus	CUT_OFF score	Overall length	LTR PAIR HOMOLOGY
112	unknown	0.33	1909 bp	98.80%
113	SC	0.34	8556 bp	89.00%
114	C	0.36	7730 bp	97.80%
115	C	0.38	2304 bp	93.10%
116	B	0.4	17106 bp	82.10%
117	unknown	0.43	1154 bp	87.10%
118	unknown	0.47	4018 bp	87.50%
119	C	0.48	3634 bp	91.60%
120	C	0.52	11484 bp	86.40%
121	C	0.57	9829 bp	82.30%
122	CB	0.67	6684 bp	98.40%
Chrom.16	Genus	CUT_OFF score	Overall length	LTR PAIR HOMOLOGY
123	B	0.34	1357 bp	90.30%
124	BC solo LTR	0.4	2192 bp	83.10%
125	C	0.42	9016 bp	91.40%
126	B	0.5	6643 bp	89.90%

127	C	0.5	6099 bp	98.30%
128	C	0.69	12194 bp	88.00%
129	BC	0.75	4817 bp	85.60%
130	C	0.92	8464 bp	90.60%
131	BC	0.99	6081 bp	97.70%
Chrom.17	Genus	CUT_OFF score	Overall length	LTR PAIR HOMOLOGY
132	B solo LTR	0.36	1738 bp	87.20%
133	C	0.43	2571 bp	90.60%
no.18	Genus	CUT_OFF score	Overall length	LTR PAIR HOMOLOGY
134	unknown	0.3	19723 bp	96.80%
135	unknown	0.36	3097 bp	93.10%
136	C	0.44	10514 bp	90.10%
Chrom.19	Genus	CUT_OFF score	Overall length	LTR PAIR HOMOLOGY
137	C	0.3	15117 bp	92.70%
138	unknown	0.3	2126 bp	74.80%
139	B	0.35	20540 bp	88.10%
140	C	0.38	6192 bp	85.10%
141	C	0.38	1667 bp	91.70%
Chrom.20	Genus	CUT_OFF score	Overall length	LTR PAIR HOMOLOGY
142	C	0.55	13425 bp	95.90%
143	C	0.55	17230 bp	84.40%
144	C	0.57	9611 bp	87.80%
145	C	0.92	17948 bp	83.30%
Chrom.21	Genus	CUT_OFF score	Overall length	LTR PAIR HOMOLOGY
146	C	0.32	16612 bp	86.20%
147	unknown	0.55	5657 bp	92.00%
148	C	0.67	6226 bp	92.40%
Chrom.22	Genus	CUT_OFF score	Overall length	LTR PAIR HOMOLOGY
149	unknown	0.32	1625 bp	61.70%
150	unknown	0.35	1496 bp	83.00%
151	unknown	0.37	2404 bp	79.40%
152	C	0.42	5192 bp	83.10%
153	C	0.48	7945 bp	83.30%
154	C	0.64	13294 bp	85.70%
155	unknown	0.75	1565 bp	77.30%
Chrom.23	Genus	CUT_OFF score	Overall length	LTR PAIR HOMOLOGY
156	unknown	0.32	1352 bp	88.00%
157	SC	0.32	17889 bp	77.00%
158	SC	0.33	1922 bp	87.10%
159	unknown	0.34	3016 bp	88.40%
160	too short	0.35	1222 bp	91.30%
161	C	0.4	1879 bp	84.10%

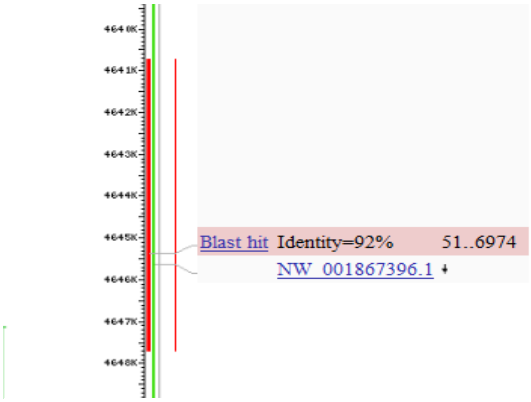
162	too short	0.4	1370 bp	58.20%
163	unknown	0.45	1447 bp	84.00%
164	too short	0.5	2107 bp	60.20%
165	unknown	0.52	2644 bp	86.90%
166	unknown	0.54	4368 bp	92.30%
167	C	0.6	12130 bp	92.30%
168	unknown	0.62	2644 bp	88.10%
169	unknown	0.62	3016 bp	85.80%
Chrom.24	Genus	CUT_OFF score	Overall length	LTR PAIR HOMOLOGY
170	C	0.4	1584 bp	85.80%
171	B	0.66	9093 bp	98.10%
Chrom.25	Genus	CUT_OFF score	Overall length	LTR PAIR HOMOLOGY
172	unknown	0.32	18199 bp	81.70%
173	unknown	0.33	14252 bp	75.20%
174	BC	0.36	9691 bp	86.40%
175	nothing	0.42	3445 bp	85.80%
176	B solo LTR	0.5	4843 bp	76.90%
177	S	0.52	6279 bp	84.10%
178	C	0.53	10648 bp	85.80%
179	C	0.56	7572 bp	97.60%
180	C	0.61	7894 bp	82.60%
181	nothing	0.68	4970 bp	86.90%
182	BC	0.68	7471 bp	99.50%
183	C	0.74	6933 bp	85.20%
184	C	0.96	16556 bp	86.60%
185	BC	1	6667 bp	93.80%
186	B solo LTR	1.05	15236 bp	88.60%
Chrom.26	Genus	CUT_OFF score	Overall length	LTR PAIR HOMOLOGY
187	C solo LTR	0.3	2776 bp	82.60%
188	C	0.39	15057 bp	82.00%
189	C	0.43	9446 bp	76.30%
190	B	0.51	7412 bp	90.50%
191	C	0.63	10370 bp	87.10%
192	S	0.66	7866 bp	97.70%
193	C	0.84	12922 bp	87.20%
194	C	0.86	3621 bp	92.30%
Chrom.27	Genus	CUT_OFF score	Overall length	LTR PAIR HOMOLOGY
195	unknown	0.4	11558 bp	83.70%
196	unknown	0.71	5629 bp	90.50%
Chrom.30	Genus	CUT_OFF score	Overall length	LTR PAIR HOMOLOGY
197	C	0.49	12503 bp	88.90%
198	BC	0.52	17268 bp	80.40%

199	nothing	0.6	4752 bp	83.30%
200	B	0.76	7373 bp	97.10%
Chrom.X	Genus	CUT_OFF score	Overall length	LTR PAIR HOMOLOGY
201	C	0.31	4297 bp	90.20%
202	B	0.32	3325 bp	94.10%
203	unknown	0.33	14252 bp	75.20%
204	BC	0.36	9691 bp	86.40%
205	nothing	0.4	2247 bp	89.10%
206	nothing	0.42	3445 bp	85.80%
207	C solo LTR	0.46	20434 bp	84.90%
208	DC	0.47	2490 bp	96.50%
209	S	0.52	6279 bp	84.10%
210	nothing	0.68	4970 bp	86.90%
211	BC	0.68	7471 bp	99.50%
212	DC	0.74	7619 bp	97.80%
213	C	0.96	16556 bp	86.60%
214	C	1	6667 bp	93.80%

Appendix 2. Unique integrations

Appendix 2 shows the unique integrations that have been found from the known functional genes or closer region to the genes.

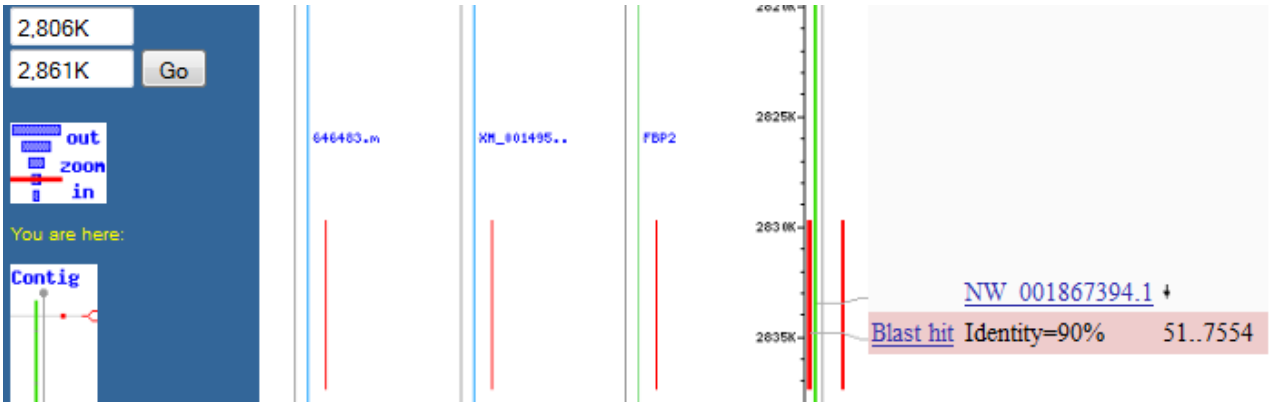
NW_001867396,1 NCBP1, nuclear cap binding protein



NCBI Reference Sequence: NW_001867396.1

GenBank FASTA

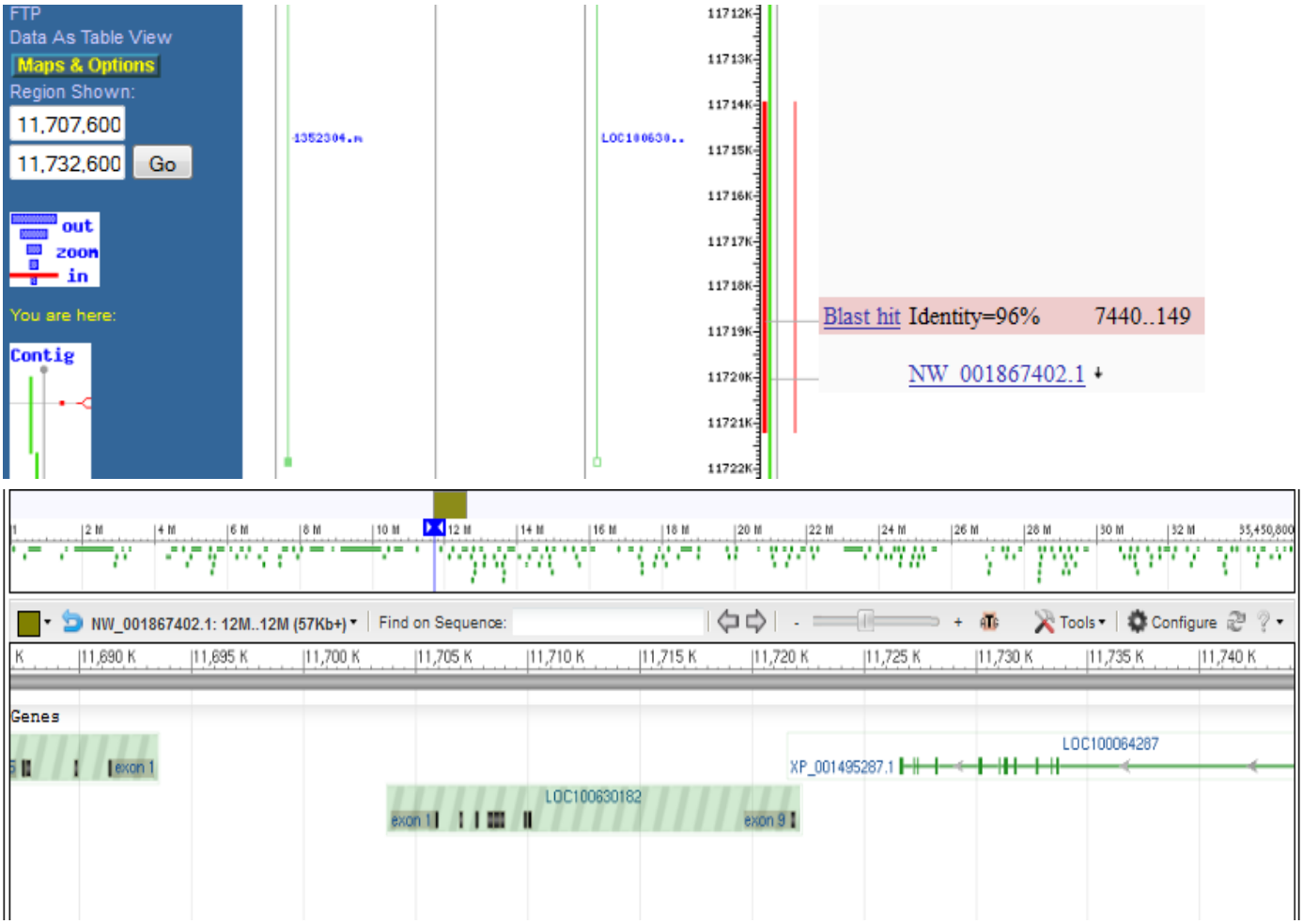
NW_001867394,1 FBP2, FBP2 fructose-1,6-bisphosphatase 2



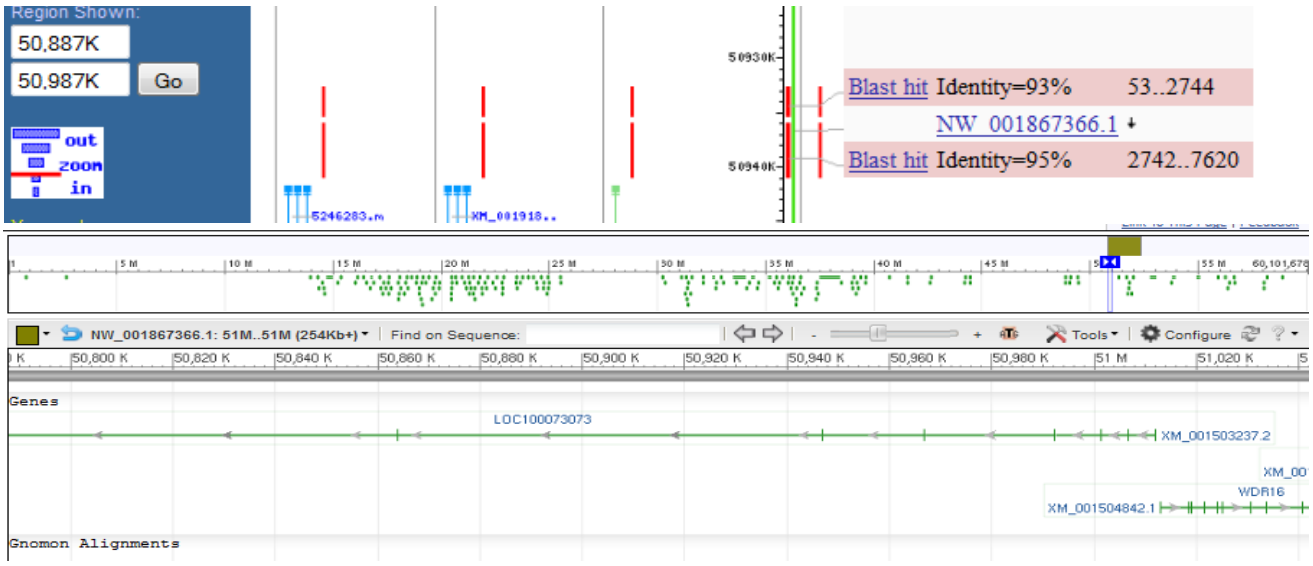
NCBI Reference Sequence: NW_001867394.1

GenBank FASTA

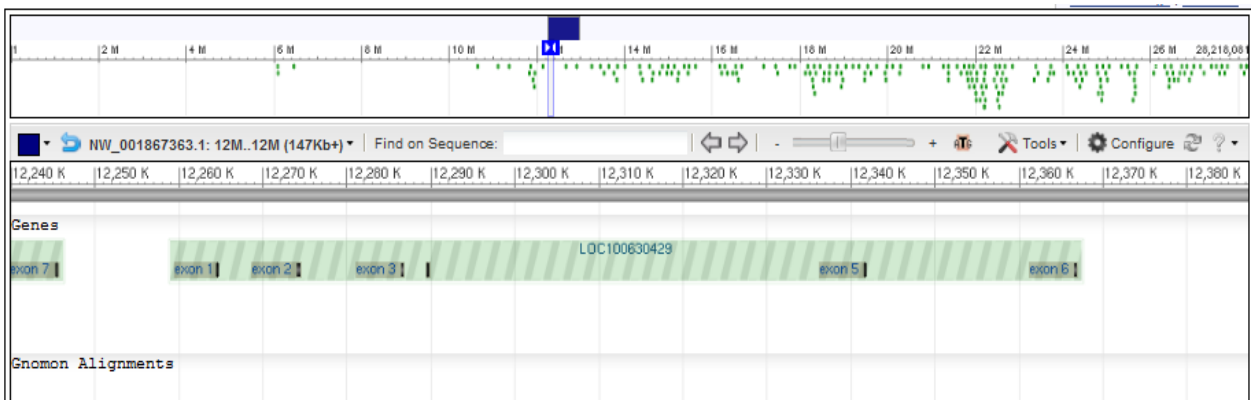
NW_001867402,1 Loc100630182



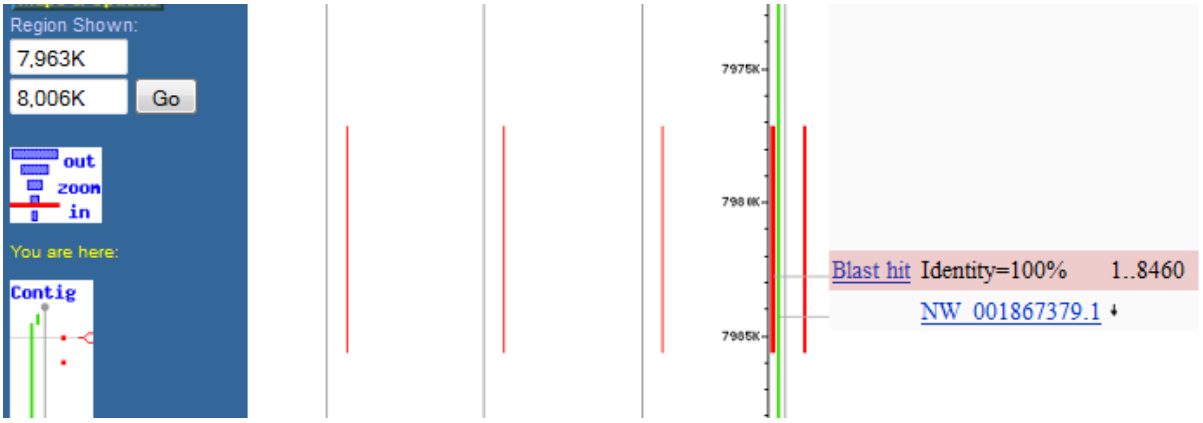
NW_001867366,1 Loc100073073 syntaxin 8 like

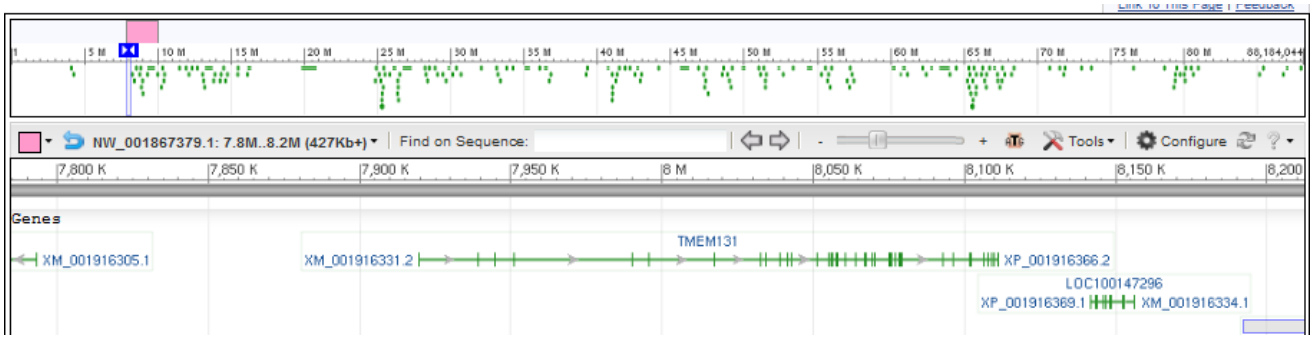


NW_001867363.1 Loc100630429 carcinoembryonic antigen related cell adhesion molecule 1 like



NW_001867379,1 TMEM131 transmembrane protein





Appendix 3. Sequencing results

>Pan_PCR_sequence

TGGGCCCTCTAGATGCATGCTCGAGCGGCCGCCAGTGTGANGGATATCTGCAGAATTCG
CCCTTGTCGGAACCAATTTATATCTCCCAAGAGCCTTTGAAAGTCATTTAAAGTTTTCAA
GGAATCTGTTCGTAATTGGATGTTTTGGGGAGTCACTAGATCAGTGTGAATGACTCGACC
AAAATAATTAATGGGTGGATTAGTCTGTATTTTTTCAGGTTCCACTTGTAGGCCATAAAG
AGTCAAAGCCTTAACTGTTTCTTGTAATACTTGCTGTAAATAACTTCTATCTTGATGTGCC
AGCAATGTCATCCATGTAATGGATCAGGTAGATAGAAGGATATTTTTGTCTGACCTGTAT
CACAGCCATCTCAACAAATTTTTGGCACAATGTTGGACTGTTTTTCATTCCTCGTGGCAA
CACTTTCCATTGAAATCTCTGATAAGGTCTTCTCAAGTTTTCTGAAGGTAAGCTGAAGGC
AAATCGGGGCCTGTCCTTAGGGTCTAAGGGGATGTTGAAAAAACAGTCTTGTAAGTCTA
TAACCACAACAGCAGCCTGAACTGGGACTGCTGCAGGAGAAGTCAGTCCCTGTTGTAGG
ACCCCCATATCTTGCATGGTCTCACTAATGGCTCTCAGATCTTGTAACAATCTCCATTTTC
CAGATTTCTTTTTGATATGAAAACGGGAGTATTCCAAGGAAGGGCGAATTCCAGCACAC
TGGCGGCCGTTACTAGTGGATCCGAGCTCGGTACCAAGCTTGGCGTAATCATGGTCATA
GCTGTTTCCTGTGTGAAATTGTTATCCGCTCACAATTCCACACAACATACGAGCCGGAAG
CATAAAGTGTAAAGCCTGGGGTGCCTAATGAGTGAGCTAACTCACATTAATTGCGTTGC
GCTCACTGCCCGCTTTCCAGTCGGGAAACCTGTCGTGC

ACKNOWLEDGEMENTS:

I would like to express my gratitude to all those who helped me to accomplish this study.

I would like to express my deep gratitude for kindly support and guidance to my supervisor Professor Göran Andersson at the Department Animal Breeding and Genetics SLU, who encouraged and challenged me through my academic program.

I would like to express my deep gratitude for kindly support and guidance to my supervisor, Associate Professor Erik Bongcam-Rudloff at the Linnaeus Centre for Bioinformatics (LCB) Uppsala University, who encouraged and challenged me through bioinformatics part of the study who guided me on bioinformatics part.

Great thanks for kindly support, guidance, to my supervisor Professor Matthew Peter Kent at the Department Animal Breeding and Genetics, UMB.

Great thanks for kindly support, guidance, and for providing all samples to Professor Sofia Mikko at the Department Animal Breeding and Genetics SLU.

Аав С.Энхбаатар, ээж Б.Батдулам, ах Э.Магнайбаяр, миний хайрт Н.Тэгшжаргал болон Б.Батцэцэг эгчдээ үргэлж тусалж дэмждэгт баярлалаа.